International Journal of Interactive Multimedia and Artificial Intelligence

June 2023, Vol. VIII, Number 2
ISSN: 1989-1660

unir LA UNIVERSIDAD EN INTERNET

"Some people call this artificial intelligence, but the reality is this technology will enhance us. So instead of artificial intelligence, I think we'll augment our intelligence."
Ginni Rometty

INTERNATIONAL JOURNAL OF INTERACTIVE MULTIMEDIA AND ARTIFICIAL INTELLIGENCE

ISSN: 1989-1660 –VOL. 8, NUMBER 2

EDITORIAL TEAM

OPEN ACCESS JOURNAL

ISSN: 1989-1660

The International Journal of Interactive Multimedia and Artificial Intelligence is covered in Clarivate Analytics services and products. Specifically, this publication is indexed and abstracted in: *Science Citation Index Expanded, Journal Citation Reports/ Science Edition, Current Contents®/Engineering Computing and Technology.*

Editor's Note

Artificial Intelligence (AI) represents one of the fastest growing areas of knowledge, sectors and fields of action globally. This growth has allowed to mark different positions, where the most favorable ones are oriented to its unquestionable contribution to facilitate decision making in various fields of society, as well as other sectors that mark a strong position for its use to be carried out in a regulated and measured way due to the scope and risks to which we are exposed. For this reason, rigorous methods are increasingly required for the design and development of AI-based computational models; methods that involve strict mechanisms for their validation, as well as the analysis of possible risks and scope that they may have on the field of application where they are being exposed. This type of aspects would definitely mark a valuable and relevant milestone to define several paths within which we can find two: 1) if it is definitely necessary to set limits on the use of AI by establishing increasingly sophisticated regulatory frameworks on various areas involving data protection and regulated use of the same, and 2) to remove all barriers so that it can be exploited openly in all its dimensions in any area of our society. Hence the importance of analysing the different risks and threats that AI may present within the particular context in which it is being applied.

Based on this panorama, this regular edition of the "International Journal Interactive Multimedia and Artificial Intelligence" presents a series of papers where proposals are oriented to different fields and sectors, which make use of diverse approaches, methods, models and AI-based systems that allow us to have a generalized idea of how these challenges are being addressed in some fields of our society. In particular, this regular issue collects research topics focusing on addressing the problems of evolving recommender systems, classification models, decision support systems, system modelling, data analytics, optimization algorithms, image retrieval, deep neural networks, social network analysis, and the relevance of the design of User Experience (UX) proposals. In order to provide a brief description of all the proposals and contributions, following is a summary of the 23 articles present in this regular issue.

In the field of health, in the proposal entitled "Exploring ChatGPT's Potential for Consultation, Recommendations and Report Diagnosis: Gastric Cancer and Gastroscopy Reports' Case", authors aim to explore ChatGPT tool as a potential tool in disseminating gastric cancer knowledge, providing consultation recommendations, and interpreting endoscopy reports. Through experimentation, authors found that GPT-4 model of ChatGPT achieved an appropriateness of 91.3% and a consistency of 95.7% in a gastric cancer knowledge test.

For its part, in the paper entitled "A Hybrid Parallel Classification Model for the Diagnosis of Chronic Kidney Disease", authors propose a fast and novel hybrid approach to diagnose Chronic Renal Disease. The proposed approach is based on the optimization of SVM classifier with the hybridized dimensionality reduction approach to identify the most informative parameters for CKD diagnosis.

Within the same subject, in the paper entitled "HDDSS: An Enhanced Heart Disease Decision Support System Using RFE-ABGNB Algorithm", authors suggest a heart disease decision-support system (HDDSS) that can predict whether or not a person has heart disease. The main goal of this research work is to use the RFEABGNB algorithm to improve HDDSS prediction accuracy.

In the same line of health and medical fields, the research entitled "ResNet18 Supported Inspection of Tuberculosis in Chest Radiographs with Integrated Deep, LBP, and DWT Features" proposes a TB detection framework using integrated optimal deep and handcrafted features in order to detect tuberculosis in chest radiography. To carry out this research, authors proposed a series of stages that include: (i) X-ray collection and processing, (ii) Pretrained Deep-Learning (PDL) scheme-based feature mining, (iii) Feature extraction with Local Binary Pattern (LBP) and Discrete Wavelet Transform (DWT), (iv) Feature optimization with Firefly-Algorithm, (v) Feature ranking and serial concatenation, and (vi) Classification by means of a 5-fold cross confirmation.

The article entitled "RGBeat: A Recoloring Algorithm for Deutan and Protan Dichromats", authors present a contribution associated with an algorithm that enhances the color perception of deutan and protan dichromats but without compromising the lifelong color perceptual learning. According to the researchers, this is the first HTML5-compliant web recoloring approach for dichromat people that considers both text and image recoloring in an integrated manner.

In the proposal entitled "Validity and Intra Rater Reliability of a New Device for Tongue Force Measurement", authors present a method to validate a new device proving that it is accurate compared to the algometer, more specific the study is oriented to determine the intra-rater reliability of a protocol to assess the maximum tongue force in asymptomatic subjects. For instance, authors propose a prototype device specifically for this study to measure tongue force through force-sensitive resistor sensors.

In the same field, in the article entitled "Deep Learning Assisted Medical Insurance Data Analytics With Multimedia System", authors present a convolution neural network-based deep learning infrastructure that performs medical imaging data analysis in various pipeline stages, including data-loading, data-augmentation, network architectures, loss functions, and evaluation metrics. The proposed deep learning approach supports both 2D as well as 3D medical image analysis. Finally, authors evaluate the proposed system's performance using metrics like sensitivity, specificity, accuracy, and precision over the clinical data with and without augmentation.

Changing to the area of optimization, in the work entitled "A Comparative Evaluation of Bayesian Networks Structure Learning Using Falcon Optimization Algorithm", authors present and evaluate a Bayesian network structure learning trough Falcon Optimization Algorithm (FOA) in order to suggest the best structural solution to create the FOA. The FOA algorithm is based on the falcon's searching technique during drought conditions. The suggested technique is compared to the score metric function of Pigeon Inspired search algorithm, Greedy Search, and Antlion optimization search algorithm. The performance of these techniques in terms of confusion matrices was further evaluated by the authors using a variety of benchmark data sets.

There are several proposals in optimization as the one in the paper entitled "Multi-Agent and Fuzzy Inference-Based Framework for Traffic Light Optimization", where the authors present a traffic simulation framework based on agent technology and fuzzy logic. The objective of this framework is to act on the phase layouts represented by its sequences and length to maximize throughput and fluidize traffic at an isolated intersection and for the whole multi-intersection network, through both inter- and intra-intersection collaboration and coordination. The system profits from agent communication and collaboration as well as coordination features, along with decentralized organization, to decompose the traffic control optimization into subproblems and enable the distributed resolution. Authors also use fuzzy technology to handle the uncertainty of traffic conditions.

DOI: 10.9781/ijimai.2023.05.005

In the field of image recognition, in the paper entitled "Digit Recognition Using Composite Features With Decision Tree Strategy", authors propose a method to identify all the characters of E13B using feature recognition in order to reduce the use of magnetic ink reader as specialized and expensive method in the baking industry. Hence authors found that the proposed method of recognition used, has relevant correlations to prove its validity and accuracy. The proposed method was also applied to an embedded device to ensure that the CPU would be used for verification instead of a high-end GPU.

For its part, in the paper entitled "Cosine Similarity Based Hierarchical Skeleton and Cross Indexing for Large Scale Image Retrieval Using Mapreduce Framework", the authors present a new method for image retrieval, named Cosine Similarity-based hierarchical skeleton and cross-indexing, that is proposed to perform the retrieval process in the MapReduce framework effectively. The feature vector of the images is converted to binary sequences. The Most Significant Bit (MSB) of the binary code is used to store the images in the mapper using the cross-indexing model.

Changing the area, in another article entitled "A Greedy Randomized Adaptive Search With Probabilistic Learning for Solving the Uncapacitated Plant Cycle Location Problem", the authors propose a mathematical formulation to model the Uncapacitated Plant Cycle Location Problem, specifically, trough the location-routing problem aimed at determining a subset of locations to set up plants dedicated to serving customers.

In the paper entitled "Resource and Process Management With a Decision Model Based on Fuzzy Logic", the authors propose a new aggregation operator in order to solve the problem of the allocation of the resources to be shared in the context of a distributed processing system that needs to be coordinated through the mutual exclusion mechanism.

In the field of cloud platforms, in the article entitled "A Hybrid Secure Cloud Platform Maintenance Based on Improved Attribute-Based Encryption Strategies", the authors introduce a hybrid data security scheme called the Improved Attribute-Based Encryption Scheme (IABES). This IABES combines two powerful data security algorithms: Advanced Encryption Standard (AES) and Attribute-Based Encryption (ABE) algorithm. These two algorithms are combined to provide massive support to the proposed approach of data maintenance over the remote cloud server with high-end security norms. This hybrid data security algorithm assures that the data cannot be attacked over the server by the attacker or intruder in any case because of its robustness.

For its part, in the paper entitled "Real World Anomalous Scene Detection and Classification Using Multilayer Deep Neural Networks", a novel methodology termed Bag of Focus (BoF) based training methodology has been proposed. BoF is based on the concept of selecting motion-intensive blocks in a long video, for training different deep neural networks (DNN's). The authors found that the methodology reduced the computational overhead by 90% (ten times) in comparison to when full-length videos are entertained.

In the field of Machine Learning Algorithms, in the paper entitled "RIADA: A Machine-Learning Based Infrastructure for Recognising the Emotions of *Spotify* Songs", authors present the RIADA infrastructure which is composed by a set of systems able to annotate emotionally the catalog of songs offered by Spotify based on the users' perception. RIADA works with the Spotify playlist miner and data services to build emotion recognition models through the use of Machine learning algorithms, music information retrieval techniques, architectures for parallelization of applications and cloud computing.

Based on a computational model, in the research entitled "Rhetorical Pattern Finding", authors research about rhetorical patterns from a musicological and computational standpoint. The above through a theoretical examination of what constitutes a rhetorical pattern is conducted. This examination includes primary sources and the study of the main composers, a formal definition of rhetorical patterns is proposed. Among the rhetorical figures, a set of imitative rhetorical figures is selected for the study, namely, epizeuxis, palilogy, synonymy, and polyptoton. Authors design a computational model of the selected rhetorical patterns to automatically find those patterns in a corpus consisting of masses by Renaissance composer Tomás Luis de Victoria.

Changing to the field of education, in the article entitled "Mapping the Situation of Educational Technologies in the Spanish University System Using Social Network Analysis and Visualization", the authors present two different maps based on the data from ICT Sectorial of CRUE Universidades Españolas. Together, they illustrate the penetration of different types of EdTech in Spain´s universities system and shed light on the strategic interest behind their adoption. The main goal of the authors for this study, is to produce self-explanatory maps that can be easily and directly interpreted.

In other side, in the field of recommender systems, in the paper "Local Model-Agnostic Explanations for Black-box Recommender Systems Using Interaction Graphs and Link Prediction Techniques", the authors propose a local model-agnostic, explanation-by-example method for recommender systems based on knowledge graphs to leverage this knowledge requirement. The system only requires information about the interactions between users and items. Through the proper transformation of these knowledge graphs into item-based and user-based structures, link prediction techniques are applied to find similarities between the nodes and to identify explanatory items for the user's recommendation.

In the field of the semantic web, in the article entitled "OntoInfoG++: A Knowledge Fusion Semantic Approach for Infographics Recommendation", the authors propose the OntoInfoG++, which is a knowledge centric recommendation approach for Infographics that encompasses the amalgamation of metadata derived from multiple heterogeneous sources and the crowd sourced ontologies to recommend infographics based on the topic of interest of the user. The approach models user topic of interest from the Query Words, Current User-Clicks, and from standard Knowledge Stores like the BibSonomy, DBpedia, Wikidata, LOD Cloud, and crowd sourced Ontologies.

Switching topics, in the field of the development of mobile applications, in the paper entitled "Adaptation of Applications to Compare Development Frameworks in Deep Learning for Decentralized Android Applications", the authors present the results of the analysis and a comparison of deep learning development frameworks, which can be adapted into fully decentralized Android apps from a cloud server.

In the field of User Experience (UX), in the paper entitled "On the Importance of UX Quality Aspects for Different Product Categories", the authors conceptualize UX as a set of semantically distinct quality aspects in order to present several studies that investigate this dependency between the product category and the importance of several well-known UX aspects.

Finally, in the same field of User Experience (UX), in the article entitled "Development of a Shared UX Vision Based on UX Factors Ascertained Through Attribution", the authors present an approach to developing a shared UX vision. This UX vision was developed by the product team while a collaborative session. The results show that the present approach for developing a UX vision helps to promote a shared understanding of the intended UX in a quickly and simply way.

Dr. Paulo Alonso Gaona-García
Universidad Distrital Francisco José de Caldas

TABLE OF CONTENTS

Exploring ChatGPT's Potential for Consultation, Recommendations and Report Diagnosis: Gastric Cancer and Gastroscopy Reports' Case

Jiaming Zhou[1], Tengyue Li[1], Simon James Fong[1,2]*, Nilanjan Dey[3], Rubén González Crespo[4]

[1] Department of Computer and Information Science, University of Macau, SAR, Macau (China)
[2] Chongqing Key Laboratory of Intelligent Perception and Blockchain Technology, Department of Artificial Intelligence, Chongqing Technology and Business University, Chongqing (China)
[3] Department of Computer Science and Engineering, Techno International New Town, Kolkata (India)
[4] Computer Science and Technology Department, Universidad Internacional de La Rioja, Logroño (Spain)

Received 20 March 2023 | Accepted 22 April 2023 | Published 25 April 2023

Abstract

Artificial intelligence (AI) has shown its effectiveness in helping clinical users meet evolving challenges. Recently, ChatGPT, a newly launched AI chatbot with exceptional text comprehension capabilities, has triggered a global wave of AI popularization and application in seeking answers through human–machine dialogues. Gastric cancer, as a globally prevalent disease, has a five-year survival rate of up to 90% when detected early and treated promptly. This research aims to explore ChatGPT's potential in disseminating gastric cancer knowledge, providing consultation recommendations, and interpreting endoscopy reports. Through experimentation, the GPT-4 model of ChatGPT achieved an appropriateness of 91.3% and a consistency of 95.7% in a gastric cancer knowledge test. Furthermore, GPT-4 has demonstrated considerable potential in consultation recommendations and endoscopy report analysis.

Keywords

Artificial Intelligence, ChatGPT, E-assessment, Gastric Cancer, Medicine.

DOI: 10.9781/ijimai.2023.04.007

I. Introduction

ARTIFICIAL Intelligence (AI) applications in the medical field are rapidly expanding, with computer vision and natural language processing technologies playing a crucial role in achieving more efficient and accurate diagnosis and treatment solutions than ever. Computer vision technology, particularly deep learning algorithms, has made profound achievements in medical image analysis. For instance, convolutional neural networks (CNNs) are widely used to detect and identify various diseases, such as cancer [1], diabetic retinopathy [2], pneumonia [3], oral anomalies [4] and brain tumor [5]. These advanced technologies assist physicians in analysing imaging data more accurately, thus enhancing diagnostic accuracy and reducing misdiagnoses.

In the field of natural language processing (NLP), AI technologies have made breakthroughs in medical record text mining, disease risk prediction, and clinical decision support [6]. For example, NLP technologies can analyse patients' electronic medical records, helping doctors quickly access patients' medical history, past diagnoses, and treatment plans, thereby improving the efficiency of diagnosis and treatment. As AI technology continues to advance, the application prospects of computer vision and natural language processing technologies in the medical field will become increasingly extensive, contributing to various domains of healthcare applications.

* Corresponding author.
E-mail address: ccfong@umac.mo

A. Gastric Cancer

Gastric cancer is a globally prevalent disease. According to data from the World Health Organization, gastric cancer is the fifth most common type of cancer globally and the fourth leading cause of cancer-related deaths. In 2020, there were over 1 million new cases of gastric cancer worldwide, and approximately 769,000 people died from the disease [7]. As the symptoms of gastric cancer are often not apparent in the early stages, many patients are only diagnosed at late stages with poor treatment outcomes and prognosis. This makes early diagnosis and screening crucial for survival. Furthermore, the causes of gastric cancer are complex and include genetic factors, environmental factors, lifestyle factors such as poor dietary habits and smoking, and Helicobacter pylori (H. pylori) infection [8]-[12].

B. ChatGPT

ChatGPT is a large-scale natural language processing model developed by OpenAI. As a general-purpose deep learning language model, it has extensive applications in all walks of life. By learning from a vast amount of text data, ChatGPT can understand, generate, and respond to a wide range of natural language questions, providing assistance and suggestions to users [13]-[18]. It demonstrates outstanding performance in many aspects, including question-responding systems, article generation, summarization, translation, and other natural language processing tasks. During the development of GPT-4, Open-AI continuously optimized and improved the model's performance to meet the needs of more users. From GPT-3.5 to GPT-4, significant progress has been made, including higher appropriateness, consistency, and error-correction capabilities.

C. Purpose of the Study

If gastric cancer is detected and treated early, the patient's five-year survival rate could be as good as 90% [19]. Unfortunately, gastric cancer is often diagnosed at late stages. This lateness of diagnosis is due to the similarity between early-stage gastric cancer symptoms and those of common gastric diseases, as well as the fact that most people are not well informed about the risk factors for gastric cancer. Therefore, they rarely proactively undergo proper examinations when experiencing gastric discomfort. The advent of ChatGPT has transformed many people's work and lifestyles, promoting a global wave of AI popularization and application. This article reports insights into exploring ChatGPT's potential in disseminating gastric cancer knowledge, providing consultation recommendations, and endoscopy report analysis, ultimately benefiting both patients and doctors.

II. Materials and Methods

A. Materials

For the medical knowledge test on gastric cancer, we tested twenty-three questions on five aspects, including concept, prevention, screening, treatment, and complications. For the consultation recommendations and endoscopy report analysis, our case materials were obtained from the Chinese Medical Case Repository, Journal of Medical Case Reports, and F1000 Research [20]-[26].

B. Methods

The tests were executed on GPT-3.5 and GPT-4, generating three responses per question. The responses were evaluated in two aspects:

- Appropriateness: If one of the responses is wrong, including obvious false statements, inaccurate diagnoses, or inappropriate suggestions, the response will be judged as inappropriate.
- Consistency: The consistency between the three responses is judged at an approximate level; there are two situations that are considered consistent in the experiments: being all correct for the same idea or making the same mistake for the same idea; if they are inconsistent, they will be considered unreliable.

There are four situations shown in Table I and Fig.1. Thumbs-up means the response is right and thumbs-down means the response is wrong. The same color means that all three responses are for the same idea and are consistent.

TABLE I. Evaluation of Appropriateness and Consistency

	Appropriate	Inappropriate
Consistent	All thumbs-up All in same color	All thumbs-down All in same color
Inconsistent	All thumbs-up Have different colors	Not all thumbs-up Have different colors

If inappropriate responses are generated, we tested the error-correction capability of ChatGPT. We classify its error-correction capabilities into two categories: automatic error-correction and guided error-correction.

- Automatic error correction refers to ChatGPT's ability to detect and correct errors upon its first attempt after receiving a prompt.
- Guided error correction refers to ChatGPT's capacity to identify errors after a user narrows down the paragraph range, even if it did not initially detect the error.

In consultation recommendations and endoscopy report analysis tests, with the support of real pathological diagnostic results, we focus more on whether ChatGPT can detect abnormalities and the consistency of the three responses.

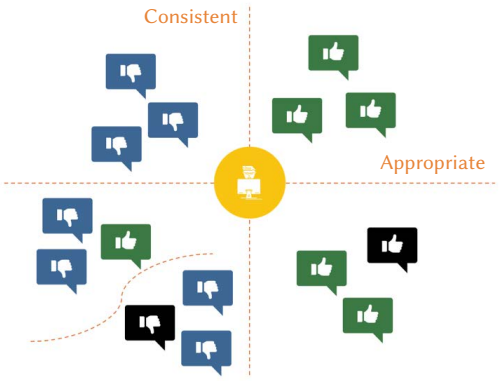


Fig. 1. Evaluation of appropriateness and consistency.

III. Experiment

In this section, we will sequentially introduce the experiments, including medical knowledge, consultation recommendations, and endoscopy report analysis tests. The general flow of the experiments is shown in Fig. 2.

A. Medical Knowledge Test

In this knowledge test about gastric cancer, we tested each question separately with GPT-3.5 and GPT-4. Three responses were generated for each question, and we evaluated the appropriateness and consistency. Table II shows the 23 fundamental questions about gastric cancer that we used for testing.

TABLE II. Evaluation of ChatGPT for Questions About Gastric Cancer

Question	GPT-4[a]	GPT-3.5[a]
1. What is gastric(stomach) cancer?	C/A	C/A
2. What are the stages of gastric(stomach) cancer?	C/NA	NC/NA
3. According to TNM stage, which stage does early gastric(stomach) cancer belong to?	C/A	NC/NA
4. How to prevent gastric(stomach) cancer?	C/A	C/A
5. How to prevent H.pylori infection?	C/A	NC/NA
6. What are the symptoms of H. pylori infection?	C/A	C/A
7. Who should seek diagnosis and treatment of an H. pylori infection?	C/A	C/A
8. Is my risk for gastric(stomach) cancer higher if my family member has it?	C/A	C/A
9. Can gastric ulcers develop into gastric(stomach) cancer?	C/A	C/A
10. What is gastric(stomach) cancer screening?	C/A	C/A
11. Are there risks associated with gastric(stomach) cancer screening?	C/A	C/A
12. How do I know if I need to be screened for gastric(stomach) cancer?	C/A	C/A
13. Is there anything I should prepare for before the gastric(stomach) cancer screening?	C/A	NC/NA
14. What are the symptoms of gastric(stomach) cancer?	C/A	C/A
15. What is cagA-positive H. pylori and how does it affect the risk of gastric(stomach) cancers?	C/A	C/A
16. Does gastroscopy require anaesthesia?	C/A	C/A
17. Is it possible for children to get gastric(stomach) cancer?	C/A	C/A
18. How to treat gastric(stomach) cancer?	NC/NA	C/NA
19. How to treat stage 0 gastric(stomach) cancer?	C/A	C/A
20. How to treat early gastric(stomach) cancer?	C/A	C/NA
21. How to treat advanced gastric(stomach) cancer?	C/A	C/A
22. What is HER2 testing?	C/A	C/A
23. Can a patient with gastric(stomach) cancer get pregnant?	C/A	C/A

[a] where C for consistent, A for appropriate, NC for inconsistent, and NA for inappropriate

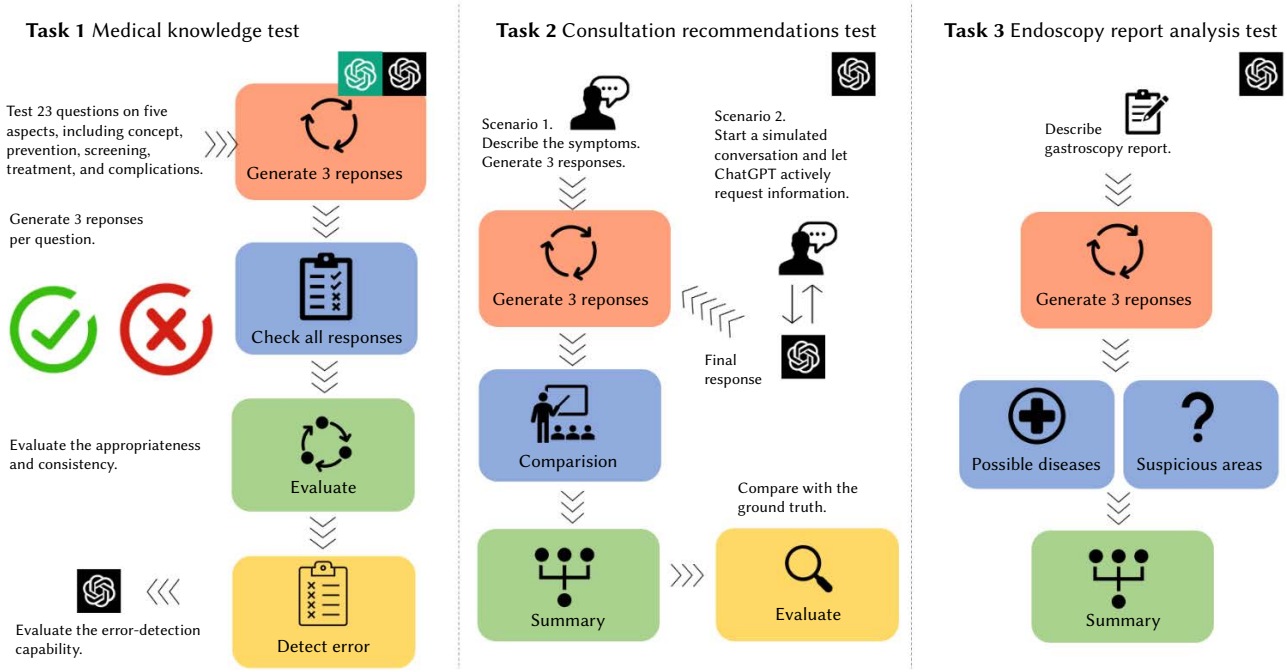


Fig. 2. General flow of the experiments.

Human: ***Preventing Helicobacter pylori (H. pylori) infection can be challenging, as it is often spread through close contact with an infected person, and can also be acquired through contaminated food or water. However, there are several steps you can take to reduce your risk of H. pylori infection: Practice good hygiene: Wash your hands frequently with soap and water, especially before eating or preparing food, and after using the bathroom or changing a diaper. Avoid contaminated food and water: Drink clean, safe water, and avoid eating food that may be contaminated, such as undercooked meat or shellfish, unpasteurized milk or dairy products, or food that has been improperly stored or handled. Avoid sharing personal items: H. pylori can be spread through saliva, so avoid sharing utensüs, cups, or other personal items with someone who may be infected. Get vaccinated: Although there is no vaccine specifically for H. pylori, vaccination against other diseases such as hepatitis A and B, and human papillomavirus (HPV) can help reduce your risk of infection and may indirectly reduce the risk of H. pylori infection. Avoid smoking: Smoking has been linked toan increased risk of H. pylori infection, so quitting smoking or avoiding exposure to secondhand smoke may help reduce your risk. If you suspect you may have been exposed to H. pylori or are experiencing symptoms such as abdominal pain, nausea, or indigestion, talk to your doctor about testing and treatment options. Early detection and treatment of H. pylori infection can help prevent complications and reduce the risk of gastric cancer. Is it correct?***

GPT-4: *Yes, the information provided in the paragraph is correct. Jt outlines various steps individuals can take to reduce their risk of H. pylori infection, including practicing good hygiene, avoiding contaminated food and water, not sharing personal items, getting vaccinated far related diseases, and avoiding smoking. Additionally, it encourages people to consult a doctor if they suspect exposure to H. pylori or experience related symptoms*

Human: ***Get vaccinated: Although there is no vaccine specijically for H. pylori, vaccination against other diseases such as hepatitis A and B, and human papillomavirus (HPV) can help reduce your risk of infection and may indirectly reduce the risk of H. pylori infection.***
Is it correct?

GPT-4: *I apologize far the confusion in my previous response. The statement about vaccination is not correct. There is currently no vaccine far H. pylori, and vaccination against other diseases like hepatitis A and B, and human papillomavirus (HPV) will not directly or indirectly reduce the risk of H. pylori infection. These vaccinations are important far preventing the diseases they target, but they have no effect on H. pylori infection risk.*

Human: ***The stages of gastric (stomach) cancerare determined based on the size and extent of the tumor, as well as the spread of cancer to nearby lymph nodes or distant organs. The stages are as follows:***
Stage 0: Also known as carcinoma in situ, the cancer cells are only in the inner lining of the stomach and have not spread to other tissues.
Stage I: The cancer has grown into the deeper layers of the stomach wall but has not spread to the lymph nodes or other organs.
Stage II: The cancer has grown through the full thickness of the stomach wall and may have spread to nearby lymph nodes.
Stage III: The cancer has spread to nearby lymph nodes and may have invaded nearby organs such as the spleen or pancreas.
Stage IV: The cancer has spread to distant organs such as the liver, lungs, or bones.
Doctors use a variety of tests to determine the stage of gastric cancer, including imaging tests like CT scans, MRI scans, and PET scans, as well as endoscopic biopsies and blood tests. Treatment options for gastric cancer depend on the stage and location of the cancer, but may include surgery, chemotherapy, radiation therapy, targeted therapy, and immunotherapy.
Is it correct?

GPT-4: *Yes, the information provided in the paragraph is correct. It accurately describes the different stages of gastric (stomach) cancer, ranging from Stage 0 (carcinoma in situ) to Stage IV (cancer that has spread to distant organs). The paragraph also lists various tests that doctors use to determine the stage of gastric cancer and mentions possible treatment options based on the stage and location of the cancer.*

Human: ***Stage I: The cancer has grown into the deeper layers of the stomach wa/1 but has not spread to the lymph nodes or other organs.***
Is it correct?

GPT-4: *The statement is not entirely accurate. Stage I gastric cancer is farther divided into two sub-stages, which consider the involvement of lymph nodes:*
Stage IA: The cancer has grown into the deeper layers of the stomach wall (either the submucosa or the muscle layer) but has not spread to the lymph nodes or other organs.
Stage IB: The cancer has either grown into the muscle layer of the stomach wall and spread to 1 or 2 lymph nodes, or it has grown through the muscle layer but has not spread to any lymph nodes or other organs.
Thus, the corree! description of Stage I gastric cancer should account far these sub-stages and their relation to lymph node involvement.

Fig. 3. Example of guided error correction.

TABLE III. Summary of Consultation Recommendations

Case ID	Pathological confirmation	Symptom Keywords	Inference results/Possible diseases[a]
1	Gastric adenocarcinoma & Pernicious anemia	1. 61-year-old female 2. intermittent social alcohol consumption 3. epigastric pain 4. heartburn 5. decreased appetite 6. weakness 7. unintentional weight loss	1. Gastritis (3) 2. Peptic ulcer disease (3) 3. Gastroesophageal reflux disease (GERD) (2) 4. Gastrointestinal malignancy (3)
2	Gastric adenocarcinoma of the fundic gland type, GA-FG	1. 73-year-old male 2. discomfort in the upper and middle abdomen 3. more than 1 month 4. H. pylori history	1. Recurrent H. pylori infection (2) 2. Gastritis (3) 3. Peptic ulcer disease (3) 4. Functional dyspepsia (3) 5. Gastroesophageal reflux disease (GERD) (3)
3	Gastric adenocarcinoma of the fundic gland type, GA-FG	1. 52-year-old female 2. vague stomach pains 3. pains worse when hungry 4. pains be relieved after eating	1. Peptic ulcer disease (3) 2. Gastritis (3) 3. Functional dyspepsia (3) 4. Gastroesophageal reflux disease (GERD) (3)
4	Gastric cancer combined with duodenal cancer	1. 81-year-old male 2. abdominal pain 3. black stools, unformed 4. persistent epigastric pain 5. abdominal distension unrelated to food intake 6. occasional nausea and vomiting 7. weight loss	1. Upper gastrointestinal bleeding (3) 2. Gastrointestinal bleeding (3) 3. Gastrointestinal malignancy (3) 4. Peptic ulcer disease (3) 5. Gastritis (2) 6. Gastric outlet obstruction (2)

[a] where the number in (*) after each possible disease indicates the number of occurrences in the three responses generated by ChatGPT.

After the experiment, the results are recorded and compared to the ground truth from the golden standards. For GPT-4, 91.3% of the responses are appropriate with a consistency of 95.7%; for GPT-3.5, the accuracy is not good, and only 73.9% of the responses are appropriate with a consistency of 82.6%. GPT-4 has a more powerful reasoning ability, but that comes with a paid subscription fee, it does not respond as fast as GPT-3.5. Prior to this, we also found that GPT-4 has higher completeness and expansiveness in its responses.

Next, we will focus only on GPT-4 because the tolerance for errors is relatively low in medical scenarios. Although GPT-3.5 is fast in response, it is unsuitable for application in medical scenarios. Based on our experimental results, a more rigorous statement would be that GPT-3.5 may be less suitable for this specific scenario of gastric cancer consultation.

In the responses obtained from the experiment, some were considered inappropriate. We used these data to examine GPT-4's error correction ability (shown in Fig. 3). The prompt we used in the error correction experiment is:

[*Paragraphs to be checked*]

Is it correct?

Unfortunately, under this prompt, GPT-4 does not perform well in error correction. When the prompt was first sent, GPT-4 did not immediately detect the paragraph error, but once the scope was artificially narrowed, GPT-4 still had the ability to correct the error (Fig. 3). Despite this, it shows that ChatGPT has the potential for error correction.

B. Consultation and Diagnosis Recommendations

In this experiment, we tested the feasibility of GPT-4 acting as if it is an "internet doctor." Again, we had ChatGPT generating three responses for each question. Table III contains some key points derived from the patients' chief complaints and ChatGPT's responses. The prompt we used is:

[*Paragraph*]

You are a specialist gastroenterologist,

please make a diagnosis of this case description.

The table presents four cases of patients with gastric cancer; however, only cases 1 and 4 were suggested as potentially having gastrointestinal malignancy. All three responses alluded to this information, emphasizing its significance. These two cases provided comprehensive information, while the other two cases did not indicate a risk of gastric cancer for two reasons: the information provided was insufficient, and there was a lack of critical information. It is important to note that many gastric diseases share similar symptoms, and complications may coexist, making diagnosis difficult with limited information. Consequently, GPT-4 listed several possible diseases for each case.

To address the issue of insufficient information, there are two possible scenarios. The first is human-driven, wherein users actively provide adequate information. This approach is crucial, as ChatGPT may not necessarily seek additional information but directly guess a diagnosis, which is beyond the user's control and may leave patients uninformed due to a lack of medical knowledge. The second scenario involves having ChatGPT proactively seeking information, allowing it to determine the necessary details (Fig. 4). It is crucial to emphasize the importance of providing sufficient information for accurate diagnosis, to consider the potential benefits of ChatGPT, and to actively seek additional information to improve diagnostic accuracy. For example:

Let's have a conversation. Suppose you are a professional gastroenterologist, and I am a patient who comes to consult you, [*What specifically do you want to consult*] *please try to ask for more information from me to analyze my disease and give me suggestion.*

In the realm of gastric cancer and other gastric diseases, GPT-4 consistently emphasizes the necessity for patients to undergo a comprehensive evaluation by a healthcare professional. This evaluation should include appropriate laboratory tests, imaging studies, and potentially an endoscopic examination to reach a definitive diagnosis and initiate suitable treatment. It is important to recognize that GPT-4 can only indicate potential risks, and a definitive diagnosis can only be made following a thorough examination. Additionally, GPT-4 can offer guidance on diagnostic procedures.

Clinician: ***Let's have a conversation. Suppose you are a professional gastroenterologist, and I am a patient who comes to consult you, please try to ask for more information from me to analyze my disease and give me suggestions.***	**Clinician:** ***Let's have a conversation. Suppose you are a professional gastroenterologist, and I am a patient who comes to consult you. I want to consult if I have gastric cancer, please try to ask for more information from me to analyze my disease and give me suggestions.***
GPT-4: *Hello, I'm Dr. ChatGPT, a gastroenterologist. I'm here tonhelp you with your gastrointestinal concerns. To better understand your situation, 1 wi/1 need to ask you a few questions.* *1. Can you please describe your main symptoms, including their severity, frequency, and duration?* *2. Have you noticed any specific triggers for these symptoms, such as certain foods or activities?* *3. Have you experienced any recent changes in your bowel habits (e.g., constipation, diarrhea, or alternating between the two)?* *4. Are you experiencing any pain or discomfort in your abdomen? If so, can you please describe the location, intensity, and any factors that seem to relieve or worsen the pain?* *5. Have you noticed any blood in your stool, or black, tarry stools?* *6. Do you have a history of gastrointestinal issues, either personally or in your family?* *7. Are you currently taking any medications or supplements?* *8. Have you recently los/ or gained weight unintentionally?* *Please provide as much detai/ as possible, as this wi/1 help me better understand your situation and provide appropriate suggestions.*	**GPT-4:** *Hello, I'm glad you've come to consult with me. As a gastroenterologist, I'm here to help you with any concerns you may have about your gastrointestinal health. To better understand your situation and give you the most accurate advice possible, I'll need to ask you sorne questions. Please remember that I am an AI language model and not a real doctor, but I will try my best to provide helpful information based on your responses.* *1. Can you please tell me your age, gender, and any relevant medical history, including any family history of gastric cancer or other gastrointestinal diseases?* *2. What specific symptoms have you been experiencing and for how long?* *3. Have you noticed any changes in your appetite, weight, or bowel habits recently?* *4. Do you have any history of gastrointestinal problems, such as ulcers, GERD, or gastritis?* *5. Are you currently taking any medications, or have you taken any medications recently, particularly nonsteroidal anti-inflammatory drugs (NSAIDs) or antacids?* *6. Do you smoke or consume alcohol? If so, how much and how often?* *Please provide as much informat ion as you can so I can better understand your situation and provide appropriate suggestions.*

Fig. 4. Example of ChatGPT requesting information.

C. Endoscopy Report Analysis

The primary objective of this experiment was to assess GPT-4's supplementary role in endoscopic examination reports. By describing the observations made during endoscopy, GPT-4 can assist in analysing suspicious regions and reconfirming biopsy sites. It is crucial to note that the focus of this experiment is not on the accuracy of GPT-4's analysis of medical mechanisms but rather on its ability to detect abnormalities. As in previous experiments, three responses were generated for each query, and if the content of the responses was not entirely consistent, they could all be used as reference points. The prompt used in this experiment is as follows:

Gastroscopy report: [*Description*]

With this description,please give me some diagnostic hints.

Fig. 5 summarizes the results of GPT-4's endoscopy report analysis. Through experiments, we find that GPT-4 does have the ability to analyse endoscopy reports.

IV. Discussion

The launch of ChatGPT has changed the way many people work and live. As a general deep learning language model, it has applications in many areas. In this experiment, we primarily focus on the medical field, using gastric cancer as a specific scenario to test ChatGPT's application. Our tests revolve around gastric cancer medical knowledge test, consultation recommendations, and endoscopy report analysis.

First, in the medical knowledge test, we compared the performance of GPT-4 and GPT-3.5 by testing both models with the same questions, generating three responses per question. GPT-4's appropriateness reached 91.3%, with a consistency of 95.7%. As a general model, it is already quite powerful without having been specifically trained on a particular type of data. However, it is important to note that we should not overly rely on ChatGPT for clinical diagnosis; as it has stated itself, it is not a doctor who can provide sound advice. According to the experiment, ChatGPT is prone to making mistakes in certain details or expressions, such as the specific staging of cancer, and its responses are not very satisfying in regard to detailed classifications. In the error correction test, GPT-4 also overlooked many details and did not detect errors as expected, only considering them after narrowing down the scope. Of course, this could also be an issue with the prompt. After ChatGPT's popularity, prompt engineering has become an important technology. One thing to note in the medical knowledge test session is that GPT-4 has tried to provide more comprehensive responses, including expanding on related but unasked content. However, by generating three responses to the same question, it sometimes still misses some key points. Therefore, one suggestion when using ChatGPT is to generate multiple responses to consider.

In the consultation recommendations test, with rich case information, we observed GPT-4's excellent diagnostic assistance capabilities. However, due to the complex gastric environment in our body, numerous complications, and similar symptoms across various gastric diseases, GPT-4's diagnostic suggestions are within a relatively broad range. This could serve as preliminary information before seeking medical attention, but a hospital examination is still necessary and very advisable for a definite diagnosis.

In the endoscopy report analysis test, GPT-4 also demonstrated outstanding reasoning abilities, identifying possible diseases based on the text description of the report, extracting information on suspicious areas, and suggesting biopsy and pathological examinations. However, there are limitations to this, as in many cases, biopsies are directly taken during endoscopic examinations before the report is available, making the diagnostic assistance a secondary confirmation. Nevertheless, it can still provide diagnostic suggestions in ordinary gastric disease endoscopic diagnoses.

In summary, through the experiment, we discovered more things that ChatGPT can do than expected. We should learn to utilize it without relying on it and always remember that it is a chatbot, not a person. Although ChatGPT has a wide range of application potential, it is important to note that as an AI model, it is not flawless [27]. In some cases, it may not provide accurate responses or suggestions. Therefore, users should exercise caution when using ChatGPT and seek professional advice from certified experts when necessary.

V. Future Perspectives

At present, most of the application of artificial intelligence in gastric cancer is primarily concentrated on the field of computer vision, such as autorecognition of gastric lesions under digital endoscopy. This task of object detection aims to identify suspicious targets from endoscopic images and classify gastric lesion types. Although the experiments in this study focus on the field of NLP, we can gain valuable insights from our endoscopic report analysis experiment. With the advancement of multimodal research, we believe that natural language processing and computer vision can be more deeply and effectively integrated in the future, such as through some automatic co-analysis and generation of unanimous reports when suspicious areas are detected during endoscopy inspection. The results of object detection and language reasoning can also be mutually verified, which would collaboratively produce a significant contribution to medical diagnosis.

Case 1 Description:
Gastroscopy report: (1) The gastric mucosa is smooth and free of erosions and ulcers. Local patchy redness is observed in the gastric antrum. The mucosa from the gastric antrum to the gastric angle is thinned, with visible blood vessels, and atrophy extends beyond the cardia. A yellowish nodular lesion is observed near the cardia at the gastric fundus. The comprehensive evaluation suggests a presentation a/ter Helicobacter pylori (HP) eradication. (2) A discolored, slightly yellowish lesion is observed on the anterior wall of the upper gastric body near the greater curvature, measuring approximately 4 mm x 5 mm. Upon closer inspection with a regular endoscope, the lesion appears slightly elevated with a yellowish hue, with visible dilated blood vessels and black dot-like changes on the surface. Under narrow-band imaging (NBI), there is a color dijference boundary between the lesion and the surrounding normal mucosa. Under narrow-band imaging-magnifying endoscopy (NBI-ME), there is no distinct boundary between the lesion and the surrounding normal mucosa. The microsurface structure of the lesion is regular/y arranged, with slightly widened crypt intervals. No significant irregularities are observed in the microvascular structure. After acetic acid staining, the whitening time of the lesion is shortened. The les ion becomes more prominent after indigo carmine staining.

Inference results (Possible diseases):
1. Atrophic gastritis (3)
2. Macular tumor near the cardia -> suggest biopsy and histopathological examination (3)
3. Discolored and slightly yellowish lesion -> suggest biopsy and histopathological examination (3)

Case 2 Description:
Gastroscopy report: A 1.0 cm diameter region with slightly red coloration is observed near the cardia on the lesser curvature ofthe stomach, with a slightly depressed center anda small amount of thin white coating on the surface. The lesion boundary is clear under narrow-band imaging magnijication endoscopy (NBI-ME), with thickened, tortuous microvessels forming a mesh-like pattern, and partial disappearance of microstructures. In the gastric angle, there is rough mucosa with nodular changes, measuring approximately 4.0 cm x 5.0 cm. NBl-ME reveals mu/tiple foci of dilated,
thickened, and irregular microvessels in the gastric angle mucosa, with partial fusion of microstructures. On the anterior wall of the gastric sinus near the greater curvature, there is a 0.8 cm diameter area of rough mucosa. The lesion boundary is clear under NBI-ME, with fused glandular ducts, partial disappearance, and malformed microvessels. On the greater curvature side of the gastric sinus and body junction, there is a 1.0 cm diameter area with a reddish center. NBI-ME reveals a brownish coloration of the lesion, with clear boundaries, partial fusion, and even absence of glandular ducts, as well as thickened, malformed microvessels. A 1.0 cm diameter area of slightly raised mucosa is observed on the anterior wall ofthe gastric sinus. NBI-ME reveals a small amount offused glandular ducts and partially linear, dilated, and deformed microvessels.

Inference results (Possible diseases):
l. Gastric lesion with slight depression -> suggest biopsy and histopathological examination (3)
2. Rough and nodular mucosa ofthe gastric hom -> suggest biopsy and histopathological examination (3)
3. Gastric sinus lesions -> suggest biopsy and histopathological examination (3)

Case 3 Description:
Gastroscopy report:
Esophagus: The mucosal surface is smooth and moist, with normal vascular patterns.
Cardia: Good closure, with smooth mucosa and no folds.
Gastric fundus: Mucus pool with bile-like fluid present. Smooth mucosa and clear, visible blood vessels. Large folds resembling gyri.
Gastric body: Smooth mucosa, with submucosal blood vessels visible. Regular arrangement of large folds along the greater curvature: normal peristaltic waves can be seen.
Gastric antrum: Intact morphology, arch-shaped, with smooth and orderly angular notches.
Pylorus: Alternating opening and closing; when open, it appears round with smooth and even edges.
Duodenal bulb: Spherical, with smooth mucosa without folds, presenting a velvety appearance.
Descending duodenum: Circular folds are clearly visible, and the papilla is normal.

Inference results (Possible diseases):
1. The presence of bile-like fluid in the mucus lake may suggest sorne reflux of bite from the duodenum into the stomach. (3)

Fig. 5. Summary of gastroscopy report analysis.

Acknowledgement

The authors are thankful for the financial support from research grants by Guangzhou Development Zone Science and Technology (2021GH10, 2020GH10), research grant by the University of Macau (MYRG2022-00271-FST) and research grant by the Science and Technology Development Fund of Macau (0032/2022/A).

References

[1] Y. Horie et al., "Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks," *Gastrointestinal Endoscopy*, vol. 89, no. 1, pp. 25–32, 2019, doi: 10.1016/j.gie.2018.07.037.

[2] A. S. Kumarasuvamy and R. S. Rajendran, "Design an Early Detection and Classification for Diabetic Retinopathy by Deep Feature Extraction based Convolution Neural Network," *Journal of Trends in Computer Science and Smart Technology*, vol. 3, no. 2, pp. 81–94, 2021, doi: 10.36548/jtcsst.2021.2.002.

[3] T. Rahman et al., "Transfer Learning with Deep Convolutional Neural Network (CNN) for Pneumonia Detection Using Chest X-ray," *Applied Sciences*, vol. 10, no. 9, Art. no. 9, 2020, doi: 10.3390/app10093233.

[4] A. Laishram and K. Thongam, "Automatic Classification of Oral Pathologies Using Orthopantomogram Radiography Images Based on Convolutional Neural Network," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 4, pp. 69-77, 2022, doi: 10.9781/ijimai.2021.10.009.

[5] M. A. Khemchandani, S. M. Jadhav, and B. R. Iyer, "Brain Tumor Segmentation and Identification Using Particle Imperialist Deep Convolutional Neural Network in MRI Images," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 7, pp. 38-47, 2022, doi: 10.9781/ijimai.2022.10.006.

[6] B. G. Patra et al., "Extracting social determinants of health from electronic health records using natural language processing: a systematic review," *Journal of the American Medical Informatics Association*, vol. 28, no. 12, pp. 2716–2727, 2021, doi: 10.1093/jamia/ocab170.

[7] H. Sung et al., "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021, doi: 10.3322/caac.21660.

[8] E. C. Smyth, M. Nilsson, H. I. Grabsch, N. C. van Grieken, and F. Lordick, "Gastric cancer," *The Lancet*, vol. 396, no. 10251, pp. 635–648, 2020, doi: 10.1016/S0140-6736(20)31288-5.

[9] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2022," *CA A Cancer J Clinicians*, vol. 72, no. 1, pp. 7–33, 2022, doi: 10.3322/caac.21708.

[10] K. Yashima, M. Shabana, H. Kurumi, K. Kawaguchi, and H. Isomoto, "Gastric Cancer Screening in Japan: A Narrative Review," *Journal of Clinical Medicine*, vol. 11, no. 15, Art. no. 15, 2022, doi: 10.3390/jcm11154337.

[11] A. Jaroenlapnopparat, K. Bhatia, and S. Coban, "Inflammation and Gastric Cancer," *Diseases*, vol. 10, no. 3, Art. no. 3, 2022, doi: 10.3390/diseases10030035.

[12] S. Kikuchi, Y. Obata, T. Sasakabe, S. Kawai, C. Wang, and Y. Lin, "Relative risk of gastric cancer between those with and without Helicobacter pylori infection history in Japan," *JGH Open*, vol. 6, no. 3, pp. 166–170, 2022, doi: 10.1002/jgh3.12714.

[13] H. L. Haver, E. B. Ambinder, M. Bahl, E. T. Oluyemi, J. Jeudy, and P. H. Yi, "Appropriateness of Breast Cancer Prevention and Screening Recommendations Provided by ChatGPT," *Radiology*, p. 230424, 2023, doi: 10.1148/radiol.230424.

[14] T. H. Kung et al., "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models," *PLOS Digital Health*, vol. 2, no. 2, p. e0000198, 2023, doi: 10.1371/journal.pdig.0000198.

[15] R. Vaishya, A. Misra, and A. Vaish, "ChatGPT: Is this version good for healthcare and research?," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 17, no. 4, p. 102744, 2023, doi: 10.1016/j.dsx.2023.102744.

[16] S. Biswas, "ChatGPT and the Future of Medical Writing," *Radiology*, vol. 307, no. 2, p. e223312, 2023, doi: 10.1148/radiol.223312.

[17] M. Sallam, "ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns," *Healthcare*, vol. 11, no. 6, Art. no. 6, 2023, doi: 10.3390/healthcare11060887.

[18] E. Kasneci et al., "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, p. 102274, 2023, doi: 10.1016/j.lindif.2023.102274.

[19] P.-H. Niu, L.-L. Zhao, H.-L. Wu, D.-B. Zhao, and Y.-T. Chen, "Artificial intelligence in gastric cancer: Application and future perspectives," *World J Gastroenterol*, vol. 26, no. 36, pp. 5408–5419, 2020, doi: 10.3748/wjg.v26.i36.5408.

[20] Z. X. Dong et al., "Endoscopic diagnosis and treatment for gastric adenocarcinoma of fundic gland type: report of 3 cases", *Chin J Dig Endosc*, vol. 39, no. 11, pp.931-934, 2022, doi: 10.3760/cma.j.cn321463-20210325-00206

[21] H. Y. Dong, X. Y. Jia, Q. P. Pang, G. D. Li, "A case of simultaneous five-origin early gastric cancer treated by gastroscopy", *Chin J Dig Endosc*, vol. 34, no. 12, pp. 913-914, 2017, doi: 10.3760/cma.j.issn.1007-5232.2017.12.018

[22] H. C. Huang et al., "[A case report of gastric adenocarcinoma of fundic gland type as suspected gastric neuroendocrine tumor]," *Zhonghua Nei Ke Za Zhi*, vol. 61, no. 6, pp. 685–687, 2022, doi: 10.3760/cma.j.cn112138-20210714-00482.

[23] S. Kamran, M. K. Dilling, N. A. Parker, J. Alderson, N. D. Tofteland, and Q. V. Truong, "Case Report: Simultaneously, diagnosed gastric adenocarcinoma and pernicious anemia – a classic association." *F1000Research*, 2020. doi: 10.12688/f1000research.24353.2.

[24] M. Moriya, A. Uehara, T. Okumura, M. Miyamoto, and Y. Kohgo, "Stress-induced hemorrhagic gastric ulcer after successful Helicobacter pylori eradication: two case reports," *Journal of Medical Case Reports*, vol. 5, no. 1, p. 252, 2011, doi: 10.1186/1752-1947-5-252.

[25] K. Barrett, M. W. Hii, and R. J. Cade, "Benign gastro-colic fistula in a woman presenting with weight loss and intermittent vomiting: a case report," *Journal of Medical Case Reports*, vol. 5, no. 1, p. 313, 2011, doi: 10.1186/1752-1947-5-313.

[26] I. B. Ismail, H. Zenaidi, R. Jouini, S. Rebii, and A. Zoghlami, "Case Report: Primary pure clear cell gastric carcinoma." *F1000Research*, 2020. doi: 10.12688/f1000research.25810.1.

[27] F. C. Kitamura, "ChatGPT Is Shaping the Future of Medical Writing However, Still Requires Human Judgment," *Radiology*, vol. 307, no. 2, p. e230171, 2023, doi: 10.1148/radiol.230171.

Jiaming Zhou

She received a B.Ec. degree in Economic Statistics from Guangdong University of Foreign Studies. She is currently pursuing an M.Sc. degree with the University of Macau, Macau China. She is a Research Assistant in Data Analytics and Collaborative Computing Laboratory (DACC). She has experience in data mining and data analysis. Her research interests are now focused on deep learning, computer vision algorithms, medical image processing, and other artificial intelligence applications in medicine.

Tengyue Li

Gloria Tengyue Li is currently a PhD student at the University of Macau. She is also the Head of Data Analytics and Collaborative Computing Laboratory, Zhuhai Institute of Advanced Technology, Chinese Academy of Science, Zhuhai, China. Ms Li is leading and managing the laboratory in R&D as well as technological transfer and incubation. She is an entrepreneur with experience in innovative I.T. contest, with her award-winning team in the Bank of China Million Dollar Cup competition. Her latest winning work includes the first unmanned supermarket in Macau enabled by the latest sensing technologies, face recognition and e-payment systems. She is also the founder of several Online2Offline dot.com companies in trading and retailing both online and offline. Ms Li is also an active researcher, manager and chief knowledge officer in the DACC laboratory at the Faculty of Science and Technology, University of Macau.

Simon James Fong

He graduated from La Trobe University, Australia, with a 1st Class Honours BEng. Computer Systems degree and a PhD. Computer Science degrees in 1993 and 1998, respectively. Simon is now working as an Associate Professor at the Computer and Information Science Department of the University of Macau, as an Adjunct Professor at Faculty of Informatics, Durban University of Technology, South Africa. He is a co-founder of the Data Analytics and Collaborative Computing Research Group in the Faculty of Science and Technology. Prior to his academic career, Simon took up various managerial and technical posts, such as systems engineer, IT consultant and e-commerce director in Australia and Asia. He published more than 200 papers in SCI journals. He serves on the editorial boards of the Journal of Network and Computer Applications of Elsevier, IEEE IT Professional Magazine, and various special issues of SCIE-indexed journals. Simon is also an active researcher with leading positions such as Vice-chair of IEEE Computational Intelligence Society (CIS) Task Force on "Business Intelligence and Knowledge Management", TC Chair of IEEE ComSoc e-Health SIG and Vice-director of International Consortium for Optimization and Modelling in Science and Industry (iCOMSI).

Nilanjan Dey

Nilanjan Dey is an Associate Professor in the Department of Computer Science and Engineering at Techno International New Town (Formerly known as Techno India College of Technology), Kolkata, India. He is a visiting fellow of the University of Reading, UK. He was an honorary Visiting Scientist at Global Biomedical Technologies Inc., CA, USA (2012-2015). He was awarded his PhD. from Jadavpur University in 2015. He is the Editor-in-Chief of International Journal of Ambient Computing and Intelligence, IGI Global. He is the Series Co-Editor of Springer Tracts in Nature-Inspired Computing, Springer Nature, Series Co-Editor of Advances in Ubiquitous Sensing Applications for Healthcare, Elsevier, Series Editor of Computational Intelligence in Engineering Problem Solving and Intelligent Signal processing and data analysis, CRC. His main research interests include Medical Imaging, Machine learning, Computer Aided Diagnosis, Data Mining, etc. He is a senior member of IEEE and fellow of IETE.

Rubén González Crespo

Dr. Rubén González Crespo has a PhD in Computer Science Engineering. Currently he is Vice Chancellor of Academic Affairs and Faculty from UNIR and Global Director of Engineering Schools from PROEDUCA Group. He is an advisory board member for the Ministry of Education at Colombia and an evaluator from the National Agency for Quality Evaluation and Accreditation of Spain (ANECA). He is a member from different committees at the ISO Organization. Finally, he has published more than 200 papers in indexed journals and congresses.

A Hybrid Parallel Classification Model for the Diagnosis of Chronic Kidney Disease

Vijendra Singh[1]*, Divya Jain[2]

[1] School of Computer Science, University of Petroleum and Energy Studies, Dehradun, 248007 (India)
[2] Computer Science and Engineering, The NorthCap University, Gurugram, 122017 (India)

Received 22 March 2021 | Accepted 1 September 2021 | Published 28 October 2021

Abstract

Chronic Kidney Disease (CKD) has become a prevalent disease nowadays, affecting people globally around the world. Accurate prediction of CKD progression over time is essential for reducing its associated mortality and morbidity rates. This paper proposes a fast, novel hybrid approach to diagnose Chronic Renal Disease. The proposed approach is based on the optimization of SVM classifier with the hybridized dimensionality reduction approach to identify the most informative parameters for CKD diagnosis. It handles the selection of features through two steps. The first one is a filter-based approach using ReliefF method to assign weights and ranks to each feature of the dataset. The second step is the dimensionality reduction of the best-selected subset by means of PCA, a feature extraction technique. For faster execution of datasets, simultaneous execution on multiple processors is employed. The proposed model achieved the highest prediction accuracy of 92.5% on the clinical CKD dataset compared to existing methods - 'CFS+SVM' (60.45%), 'ReliefF + SVM' (86%), 'MIFS + SVM' (56.72%), 'ReliefF + CFS + SVM' (54.37). The proposed work is also examined on the benchmarked Chronic Kidney Disease Dataset and achieved classification accuracy of 98.5% compared to the accuracy with other methods -'CFS+SVM' (92.7%), 'ReliefF + SVM' (89.6%), 'MIFS + SVM' (94.7%). The experimental outcomes positively demonstrate that the proposed hybridized model is effective in undertaking medical data classification tasks and is, therefore, a promising tool for the diagnosis of CKD patients. The proposed approach is statistically validated with the Friedman test with significant results compared to other techniques. The proposed approach also executes in the least time with improved prediction accuracy and competes with and even outperforms other methods in the literature.

Keywords

Chronic Kidney Disease Diagnosis, Clinical Dataset, Hybrid Approach, SVM Classifier, Dimensionality Reduction, Fast Execution.

DOI: 10.9781/ijimai.2021.10.008

I. Introduction

CHRONIC Kidney Disease (CKD), as known as Chronic Renal Failure has become a global health problem that results in high morbidity, mortality, and health care costs. It is a long-term condition that includes gradual loss of kidney function over time and can be caused by diabetes, high blood pressure, and other disorders [1]. Chronic Renal Failure leads to difficulties in removing extra fluids from the body and if this disease gets worse, wastes can build to high levels in the blood and may develop complications like high blood pressure, anemia, weak bones, and nerve damage [2]. So, damage to the kidneys and progression of this disease can potentially lead to renal failure. Often CKD is detected in individuals at later stages that are at high risk through advanced screening processes which require dialysis or a kidney transplant to sustain life [3]. Early diagnosis would facilitate in-time treatment, and is, therefore, essential to prevent complications. Data mining techniques can help in predicting the most significant risk factors related to CKD by using their medical history and plays a key role in the medical field [4].

* Corresponding author.
E-mail address: vsingh.fet@gmail.com

Moreover, the prevalence of CKD is rising in both developed and developing countries which is a matter of serious concern. At present, an estimated one in ten people is suffering from CKD worldwide [5]. Moreover, in the last decade, the US has seen a 30% increase in the prevalence of CKD [6]. The Global Burden of Disease (GBD) 2015 study by WHO estimated that approximately 5–10 million people die annually from kidney disease [7]. According to the National Chronic Kidney Disease Fact Sheet, 2017, 30 million adults in the US are estimated to have CKD [8]. The NHS Kidney Care examined the impact of CKD in England and estimated that approximately 1.8 million people are suffering from CKD and that around 40,000 to 45,000 premature deaths each year in people with CKD [9]. Similar statistics are found in America where 30 million adults are suffering from Chronic Renal Disease and millions of others are at significant risk of CKD [10]. In addition, the National kidney foundation [11], 10% of the population worldwide is affected by chronic kidney disease (CKD), and millions die each year due to restricted access to affordable treatment. Therefore, effective diagnosis and in-time treatment of patients are of prime significance. It is crucial to identify the presence of CKD in individuals at an early stage so that treatments that delay the progression of renal failure can be applied.

Diagnosis of CKD, which is dependent upon various symptoms, is a critical task in the medical field. It is an intricate process and

prone to false assumptions. When diagnosing diseases, the clinical decision is largely based upon the patient`s symptoms as well as on the knowledge and experience of the physicians [12]. Also, with the advancement in medical systems and the availability of new drugs, it becomes challenging for physicians and doctors to keep up-to-date with the latest developments in clinical practice [13]. Moreover, a computer-aided diagnostic system can assist even experienced physicians in taking medical correct decisions [14]. Thus, automating the diagnostic process by combining both machine learning techniques and physician`s experience is of large interest to medical professionals [15]. Machine learning and data mining techniques are playing substantial efforts to intelligently convert available data into useful information to increase the efficiency of the diagnostic process.

In the medical domain, classification techniques are typically useful for diagnostic problems that have been applied particularly in the area of disease diagnosis [16]-[18]. The classification system facilitates correct and in-time diagnosis of diseases which, in turn, enhances the success rate and reduces the decision-making time [19]. Support Vector Machine (SVM) Classifier, developed by Vapnik in 1995 [20], is a supervised machine learning technique that has been widely studied and implemented by researchers due to its outstanding generalization performance. SVM classifier has the potential for classifying large-scale datasets because it is robust and less sensitive to the curse of dimensionality. In addition, the appropriate setting of SVMs' parameters is extremely important to improve the classification accuracy [21], [22]. The optimization settings must ensure the accuracy of the SVM classifier but not increase the computational cost too significantly. The grid search technique, a widely used method for optimizing SVM classifier, helps in finding the best parameters to tune the performance of SVM [23]. With the optimization of parameters used in kernel functions, the SVM classification accuracy increases at a significant rate.

Furthermore, to design an efficient diagnostic system, the primary challenge lies in the identification of the most significant features from medical datasets. Feature Selection and feature extraction methods have been extensively used for medical diagnosis to tackle dimensionality reduction problems [24], [25]. They extract the most influential features and eliminate irrelevant ones from the data set, to reduce feature dimensionality and enhance the classification accuracy. Reducing the dimensionality of datasets helps in lowering the computational cost and improving the overall computational efficiency of the learning algorithms. Today hybridized dimensionality reduction methods with classification methods are being used by researchers that take advantage of two or more techniques that accelerate the removal of useless and extraneous features resulting in better diagnostic accuracy of the classifier [26]-[29].

The core objective of this long-term research work is to improve the diagnostics of CKD from a computational perspective. For the effective diagnosis of CKD, this work presents a fast classification technique based on the optimized SVM method with the inclusion of hybridized dimensionality reduction approach to identify the most informative parameters for CKD diagnosis, named RFP-SVM. As a first process in the proposed approach using the RFPCA method, the high dimensionality of the dataset is reduced using the hybridized technique based on ReliefF and PCA method. ReliefF method is applied in this research work as this method includes interaction among attributes and captures local dependency between features. This method is also robust to noisy and incomplete data and can deal with multiclass problems. PCA is used after the application of the ReliefF method as it forms uncorrelated variables that maximize the variability of the data. Furthermore, it reduces the dimensionality of data, while keeping as much variation as possible. For classification purposes, efficient optimization of SVM parameters is done using the grid-search method. SVM has good generalization capability with the ability to learn with very few samples. So, it is selected for the proposed technique. Hence, the proposed system is developed with the blending of dimensionality reduction techniques and optimized SVM results for the effectual and powerful classification of the CKD dataset. Thereafter, for faster execution, the proposed method is used with multiple processors which simultaneously process the CKD dataset using GPUs and machine workers.

The contribution of the research work can be stated three-fold: 1) first, high classification accuracy is achieved in the diagnosis of CKD for both clinical dataset and repository dataset. In addition, the model performs outstandingly well in terms of other evaluation measures such as precision, specificity, recall, and f-measure 2) second, the most significant risk factors related to CKD are identified and the least significant parameters are eliminated from the dataset using the proposed dimensionality reduction method 3) third, the diagnostic model executes in the least time as each task is executed simultaneously with multiple processors.

This research work is presented in the paper under the following sub-sections. Section II reviews previous research relevant to the CKD diagnosis and prediction of other chronic diseases using machine learning techniques. The next section presents the materials and approaches employed for the research. This also includes the description of the real-time clinical data and repository dataset considered for this work. The succeeding section presents a detailed discussion on the methodology of the proposed system with the design of the model used for the diagnosis of chronic kidney disease. The subsequent section illustrates the findings of the work and analysis of results using various performance evaluation measures. This is then followed by the benchmarking of the proposed model and statistical test used for the validation of the proposed technique. Finally, the closing remarks are provided in the discussion section followed by the conclusion.

II. Literature Survey

Machine Learning techniques have shown success in the prediction and diagnosis of numerous critical diseases. In recent years, early diagnosis of the disease, especially finding the best methods to apply medical treatments for CKD has received great attention among clinicians and researchers. Many recent studies have demonstrated the potential of using machine learning classification techniques to aid in the successful diagnosis of CKD.

In [30], the authors proposed an algorithm to diagnose CKD using classification algorithms. The authors compared the results of the proposed approach with different machine learning algorithms such as KNN, SVM, Naïve Bayes and showed the results in terms of accuracies of different classifiers.

In [31], authors applied six machine learning algorithms, namely: Random Forest (RF) classifiers, SMO, Logistic Regression, Radial Basis Function (RBF), Naïve Bayes, and Multilayer Perceptron Classifier (MLPC) to predict CKD and applied ten-fold cross-validation for validation of the dataset. The authors concluded that the Random Forest classifier outperforms other classifiers in terms of Area under the ROC curve (AUC), accuracy, and MCC metrics.

In [32], the authors applied the K-Means Clustering Algorithm with a single mean vector of centroids, to classify and make clusters of the varying probability of likeliness of suspect being prone to CKD. The methodology was demonstrated a dataset from UCI Machine Learning Repository.

In [33], three machine learning algorithms, namely: Logistic Regression, Radial Basis Function (RBF), and Multilayer Perceptron

Classifier (MLPC) were applied by the authors to predict CKD. The obtained results concluded that the Multilayer Perceptron Classifier outperforms other classifiers in terms of type I error, type II error, sensitivity, and accuracy.

In [34], the authors compared the performance of SVM and KNN classifier on CKD dataset and concluded that KNN outperforms SVM classifier in terms of accuracy, precision, recall, and f-measure metrics.

In [35], authors predicted CKD through two algorithms Naïve Bayes and Support Vector Machine, and concluded that the SVM classifier outperforms the Naïve Bayes classifier in terms of accuracy, precision, recall, and specificity. They also compared the execution time of both algorithms and the SVM classifier executes in less time compared to the Naïve Bayes algorithm.

Almansour et al. [36], for example, predicted CKD using two classification techniques that include SVM and ANN classifier. Before applying these algorithms, an appropriate setting is made to search for optimized parameter values. Subsequently, the classification models created from the two proposed techniques were developed using the best-obtained parameters and characteristics. The empirical results showed remarkable results with a predictive accuracy of 99.75% and 97.75% with the ANN and SVM classifiers, respectively. Sahu et al. [37] discovered the most significant parameters related to CKD with a genetic-search-based feature selection technique, named GSBFST. The authors employed various classifiers for evaluating the performance of the model. The proposed approach obtained better results in comparison to the existing algorithms. Akben [38] proposed a novel method for the early and automatic diagnosis of CKD. In the first phase, the pre-processing technique was applied to CKD data and in the second phase, three classification approaches (KNN, SVM, and Naïve Bayes) were applied to the resulting data to diagnose CKD. The results demonstrated a success rate of the proposed system with the highest diagnostic accuracy between 96% and 98% of the classifier. Misir et al. [39] presented an approach to predict CKD using correlation feature selection with classification approach and achieved good results in terms of classification accuracy, sensitivity and specificity, and AUC analysis. The proposed approach produced a reduced set of features and identified eight significant risk factors related to CKD. Norouzi et al. [40] predicted the renal failure timeframe of CKD using an adaptive neuro-fuzzy inference system. The authors used real clinical data using the ANFIS model to predict GFR values. The results concluded that the presented model accurately predicts the GFR variations for the prediction of renal failure. Serpen [41] diagnosed CKD using C4.5 decision trees, formulating a set of diagnostic rules to determine the highly significant risk factors related to the disease. Authors attained 98.25% accuracy using 3-fold cross-validation approach and identified primary and secondary indicators associated with the disease.

Machine learning models using appropriate feature selection and classification methods have been developed from time to time to support various medical decision-making tasks for the diagnosis of chronic diseases. The most recent significant work in this area has been done by Li et al. [42] in which they proposed a fast filter-based feature selection known as Coefficient of Variation to diagnose diabetes. This feature selection scheme discarded those attributes that degrade the performance of the model. The simulation experiments indicated the superiority of the approach in comparison to nine other traditional feature selection methods. Shukla et al. [43] developed a two-stage hybrid method for the classification of six cancer diseases. The proposed hybrid method (CMIMAGA) aggregates two techniques – CMIM (Conditional Mutual Information Maximization) and AGA (Adaptive Genetic Algorithm) to determine highly discriminating genes from cancer datasets. While CMIM was employed to filter out irrelevant genes, the AGA method was used as a wrapper that combined the learning algorithms as a fitness function for finding a small number of genes with maximum accuracy. The experimental results demonstrate that the proposed approach with Extreme Learning Machine (ELM) obtained fairly promising results by significantly reducing the original dataset with the selection of the most informative subset of genes and attaining high classification accuracy compared to other classifiers. The proposed hybrid strategy also reduced over-fitting and outperformed other filter and wrapper approaches.

Park et al. [26] diagnosed hypertension using a hybrid feature selection and classification technique. The hybridization aggregated symmetrical uncertainty and correlation feature selection with Bayesian classification. The experimental results concluded that the presented approach significantly improved the robustness and performance of the classifier to diagnose hypertension problems. Mert et al. [44] examined the effects of feature reduction techniques with the probabilistic neural network using a hybrid approach for classifying breast cancer datasets. The obtained results indicated that the proposed method reduced the computational complexity and enhanced the distinguishing performance of the classifier, showing the accuracy of 96.31% and 97.01% for ten-fold cross-validation and leave-one-out cross-validation techniques respectively. A computer-aided technique using feature selection and classification for the early diagnosis of Alzheimer-type dementia (ATD) was employed by Salas-Gonzalez et al. [45]. Researchers also compared the results of support vector machines and classification trees (CT) using the values of sensitivity, specificity, and accuracy rate. The analysis of results indicated that the presented diagnosis technique reached more than 95% accuracy during classification.

To diagnose CKD, the researchers have found the Support Vector Machine (SVM) classifier to be propitious in improving the diagnostic performance of the model. Polat et al. [12], for example, employed Support Vector Machines with effective feature selection methods to diagnose CKD and achieved 98.5% accuracy with this dataset. Al-Hyari et al. [46] designed a clinical decision system with an SVM classifier and obtained 93.14% accuracy to diagnose Chronic Renal Failure. To increase the diagnostic success rate, an SVM classifier has been used together with different feature selection and feature extraction algorithms to reduce the dimensionality of the datasets [47], [48]. The Principal Component Analysis (PCA) is one of these feature extraction algorithms that has been used with an SVM classifier for disease diagnosis [49], [50]. Many researchers have applied PCA with the ReliefF feature selection to eliminate extraneous features from the dataset [51]. In most of diagnostic systems, pre-processing before introducing the training data is recommended to increase the diagnosis success rate of the system.

The extant literature cumulates numerous studies demonstrating outstanding results from authors who have researched in the field of SVM classification with dimensionality reduction techniques with both text and microarray datasets. Besides, researchers are focusing on hybrid feature selection approaches to reduce the dimensionality of datasets. The most recent literature contains many studies, which have been implemented using hybrid structures. Pang et al. [52] applied ReliefF-SVM based method for the computer-aided diagnosis of breast tumors, yielding positively appealing results with a 90.0% accuracy rate, 98.7% sensitivity, and 73.8% specificity rate. Uğuz [53] presented a hybrid system with the aggregation of information gain, PCA, and SVM classifier. The information gain method was used to rank each feature based on its importance. Consequently, the most significant features were identified and passed to the PCA method for dimensionality reduction. Next, the reduced sets of features are passed as input to the classifier. The classification performance of the presented method when compared and evaluated with existing studies was found to be best performed with an SVM classifier. Chen et al.

[54] diagnosed hepatitis disease with a hybrid method integrating Fisher Discriminant Analysis Algorithm and SVM Classifier. The proposed method was compared with existing methods and the results demonstrated that the hybrid method outperformed other methods, obtaining the best classification accuracy of 96.77%. The literature provides numerous hybrid feature selection models with the usage of support vector machines with excellent results for the diagnosis of chronic diseases such as Breast Cancer [55], Diabetes [14], Lung Cancer [56], CKD [12], Heart Disease [57], Hepatitis [54] and many more. Table I gives a glimpse of previously used methods in the literature for the diagnosis of chronic diseases.

TABLE I. Accuracy Achieved by Other Researchers for the Diagnosis of Chronic Diseases

Source	Disease Dataset Considered	Method Applied	Accuracy Achieved (%)
[38]	Chronic Kidney Disease	Pre-processing + k-NN ANN	96 81
[58]	Heart Disease	Vote Technique	87.4
[12]	Chronic Kidney Disease	SVM	98.5
[57]	Heart Disease	Rule-Based Fuzzy Classifier	78
[52]	Breast Tumor	ReliefF-SVM	90
[59]	Chronic Kidney Disease	KNN SVM	78.75 78.35
[14]	Diabetes Breast Cancer	SVM	100 100
[56]	Lung Disease	Genetic Algorithm Based Feature Selection	99
[55]	Breast Cancer	SFSP + NN SBSP+ NN	97.57 98.57
[54]	Hepatitis Disease	FDA + SVM	96.77
[60]	Lymph Disease	PCA + Fuzzy Weighting Pre-Processing + ANFIS	88.83

III. Datasets and Algorithms Used

This section presents a brief overview of the datasets and materials used for this research. The first and second subsection discusses the clinical CKD dataset and repository CKD dataset used for this work. The subsequent subsections provide a brief overview of the techniques used for this work.

A. Clinical Dataset Description (CKD)

Clinical data of 337 suspected CKD patients were collected at Vasu Diagnostic Centre, Gurugram, India, and is summarized in Table II. 23 features were recorded for each patient including Age, Gender, Serum_Urea, Serum_Creatinine, Serum_Uric_Acid, Sodium, Potassium, Calcium, Total_Protein, Albumin, Hemoglobin, TLC, DLC_Polymorph, DLC_Lymphocites, DLC_Eosinophil, DLC_Monocite, Platelet, RBC, PCV, MCV, MCH, MCHC, and CKD. There are 86 missing values in this clinical dataset. All attributes are numerical except one attribute (Gender) which is categorical. The second column of Table II shows the units corresponding to each attribute. The third column depicts the normal range value of each parameter related to CKD. The fourth column shows the range of values present in the clinical data corresponding to each parameter in the CKD dataset.

TABLE II. Details of Clinical CKD Dataset

Features	Units	Normal Values	Range
Age	Years	–	02 - 90
Gender	–	–	M-Male, F-Female
Serum_Urea	mg/dL	15 – 39	4.75 – 183.3
Serum_Creatinine	mg/dl	0.60 – 1.30	0.03 – 11.4
Serum_Uric_Acid	mg/dl	2.6 - 6.0	1.55 -12.81
Sodium	mmol/L	136.0 – 149.0	120.8 – 155.3
Potassium	mmol/L	3.5 – 5.0	2.78 – 8.22
Calcium	mg/dl	8.6 – 10.3	5.74 – 10.22
Total_Protein	gm/dl	6.40 – 8.30	4.26 – 8.81
Albumin	gm/dl	3.5 – 5.0	2.34 – 4.99
Haemoglobin	g/dl	11 – 15	3.9 – 17.5
TLC	/cumm	4000 – 11000	3000 - 30900
DLC_Polymorph	%	40 – 75	20 - 92
DLC_Lymphocites	%	20 – 45	5 – 75
DLC_Eosinophil	%	1 – 6	1 – 12
DLC_Monocite	%	2 – 10	0 – 9
Platelet	100000/cumm	1.5 – 4	0.22 – 30.3
RBC	Millions/cumm	4.5 – 5.5	1.04 – 8.13
PCV	%	37 – 47	3.82 – 51.5
MCV	fl	78 – 94	22.2 – 120.2
MCH	Picogram	27 – 32	15.6 – 38.9
MCHC	gm/dl	30 – 35	21.5 – 41.9
CKD	-	-	Y, N

B. CKD Repository Dataset Description

The CKD data was collected from the UCI machine learning repository [61]. This database was selected for this research because it is a commonly used database by machine learning researchers with records that are most complete. The dataset contains 400 records with some missing values. Table III describes the description and type of attributes. There are 25 attributes (11 numeric plus 14 nominal) that feature in CKD prediction and one attribute serves as the output or the predicted attribute for the presence of CKD in a patient. The second shows the type of each attribute and the third column shows the units or values corresponding to each attribute.

C. Data Pre-Processing

Pre-processing is the process of converting raw data into a purposeful and relevant format. The actual data generally consists of inconsistent, irrelevant, surplus data containing a large number of null values. It is crucial to pre-process the dataset before training it on a classifier in order to improve its prediction ability.

Prior to applying the classification model, both CKD datasets are first pre-processed and then subjected to dimensionality reduction techniques. Fig. 1 show the pre-processing techniques used in this research work.

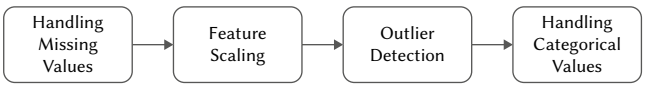


Fig. 1. Steps of Pre-processing.

In this work, the pre-processing is done in the following steps –

1. Handling Missing Data

Medical data is generally incomplete with missing data, noisy with errors or outliers and inconsistent containing discrepancies in names of features. The missing values in the dataset can be handled using some imputation techniques.

TABLE III. Details of Repository CKD Dataset

Features	Description	Range
age	Age (In Years)	0 - 90
bp	Blood Pressure	0 - 180
sg	Specific Gravity	0 - 1.025
al	Albumin	0 - 5
su	Sugar	0 - 5
rbc	Red Blood Cells	normal, abnormal
pc	Pus Cell	normal, abnormal
pcc	Pus Cell Clumps	present, notpresent
ba	Bacteria	present, notpresent
bgr	Blood Glucose Random	0 - 490
bu	Blood Urea	0 - 391
sc	Serum Creatinine	0 - 76
sod	Sodium	0 - 163
pot	Potassium	0 - 47
hemo	Haemoglobin	0 - 17.8
pcv	Packed Cell Volume	0 - 54
wc	White Blood Cell Count	0 - 26400
rc	Red Blood Cell Count	0 - 8
htn	Hypertension	yes, no
dm	Diabetes Mellitus	yes, no
cad	Coronary Artery Disease	yes, no
appet	Appetite	good, poor
pe	Pedal Edema	yes, no
ane	Anaemia	yes, no
class	Class	ckd, notckd

2. Feature Scaling

It refers to putting the values in the same range or same scale so that no variable is dominated by the other. If it is not done, then the learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. There are essentially different types of feature scaling. However, the most widely used are standardization and normalization. In standardization, we compute the transformed values by computing the difference of each feature value from the mean of all the values of that feature and then divided by the standard deviation for that feature. This transforms the data between the range of -1 and +1. The transformed data has means of 0 and a standard deviation of 1. In normalization, we compute the transformed values by computing the difference of each feature value from the minimum of all the values of that feature and then divide by the difference between the minimum and maximum value for that feature. This transforms the data between the range of 0 and 1. Normalization is generally not a good option especially when the data contains a lot of noise and outliers. In particular, when there are out, normalization will transform the normal data i.e., the data without out into a very small range of values which is not very desirable for machine learning models. So, standardization is used in this work for scaling of features.

3. Outlier Detection

An outlier is an observation point that is distant from other observations. An outlier may be due to variability in the measurement, or it may indicate an experimental error. Many machine learning algorithms are sensitive to the range in distribution of attribute values in the input data. An outlier in input data can skew and mislead the training process of machine learning algorithms. Thereby, resulting in longer training times, less accurate models, and ultimately poorer results. Therefore, the generally used approach is to get rid of outliers before passing the data to the learning algorithms.

4. Handling Categorical Data

Categorical data needs to be encoded as the majority of the machine learning models are based on mathematical equations, it will cause some problems if we keep the categorical variables in equations because we would only want numbers to be there so that we can meaningfully compute the equations, in other words, we need to encode these categorical variables into numbers. This can be done by introducing dummy variables in the dataset in which we have several columns equal to the number of categories.

D. Dimensionality Reduction Techniques

The performance of the SVM classifier is largely affected by the usage of dimensionality reduction techniques. Two general approaches to solve the problem of dimensionality are – a) feature extraction that transforms the existing features into a lower-dimensional space and b) feature selection that selects a subset of the existing features without a transformation [19]. To deal with the issue of "curse of dimensionality" and to speed up the classification tasks, researchers have proposed various methods to improve the accuracy of results.

The performance of the SVM classifier significantly improves if dimensionality reduction techniques are applied before the classification of data. Hence, researchers use feature selection and feature extraction techniques extensively to reduce high data dimensionality. For disease diagnosis, feature selection eliminates the attributes that are least significant to a particular disease. As less significant features are removed with dimensionality reduction techniques, SVM would now be working on features that affect a particular disease. Due to this, the diagnostic accuracy of SVM also increases at a significant rate.

PCA is a widely-used feature extraction technique used for dimensionality reduction that projects data from original m-dimensional space to a new dimensional space (d<m) with minimal loss of data. PCA calculates the eigen vectors of the covariance matrix of the input data. For variance to be maximum, the eigen vector with the largest eigen value is chosen as the first principal component. The second principal component is orthogonal to the first one but with slightly less variance [62]. In a nutshell, PCA is a linear algebra method that is used for continuous attributes that find new principal components that are perpendicular to each other and captures maximum variance of the data.

Suppose we have a set of n-dimensional features X=X1, X2, X3…XN and want to map it to a lower-dimensional space which is m-dimensional. The objective of using PCA is to get the features Z=Z1, Z2, Z3….ZM where M<N and each of these features is some function of the original feature set f (X1, X2, X3…XN). So, it is the projection of a higher-dimensional feature space to a lower-dimensional feature space so that the smaller dimensional feature set can help in better classification. Therefore, we need to find a projection matrix W

$$\bar{Z} = W^T \bar{X} \tag{1}$$

where W^T is a projection from N-dimensional space to M-dimensional space.

The new projection should contain uncorrelated features. The mapping to smaller spaces ensures that features are not redundant and cannot be reduced further. In addition, the features should have a large variance because if the feature takes a similar value for all the instances that feature cannot be used as a discriminant. Since we want the features to be able to distinguish between the different instances, it is better to have a larger variation between the features.

Another popular method used for dimensionality reduction is the ReliefF feature selection method proposed by [63] that assigns relevance scores to each attribute by randomly sampling an instance

from the data and then finding its nearest neighbor from the same and opposite classes. The scores corresponding to each attribute are updated by comparing the attribute values of the nearest neighbors to the sampled instance. This research focuses on proposing a novel hybrid dimensionality reduction consisting of the ReliefF method (applied with an appropriate threshold value) and the Principal Component Analysis method and is discussed in detail in the next section.

E. Support Vector Machine Classifier

Support Vector Machine classifier (SVM), developed by Vapnik in 1995, is a widely-used technique in which classification is done by projecting the input data points into n-dimensional vector space and finding the best hyperplane that maximizes the margin between two classes. An un-optimized decision boundary could result in greater misclassifications on the new data. The main goal of SVM is to identify a separating hyperplane between the positive and negative classes and to keep the boundary as far as possible. SVM operates by building a suitable model from the training data and then applying the constructed model to estimate the class values of the test data. For non-linear problems, it works by mapping the training data from low-dimensional space into high-dimensional space with the usage of kernels. It efficiently solves the quadratic optimization problem and maximizes its generalization performance for finding the best separating hyperplane [56]. Although support vector machine classifier has several benefits, yet its classification performance is often influenced by the 'curse of dimensionality'. As the data in medical datasets are increasing voluminously in terms of several features and instances, insignificant and redundant features must be removed before being passed to the appropriate classifier.

The equation of separating hyperplane is given by:

$$D(x) = (w * x) + w_o \tag{2}$$

where w and w_o are parameters of the classifier model to be evaluated given a training set D

The hyperplane should satisfy the following inequality:

$$y_i(w * x_i) + w_o \geq 1 \tag{3}$$

Given a training set of labeled pairs (x_i, y_i) where i =1, 2…., m. The SVM classification determines the solution of the following optimization problem:

$$\min_{w,b,\varepsilon} \frac{1}{2} w^T w + C \sum_{i=1}^{N} \varepsilon_i \tag{4}$$

subject to

$$y_i(w^T \emptyset(x_i) + b) \geq 1 - \varepsilon_i \quad \text{for } \varepsilon_i \geq 0 \tag{5}$$

where ε is the slack variable, C is the user-specified penalty parameter, $\emptyset$ is the Radial Basis kernel function.

Besides, the generalization ability of the SVM classifier highly depends on the appropriate model selection. The performance of the SVM classifier is highly dependent on the selection of the kernel function and the kernel function parameters, and the key to enhancing the classification accuracy is to select the appropriate values of the parameters. The grid search technique, a widely used method for optimizing SVM classifier, finds the optimal parameters to tune the performance of SVM. In this work, the RBF kernel function is used and the parameters that should be optimized for the RBF kernel function are the penalty parameter C and the gamma parameter.

IV. Proposed Diagnostic Model Design

In this paper, a fast hybrid model, i.e., R_FP-SVM, is developed to undertake CKD classification problems. The diagnostic system possesses two important implications, i.e., fast learning with high performance and identification of the most significant factors related to CKD. The framework of the proposed hybrid approach (R_FP-SVM) is described in two phases. The proposed approach consists of two phases. In the first phase shown in Fig. 2, the proposed dimensionality reduction approach is applied to discover the most informative parameters related to CKD and to eradicate extraneous features form the CKD dataset. The second phase (Fig. 3) basically improves the SVM learning and classification accuracy through efficient parameter optimization. Both phases are applied using parallel execution functionality using GPUs and machine workers to speed up the computation.

A. Data Pre-processing

To perform data pre-processing, the original CKD dataset is verified for the management of missing data, detection of outliers, scaling of features and management of categorical values. While the clinical CKD dataset contains 337 instances and 23 features, the CKD dataset taken from online repository contains 400 instances and 25 attributes. The primary aim of this work is to determine whether the person is diagnosed with CKD or not.

As seen in Fig. 1, the CKD dataset is first taken as input into the system. Next data pre-processing techniques are applied to convert it into an appropriate format.

First, both CKD datasets are checked for missing values. There are 242 patients with missing values in their records in the repository dataset and a total of 86 missing values in the clinical dataset. The features containing more than 30% missing values have been eliminated. To remove numerical missing values, first, the mean values across the column are calculated and then the missing data is replaced by the mean of the values in the column containing the missing data. For removing non-numerical values, authors first find the most frequently occurring value and use that value in place of the missing value.

Then, authors apply standardization on both sets of data to put all the entities on the same scale. For feature scaling, standardization is applied on both CKD datasets to transform the data between the range of -1 and +1 using the formula shown in equation (6).

$$x_{transformed} = \frac{x - mean(x)}{standard\ deviation(x)} \tag{6}$$

Subsequently, both datasets are checked for outliers. For the eradication of outliers, first, the median across the column is calculated and then the values that are three times away from the median are discarded i.e., those values that are three times greater or smaller than the median are excluded from the dataset.

The last step in data pre-processing is the handling of categorical variables. While the repository dataset contains 14 nominal values, clinical data contains two categorical variables (Gender and Class). Categorical values are converted into numerical form by introducing dummy variables in the dataset in which authors have a number of columns equal to the number of categories. For variables that contain two values such as "gender", the corresponding values are replaced by binary values - 0 and 1.

B. Elimination of Extraneous Features

With the rapid growth of large-sized medical data sets in recent years, the need for diminishing the dimensionality of data has risen significantly. Feature selection and feature extraction play a vital role in reducing surplus and extraneous features from disease datasets to speed up the computation as well as to enhance diagnostic accuracy. The main idea of using feature selection is selecting a subset from the original set of attributes to eliminate those parameters that do not contribute to the medical diagnosis. This research presents R_FPCA-

SVM based hybrid approach in which the dataset is reduced using appropriate feature selection using the ReliefF method and feature extraction using the PCA method. The approach presented for the elimination of extraneous features is named as R_FPCA method. This has been done to achieve good performances on running speed or/and classification accuracy.

In the first stage, the first R_FPCA method is employed for feature selection, which relies on the ReliefF method for ranking the importance of each feature based on weight values and help to reduce the computing complexity of the method, and then redundant and unrelated features are filtered out using the PCA method, and thereby extracting the dataset with most significant features. The dataset with a reduced set of features is then passed to the learning algorithm.

Fig. 2 shows the model for the proposed hybrid dimensionality reduction method R_FPCA which takes pre-processed CKD dataset as input into the system. Thereafter, the ReliefF method is applied to the dataset that yields a Weight Matrix (W). The weight matrix contains the weights respective to each attribute of the CKD dataset. Based on the weight of each feature, a rank is assigned to all the features of the dataset. Subsequently, the appropriate threshold is applied to select only relevant features from the dataset. The threshold value is dynamically calculated from weights generated by the ReliefF method. After repeated experiments, the threshold 'theta' is taken as mean (W). Then, oust those features whose weight (Wx) is lesser than the threshold (theta). By choosing among those with a large W value (i.e., those that exceed a threshold 'theta'), the final selection of attributes is performed. The resultant creates a list of features (L) with the removal of irrelevant features from the CKD dataset.

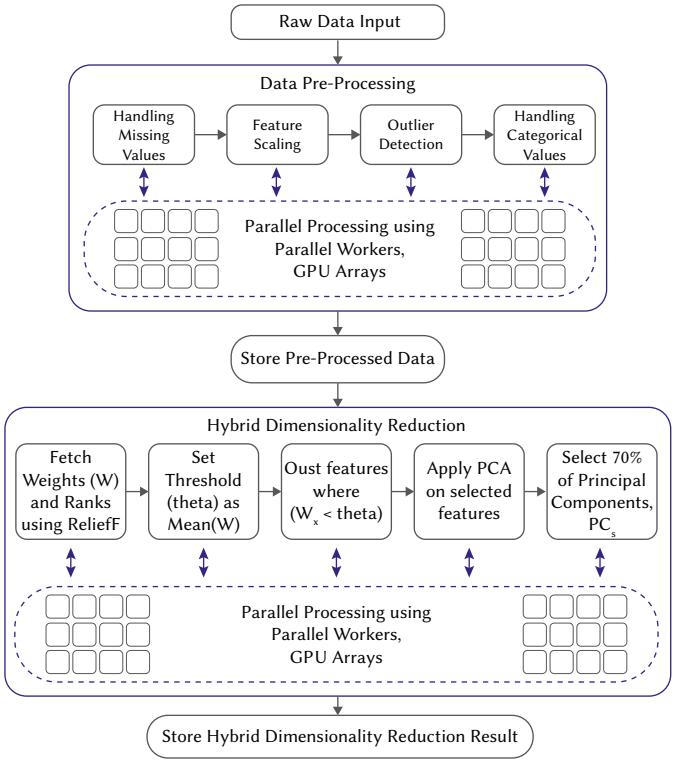


Fig. 2. Proposed Model (Phase I).

This generated list (L) with the selected features is then passed to the PCA method to remove redundant features from the dataset. PCA provides principal components corresponding to each feature and forms uncorrelated variables that maximize the variability of the data. The redundant attributes are eliminated by considering 70% of the principal components.

The proposed dimensionality reduction method is implemented on parallel processors using GPUs and machine workers to decrease the computational time. The whole work is divided among n-workers where n ranges from 2 to 16.

Finally, the CKD dataset with reduced dimensionality is passed onto the learning algorithm, that is, the second phase of the research.

Then, the weight matrix is divided into n-equal sets which are further assigned to machine workers ranging from 1 to n. After that, machine workers are parallelly executed to faster computations. In each set of execution, the weights higher than the set-out threshold value are selected and features are selected corresponding to selected weight values. Next, the selected columns which satisfy the threshold are merged. In all, the work is divided among n-workers in which computations are done simultaneously and features whose weights exceed the threshold are selected and sent to further stages.

C. Speeding Up and Tuning SVM Classification

SVM classification is a supervised learning method that identifies the hyperplane separating the two classes. The primary role of SVM is to separate labelled data based on a line maximizing the distance between the two classes. It uses the kernel trick to handle the non-linear cases by projecting the data to a high-dimensional feature space. To develop an accurate classification model, it is crucial to select a powerful machine learning algorithm and to tune up its parameters. The SVM in this work uses a Radial Basis Function (RBF) kernel and employs a grid-search method to gather and process all possible combination of hyper-parameters – Cost, Epsilon, and Gamma.

Hyper-parameters are the parameters that can`t be directly learned in the regular training process. For example – learning rate for logistic regression, number of trees in random forest classifier, cost and gamma values in SVM classifier, number of hidden layers in a neural network. The optimization of hyperparameters ensures high accuracy of SVM classifier and it also don`t increase the computational cost too significantly They help us find the balance between bias and variance and thus, prevent the model from overfitting or underfitting. Grid search is a method to perform hyper-parameter optimization, that is, it is a method to find the best combination of hyper-parameters. It is usually applied with a cross-validation method with different combinations of hyper-parameters. Each of these combinations of parameters, which correspond to a single model, can be said to lie on a point of a grid. The goal is then to train each of these models and evaluate them using cross-validation. The hyper-parameter combination which performs best is selected for training and testing the model. In this work, the best combination of SVM hyper-parameters is the one that produces maximum classification accuracy, minimum Mean Absolute Error (MAE), and least execution time. Since the RBF kernel function is applied in this work, the diagnostic performance of SVM depends heavily on an appropriate choice of its parameters. Tuning the kernel parameter gamma (σ), epsilon (ϵ) and the penalty parameter (C) would increase the efficiency of the SVM classifier. Fig. 3 shows phase II of the proposed model.

As shown in the Fig. 3 model, the resultant CKD dataset after the application of the proposed hybrid dimensionality reduction approach is passed onto the Phase II of the research. In this phase, the optimization of SVM parameters is done to obtain the best values of 'cost', 'epsilon', and 'gamma'. A grid search is conducted to find the best parameter values using 10-fold cross-validation. Every single possible combination of hyper-parameters - cost, gamma, and epsilon was tried. Use a set of possible values for each parameter and create a variable to store the model`s accuracy for each set. Then create a nested for-loop where for every value of C, authors tried every value of epsilon and gamma. A similar process is used with the other two parameters. Inside the loop, train the model and score it, and then

compare its score to the best score. If it is better, update the values accordingly. The best parameter values are then applied to the testing subset and the highest classification accuracy is recorded. This process would run for every hyper-parameter value until it finds the optimal ones. In this work, 10-fold cross-validation method is used to calculate classification accuracy. In the 10-fold cross-validation method, the training set is divided into ten subsets of equal size. Subsequently, the 10th subset is tested while the classifier trains the remaining 9 subsets. Various combinations of (C, gamma, epsilon) are tried, and the one with the best cross-validation accuracy, minimum Mean Absolute Error (MAE), and least execution time is used to create the model for training. After obtaining the predictor model, the prediction is conducted on each testing set accordingly. Fig. 4 shows the pseudo-code of the model.

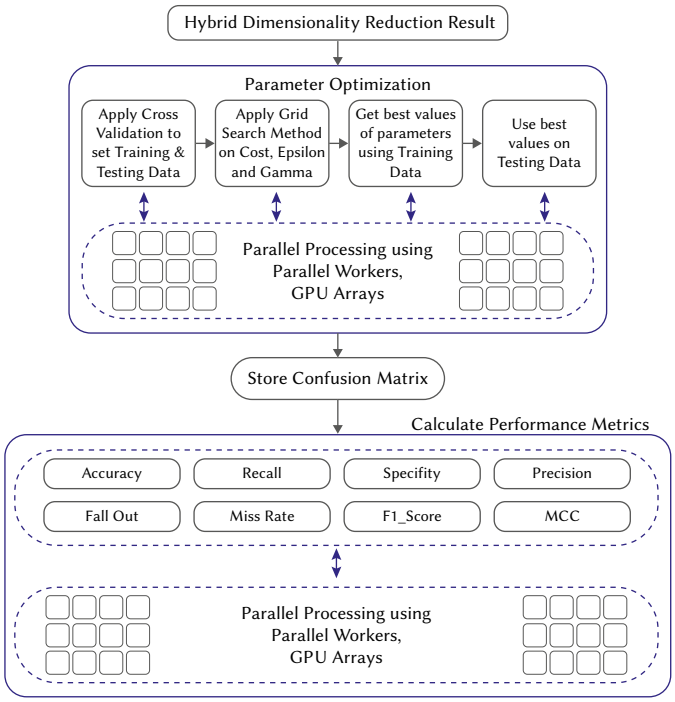


Fig. 3. Proposed Model (Phase II).

The procedure is shown as follows:

i) Consider a grid space of (C, gamma, epsilon) with C belongs to {1/2,1,2,8,10,50,100}, gamma {1/50,1/10,1/5,1/2,1.5,2,5,10} and epsilon {0.001, 0.01, 0.1}.

ii) For each combination (C, gamma, epsilon) in the search space, conducting k-fold CV on different training–testing partitions;

iii) Choose the parameter (C, Gamma, and Epsilon) that leads to the highest classification rate, least mean squared error, and least execution time.

iv) Use the best parameter to create a model for training the data set and then later use it for prediction.

For the grid search method, to execute with the least execution time, the concept of machine workers executing in parallel has been applied. Simultaneous computations using GPUs and machine workers are carried out while optimizing the SVM parameters- cost, epsilon, and gamma. To obtain the best values of SVM parameters, first, a GPU array is created for each of the parameters - cost, epsilon, and gamma, and a mesh grid is formed using these values. After that, the workers are assigned to GPU arrays that are executed parallelly to get the values of the best parameters. The best parameter selected is based on the calculation of Mean Absolute Error (MAE) and Time. For

```
Step 1: Load and Store Raw Dataset in variable DS
Step 2: Set wk as number of machine workers
Step 3: Divide DS to prepare array of sets SD with length wk
Step 4: Apply Pre-Processing method, preProcess(preprocess_name)
        Initiate Parallel Execution
            Create temporary blank List, LTemp
            If (no preprocess_name)
                preprocess_name = Missing_Values_Removal
            EndIf
            ForEach set Array SD(SD_y)
                Iterate every column, CTemp, in set SD_y
                If (Pre-process required CTemp)
                    Pre-process the data in colymn with methodName
                    Update CTemp and push in LTemp
                Else
                    Push CTemp in LTemp without pre-processing
                EndIf
            EndForEach
            Store LTemp as PDS (Pre-processed complete dataset)
            Remove existing data from DS and store PDS in DS
        Terminate Parallel Execution
Step 5: Execute Step 3 & Step 4 for below processes
            preprocess_name = Feature_Scaling
            preprocess_name = Outlier_Detection
            preprocess_name = Categorical_Vaues
Step 6: Hybrid Dimensionality Reduction
        Initiate Parallel Execution
            Create temporary blank List, HDRTemp
            Apply ReliefF method on PDS. Store output weights as W
            Calculate dynamic threshold (theta) using weights
            Divide W in array of set W_SET with wk as length
            ForEach set in Affay W_SET (W_i)
                If (weight of column >= theta)
                    Push con-esponding column in HDRTemp
                Endlf
            EndForEach
            Apply PCA method on HDRTemp
            Select 70% of Principal Components. Store result as HDR
        Terminate Parallel Execution
Step 7 : Initialize GPU Array gCost/gEps/gGamma for Cost/Eps ilon/Gamma
Step 8 : Parameter Optimization
        Initiate Parallel Execution
            Apply Grid_Search_Method on gCost/gEps/gGamma
            Apply Cross_Validation to prepare training and testing data of HDR
            Divide training data to prepare all'ay of sets TD with length wk
            ForEach set in Array TD (TD_y)
                Apply above set of values on training data
                Store best parameters values pair as (bCost/bEps/bGamma), if
                    Mean Absolute Error, Time are least and Accuracy is high
            EndForEach
        Terminate Parallel Execution
Step 9 : Use best parameters on testing data to store performance results
```

Fig. 4. Pseudo-code of the model.

the least value of MAE and time, the parameters are selected. Once all the workers are executed, n-sets of the best parameters are received. Out of those n-sets of values, the best pair out of all pairs is selected. Using the best pair, SVM classification is applied on the resultant dataset from Stage 1 and various performance metrics are recorded.

Therefore, in this phase, an optimized SVM algorithm is developed that works efficiently with both CKD datasets. Instead of running it on one single processor, the model is extended to build a parallel variant using GPUs. The parallel implementation performs all matrix computations on the GPUs. The GPUs can run multiple concurrent processes at a time. Therefore, parallel computations of the optimized SVM classifier are done implicitly.

Using this method, all the possible combinations of parameter values are evaluated and the best combination yielding maximum accuracy is retained. With the optimized SVM parameters, classification is applied on both CKD datasets and assessed based on accuracy.

V. Experimental Analysis

The experiments have been performed on the Chronic Kidney Disease Dataset for the diagnosis of patients suffering from CKD. The datasets have been described in the previous sections in detail. Datasets are obtained from two different sources. One is a disease clinical CKD dataset that contains 337 instances and 23 features and is obtained from the 'Vasu Diagnostic Centre, Gurugram, India'. The other CKD data consisting of 400 instances and 25 features are taken from UCI machine learning repository. To validate the efficacy of the proposed diagnostic model, several useful performance metrics in medical applications that include accuracy, precision, recall, f-measure, specificity are computed. The produced results are analyzed and compared with those from other methods published in the literature. The parameters used to evaluate and compare methods are the Number of Selected Features, Execution Time, and Classification Accuracy. Execution time is machine-dependent, so the algorithms have been implemented and compared on the same machine. The classification accuracy is calculated using 10-fold cross-validation strategy for the training and testing sets. The training set consists of 70% of the values and the test set consist of 30% of values. For each method, obtain the average classification accuracy, several selected features, runtime found under each algorithm and each dataset. The outcomes positively demonstrate that the hybrid diagnostic model is effective in undertaking medical diagnostic tasks.

The proposed method is implemented in MATLAB 2018a software using parallel processors. The processor(s) used for the experiments is '2 x Intel Xeon E5-2650V2' and 'Matrox G200eW' as GPU.

A. Evaluation Parameters

The diagnostic accuracy of the proposed model is measured in terms of the following evaluation measures the details of which are described below. The proposed approach is measured based on performance measures that are computed from a confusion matrix that contains four terms - True Positive (T_p), True Negative (T_N), False Positive (F_p), and False Negative (F_N) where,

- T_P: stated as the number of instances estimated positive that is actually positive.
- F_P: stated as the number of instances estimated positive that are actually negative.
- T_N: stated as the number of instances estimated negative that are actually negative.
- F_N: stated as the number of instances estimated negative that are actually positive.

Most charts graphs and tables are one column wide (3 1/2 inches or 8.89 cm) or two-column wide (7 1/16 inches or 17.93 cm). We recommend that you avoid sizing figures less than one column wide, as enlargements may distort your images and result in poor reproduction. Therefore, it is better if the image is slightly larger, as a minor reduction in size should not have an adverse effect on the quality of the image. If the size is changed, keep the proportion so that images and graphics do not distort.

1. Accuracy (A_C)

This evaluation metric estimates the proportion of exact predictions and the total number of predictions made by the classifier. It is stated as:

$$A_C = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \quad (7)$$

2. Recall (R_C)

This evaluation measure estimates the proportion of positive patterns that are perfectly classified. The larger value of recall implies that the classifier returns most of the positive results. It is defined as:

$$R_C = \frac{T_P}{T_P + F_N} \quad (8)$$

3. Specificity (S_F)

This performance measure estimates the proportion of negative patterns that are perfectly classified. The larger value of specificity implies that the classifier returns most of the negative results. It is stated as:

$$S_F = \frac{T_N}{T_N + F_P} \quad (9)$$

4. Precision (P_R)

This performance measure estimates the fraction of perfectly predicted positive observations to the total predicted positive observations. It is defined as:

$$P_R = \frac{T_P}{T_P + F_P} \quad (10)$$

5. F-Measure (F_M)

It is a single evaluation metric that merges both precisions and recalls via their harmonic mean. The mean inclines towards the smaller of the two components. So, if the value of either precision or recall is small, the value of F_M will be small. It is stated as:

$$F_M = \frac{2 \times (R_C \times P_R)}{(R_C + P_R)} \quad (11)$$

where F_M lies in the range 0,1.

6. Mathew`s Correlation Coefficient (M_{CC})

This is one of the powerful performance metrics which, in essence, is a correlation coefficient between the observed and predicted binary classifications. It is considered a balanced measure as it involves values of all the four quadrants of a confusion matrix. The range of values of M_{CC} lies between -1 to +1. A model with a score of +1 indicates a completely correct classifier and a score of -1 indicates a completely wrong classifier. It is stated as:

$$M_{CC} = \frac{T_P * T_N - F_N * F_P}{\sqrt{(T_P + F_N)(T_P + F_P)(T_N + F_N)(T_N + F_P)}} \quad (12)$$

7. Fall Out (F_{PR})

This performance metric, also known as False Positive Rate, signifies the proportion between the incorrectly classified negative samples to the total number of negative samples. In other words, it is the proportion of the negative samples that were incorrectly classified. It is stated as:

$$F_{PR} = \frac{F_P}{F_P + T_N} \quad (13)$$

8. Miss Rate (F_{NR})

Also known as False Negative Rate, miss rate implies the percentage of positive samples that were incorrectly classified. It is stated as:

$$F_{NR} = \frac{F_N}{F_N + T_P} \quad (14)$$

B. Results and Analysis

This section depicts the results and their analysis based on various factors on the CKD dataset. A detailed analysis has been done to determine the efficacy of the approach based on the proposed dimensionality reduction method, proposed classification technique, and proposed parallel execution functionality for both CKD datasets. The results have also been compared with existing feature selection and classification techniques in the literature.

1. Assessment of the Efficacy of Proposed Approach on Both CKD Datasets

Table IV and Table V exhibits the effect of applying the proposed approach on both the clinical CKD dataset and repository CKD dataset. The proposed dimensionality reduction approach yields excellent results by reducing irrelevant and redundant features from both CKD datasets. As Table IV unveils, the presented method eliminates approximately 68% from the clinical CKD dataset. Likewise, for the repository dataset, the dimensionality significantly reduces to 41.6% with the presented approach.

The diagnostic performance of the proposed classification approach is assessed based on various evaluating metrics – Accuracy, Recall, Specificity, and Precision. The proposed model yields excellent results in terms of all evaluating metrics with both CKD datasets. The columns 5 through 8 of Table IV and Table V depicts the accuracy of the proposed system, which in turn, shows the ability of the classifier to meaningfully classify positive and negative classes. For the repository dataset, the proposed approach achieves a classification accuracy of 98.5% with specificity and precision as 96.29% and 98.11% respectively. Likewise, the results are outstanding with a clinical dataset with a classification accuracy of 92.5%. The proposed model correctly classifies 311 instances out of 337 instances with recall and precision values as 96.49% and 94.82% respectively.

The performance of the proposed technique is also compared with existing feature selection techniques that have been applied with standard SVM classifier on both CKD datasets. The existing feature selection techniques considered are - 'Correlation Feature Selection (CFS)', 'Mutual Information-Based Feature Selection (MIFS)', 'ReliefF Feature Selection', 'ReliefF + CFS' method. The results of the proposed technique are found to be better compared to other existing techniques in terms of all evaluating metrics.

a) Analysis on Clinical CKD Dataset

For clinical datasets, the proposed approach yields superior results compared to other existing well-known methods in the literature. It selects the seven most significant risk factors related to CKD. The details of identified risk factors are given in section V.C. The diagnostic accuracy of the proposed approach is very high (92.5%) compared to the accuracy with other methods - 'CFS+SVM' (60.45%), ReliefF + SVM (86%), 'MIFS + SVM' (56.72%).

While the 'CFS' and 'ReliefF + CFS' method selects three and four important features respectively from the clinical dataset, the proposed technique identifies the seven most significant risk factors related to the CKD. Although MIFS and ReliefF feature selection methods select seven and eight important features respectively, the performance of the classifier is not good with the two methods. As can be seen from Table IV and Fig. 5, the performance of the 'MIFS+SVM' method is very poor compared to the proposed approach. It executes with accuracy, recall, specificity, and precision values as 56.72%, 46.43%, 64.1%, and 48.15% respectively, which are very less compared to the proposed approach. Similarly, the 'ReliefF + SVM' method executes with precision and accuracy as 47.3% and 86% respectively which are very less compared to the proposed approach (Precision: 94.82%; Accuracy: 92.5%). Likewise, the 'ReliefF + CFS + SVM' method executes with recall and precision values as 64.71% and 38.6% respectively which is very less compared to the proposed approach.

In all, it can be concluded that the proposed approach selects the most significant factors related to CKD that are verified with pathologists. This shows the effectiveness of the proposed approach in terms of dimensionality reduction and classification metrics compared to the existing techniques presented in the literature.

TABLE IV. Results Corresponding to Clinical CKD Dataset

Feature Selection Technique	Total Features	Selected Features	%age of Features Eliminated	A_c	R_c	S_F	P_R
Correlation Feature Selection + SVM	22	11	50	60.45	51.22	64.52	38.89
Mutual Information Based Feature Selection + SVM	22	12	45.45	56.72	46.43	64.1	48.15
ReliefF Feature Selection + SVM	22	8	63.63	86.01	97.22	84.4	47.3
ReliefF + CFS + SVM	22	7	68.18	54.37	64.71	49.28	38.6
R_FP-SVM (Proposed)	22	7	68.18	92.53	96.49	70	94.82

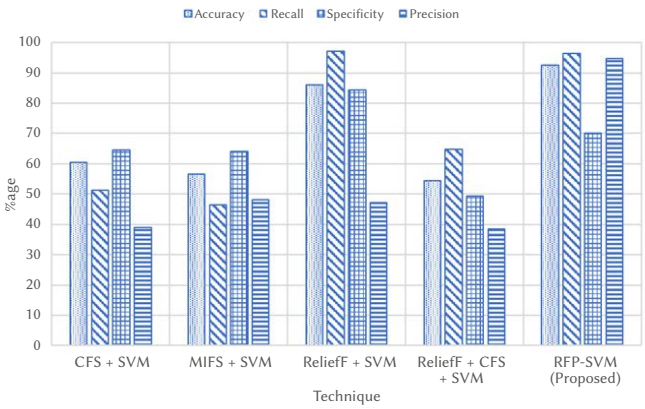


Fig. 5. Graphical representation for Clinical CKD Dataset.

b) Analysis on Repository CKD Dataset

Table V exhibits the results on the CKD dataset taken from the UCI repository. Similar to the clinical dataset, experimental results demonstrate the superiority of the proposed approach for this dataset compared to other existing methods in the literature. With the proposed dimensionality reduction approach, the size of the dataset reduces from 24 features to 10 features, thereby, eliminating approximately 58% of features from the CKD dataset. The diagnostic accuracy of the proposed approach for this dataset is extremely high (98.5%) compared to the accuracy with other methods - 'CFS+SVM' (92.7%), ReliefF + SVM (89.6%), 'MIFS + SVM' (94.7%).

The presented approach performs outstandingly well in terms of all evaluating metrics compared to other methods. As can be seen from Table V and Fig. 6, the values of specificity and precision with 'ReliefF + SVM' method are 84.38% and 89.36% respectively, while the values with the proposed approach are found to be higher with 96.29% and 98.11% respectively.

TABLE V. Results Corresponding to Repository CKD Dataset

Feature Selection Technique	Total Features	Selected Features	%age of Features Eliminated	A_c	R_c	S_F
Correlation Feature Selection + SVM	24	15	37.50	92.71	94.83	89.47
Mutual Information Based Feature Selection + SVM	24	16	33.33	94.79	93.55	97.06
ReliefF Feature Selection + SVM	24	11	54.16	89.61	93.33	84.38
ReliefF + CFS + SVM	24	10	58.33	91.67	93.22	89.19
R_FP-SVM (Proposed)	24	10	58.33	98.50	100	96.26

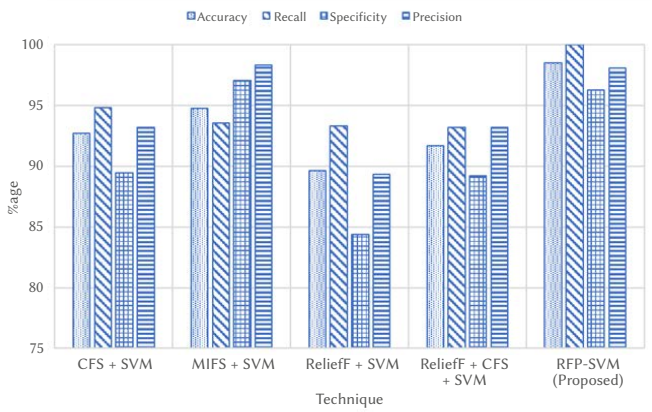


Fig. 6. Graphical representation for Repository CKD Dataset.

While the specificity values with 'CFS+SVM', 'ReliefF + SVM' and 'ReliefF + CFS + SVM' method are 89.47%, 84.38% and 89.19% respectively, the proposed approach executes with higher value of specificity i.e., 96.29%. Similarly, the proposed approach outperformed 'ReliefF+ SVM' and 'CFS+ SVM' techniques in terms of precision with 98.11% precision.

Finally, it may be concluded from Tables IV and V that the proposed hybrid approach performs better than the existing techniques in the literature.

2. Comparison of the Proposed Hybrid Approach With Other Classification Techniques Based on Various Evaluation Metrics for CKD Dataset

The performance of the proposed hybrid approach is also compared with other classification techniques based on various evaluation metrics for both CKD datasets. For the clinical dataset, the logistic regression technique and k-Nearest Neighbor (KNN) classifier yield an accuracy of 88.12% and 83.17% respectively. The proposed approach outperformed these techniques with an approximate increase of 4.4% and 9.1% respectively. Likewise, for the repository dataset, the proposed classification technique increases the accuracy of the SVM classifier from 95.65% to 98.5%. With the application of Ensemble -Boosted Trees, the accuracy comes out to be 60.87% which is very less compared to the proposed approach. Lastly, it may be concluded from Table VI and Table VII, the classification accuracy turns out to be best with the proposed system compared to other classification systems. Not only accuracy but it can also be seen that the proposed approach outperforms other techniques in terms of all performance metrics for both CKD datasets.

TABLE VI. Comparison of Proposed Approach With Existing Classification Algorithms on Clinical Dataset

Algorithm	A_C	R_C	S_F	P_R	F_{PR}	F_{NR}	F_M	M_{CC}
Logistic Regression	88.12	87.95	88.89	97.33	11.11	12.05	92.41	67.26
K-Nearest Neighbor	83.17	84.52	76.47	94.67	23.53	15.48	89.31	52.2
Ensemble -Boosted Trees	74.26	74.26	-	100	-	25.74	85.23	-
Support Vector Machines	87.13	86.05	93.33	98.67	6.67	13.95	91.93	64.56
R_FP-SVM (Proposed)	92.53	96.49	70	94.82	30	3.50	95.65	69.48

TABLE VII. Comparison of Proposed Approach With Existing Classification Algorithms on Repository Dataset

Algorithm	A_C	R_C	S_F	P_R	F_{PR}	F_{NR}	F_M	M_{CC}
Logistic Regression	95.65	97.56	92.86	95.24	7.14	2.44	96.39	90.97
Ensemble -Boosted Trees	60.87	60.87	-	100	-	39.13	75.68	-
Support Vector Machines	95.65	95.35	96.15	97.62	3.85	4.65	96.47	90.85
R_FP-SVM (Proposed)	98.50	100	96.29	98.11	3.7	0	99.04	97.2

3. Analyzing of the Results of Parallel Execution With N-processors in Multiple Iterations for CKD Dataset With Respect to Time

For faster processing of CKD datasets, the proposed hybrid approach is executed parallelly with n-processors. The value of n varies between 2 and 16. The results are iterated 5 times to record execution time for both CKD datasets corresponding to each value of the processor. The clinical CKD dataset contains 23 features and 337 instances. For each value of processor ranging from 2 to 16, the execution time is recorded for five iterations. In each iteration, it took approximately 4 sec to execute this dataset. Similarly, the repository dataset with 400 instances and 25 features, the proposed approach took approximately 5.2 sec for the execution of the dataset. As Table VIII and Table IX unveils, for most of the iterations, the proposed approach works best with 4-processors in each iteration for this dataset and shows the remarkable performance in terms of reduction of execution time with the proposed parallel execution approach.

TABLE VIII. Analysis of Parallel Execution on Clinical CKD Dataset (Time in Secs)

Iteration / Workers	Worker-2	Worker-4	Worker-6	Worker-8	Worker-10	Worker-12	Worker-14	Worker-16
Iter I	4.04	4.03	4.18	4.11	4.11	4.32	4.51	4.52
Iter II	4.16	3.95	4.06	3.9	4.24	4.11	4.29	4.36
Iter III	4.05	4.08	4.02	4.2	4.07	4.13	4.89	4.26
Iter IV	4.08	4.05	4.05	4.06	4.13	4.29	4.21	4.16
Iter V	4.13	4.1	4.02	4.05	4.32	4.05	4.15	4.04

TABLE IX. Analysis of Parallel Execution on Repository CKD Dataset (Time in Secs)

Iteration / Workers	Worker-2	Worker-4	Worker-6	Worker-8	Worker-10	Worker-12	Worker-14	Worker-16
Iter I	5.39	5.28	5.29	5.29	5.45	5.21	5.21	5.3
Iter II	5.32	5.22	5.23	5.27	5.39	5.22	5.24	5.21
Iter III	5.44	5.29	5.22	5.26	5.49	5.34	5.34	5.23
Iter IV	5.56	5.01	5.43	5.39	5.27	5.47	5.48	5.26
Iter V	5.33	5.31	5.23	5.32	5.43	5.52	5.53	5.29

C. Significant Risk Factors Identified Using the Proposed Approach

This section of the paper discusses the most significant risk factors related to CKD that are determined with the proposed approach. The most critical risk factors are identified corresponding to both CKD datasets. The identified parameters related to the clinical dataset are also confirmed with senior pathologists. Table X indicates the most crucial factors for clinical data as well as repository dataset that should be considered while diagnosing CKD disease.

TABLE X. Most Significant Risk Factors

Clinical CKD dataset	Repository CKD dataset
Age	Blood Pressure
Serum_Urea	Specific Gravity
Serum_Creatinine	Albumin
Potassium	Red Blood Cells
TLC	Pus Cell
DLC_Polymorph	Serum Creatinine
DLC_Lymphocites	Packed Cell Volume
	Red Blood Cell Count
	Hypertension
	Diabetes Mellitus

VI. Benchmarking of the Proposed Approach

Benchmarking is a widely used method that compares the performance of a model against the performance attained by state-of-the-art models. This technique is used in this research to determine whether the presented diagnostic framework has attained acceptable accuracy as compared to the accuracy achieved by the already existing studies. The classification accuracy of the proposed hybrid approach on the CKD dataset gathered from UCI machine learning repository was compared against the other four studies used in the existing work. Table XI shows the comparison of the accuracy of the proposed approach against the accuracy of the approaches used in the existing studies. Based on Table XI, it can be deduced that the presented classification model has performed better as compared to the state-of-the-art models. Based on the comparison, it is apparent that this research has generated higher accuracy with using the proposed hybrid technique.

TABLE XI. Most Significant Risk Factors

Source	Approach Used	Attained Accuracy (%)
The Proposed Model		98.5
Rady et al. [64]	Multi-Layer Perceptron	77.29
Akben [38]	Unit Synchronization + k-NN	96
Sinha et al. [59]	Support Vector Machine	78.35
Sinha et al. [59]	k-Nearest Neighbor	78.75

VII. Statistical Validation Test

The effectiveness of the proposed approach used in this work is validated through 'Friedman-Test' [65]. It is a widely used non-parametric approach that efficiently tests the null hypothesis of identical populations.

In this test, first, the null hypothesis (H_{nu}) and alternative hypothesis (H_{at}) are formulated in the beginning. Here, they are stated as follows:

H_{nu}: No difference between all approaches

H_{at}: Difference between all approaches

Next, the significance level (alpha) and test statistics are stated. Here alpha value is 0.05 i.e., 5% significance level. The test statistics are used to compare the rank of p-algorithms over d-datasets and is defined as:

It ranks all the models as mentioned in Table XII, Table XIII, and depending on the test statistics and calculations, it determines the value of F_R from equation (15). Next, from the critical value of the chi-squared table, the null hypothesis is either accepted or rejected. The decision rule then states that the null hypothesis should be rejected if F is greater than the critical value.

Table XII shows the ranking of different classification algorithms based on different evaluation parameters of the Clinical CKD Dataset. Likewise, Table XIII depicts the ranking table of the Repository CKD Dataset. The Friedman statistical test is applied separately on both CKD datasets to check their validity.

For Clinical CKD Dataset, putting values of *d=8, q=5, R (30, 18, 12, 26, 34)* in equation (15), F_R is achieved as 16. From the chi-squared table for the value of q and degree of freedom (0.05), the critical value is 9.49. Since F_R is higher than the critical value (9.49), the null hypothesis (H_{nu}) is rejected; hence, there exists a statistically significant difference between all approaches.

For Repository CKD Dataset, putting values of *d=8, q=4, R (21, 11, 19, 29)* in equation (15), F_R is achieved as 12.3. From chi-squared table for the value of q and degree of freedom (0.05), the critical value is 7.5. Since F_R is higher than the critical value (7.5), the null hypothesis (H_{nu}) is rejected; hence, there exists a statistically significant difference between all approaches.

TABLE XII. Ranking for the Clinical CKD Dataset

Algorithm	Logistic Regression	K-Nearest Neighbor	Ensemble -Boosted Trees	Support Vector Machines	R_FP-SVM
Accuracy	4	2	1	3	5
Recall	4	2	1	3	5
Specificity	4	3	1	5	2
Precision	3	1	5	4	2
Fall out	3	4	1	2	5
Miss Rate	4	2	1	3	5
F-Measure	4	2	1	3	5
MCC	4	2	1	3	5
Ranks	**30**	**18**	**12**	**26**	**34**

TABLE XIII. Most Significant Risk Factors

Algorithm	Logistic Regression	Ensemble -Boosted Trees	Support Vector Machines	R_FP-SVM
Accuracy	3	1	2	4
Recall	3	1	2	4
Specificity	2	1	3	4
Precision	1	4	2	3
Fall out	4	1	3	2
Miss Rate	3	1	2	4
F-Measure	2	1	3	4
MCC	3	1	2	4
Ranks	**21**	**11**	**19**	**29**

VIII. Discussion

This research work has proposed an influential method for diagnosing CKD that can be used as a screening tool to assist in decision-making for preliminary medical diagnosis. The research has been carried out on the clinical CKD data collected from a diagnostic center that contains 337 instances and 23 features. The benchmarked CKD dataset from the UCI repository that contains 400 instances, and 25 attributes is also considered in this work and results are compared with the algorithms used in the literature. The presented model using hybridized dimensionality reduction method along with parallel classification model can diagnose CKD by capturing the knowledge

inherent in the CKD dataset accurately as indicated by all performance metrics described in section V.A.

The dimensionality reduction and classification results on the clinical and repository dataset presented in Tables IV and V signify that the identified significant features have enhanced the accuracy compared to existing machine learning techniques. Table X depicts the most critical risk factors that are determined from the proposed approach corresponding to both CKD datasets. The identified significant risk factors are confirmed with pathologists to confirm the effectiveness of the results. This confirms the findings presented in Section V on the significant attributes in the prediction of CKD.

According to the results depicted in tables IV through VII, the prediction model developed using the hybridized approach, R_FP-SVM, achieved the highest accuracy of 98.5% for the repository dataset and 92.5% accuracy for the clinical dataset. Tables VI and VII signify the superiority of the proposed approach over other classification algorithms used in the past by various researchers. Tables VIII and IX present the analysis of the results of proposed parallel execution functionality with n-processors in multiple iterations for both CKD datasets. Table XI shows the evaluation results which compare the proposed model with state-of-the-art algorithms. Finally, the predictive model is statistically checked in section VII to confirm the validity of the results.

Overall, this research work demonstrates that the proposed hybrid parallel classification model identified significant features and has significantly improved the diagnostic performance of CKD. Since R_FP-SVM outperforms other existing feature selection and classification methods, it was identified as the best performing technique among all other techniques. The experimental results have encouraged further research to examine other hybrid methods using different combinations of machine learning algorithms to improve the performance of the prediction models. Furthermore, the proposed methodology used in this work through machine learning techniques is readily applicable to many other realms of medicine as well. Other diseases, such as diabetic kidney disease, may also be predicted by considering the diabetic dataset and analyzing the attributes of the patients who are suspected to be positive with the algorithm. This research work reaffirmed the potential ability of machine learning algorithms to classify patients into appropriate categories to assist with the assessment process for their risk of developing a particular disease.

IX. Conclusion and Future Scope

As Chronic Renal Failure progresses slowly, early diagnosis and in-time treatment are the only ways to reduce the mortality rate. Classification techniques are gaining significance in the healthcare field because of their ability to classify disease datasets with high precision. This research work presents a fast, novel classification system to diagnose renal disease based on real clinical data. This diagnostic system is based on the efficient optimization of the SVM classifier with the hybridized dimensionality reduction approach to identify the most significant risk factors parameters related to CKD. The performance of the developed model is assessed in terms of diagnostic accuracy, recall, precision, specificity, and decisions made by experienced physicians. The obtained results showed the proposed approach to be the most accurate for the repository dataset (98.5%) when compared to state-of-the-art algorithms. The results are also outstanding with a clinical dataset with a classification accuracy of 92.5%. The best prediction model was created using the seven significant parameters for clinical data when insignificant features are removed from the dataset. Therefore, it may be concluded that the proposed approach executes in the least time with high classification accuracy and competes with and even outperforms other methods in the literature.

The contribution of the research work can be stated three-fold: 1) first, high classification accuracy is achieved in the diagnosis of CKD for both clinical dataset and repository dataset. In addition, the model performs outstandingly well in terms of other evaluation measures such as precision, specificity, recall and f-measure 2) second, the most significant risk factors related to CKD are identified and the least significant parameters are eliminated from the dataset using the proposed dimensionality reduction method 3) third, the diagnostic model executes in the least time as each task is executed simultaneously with multiple processors. The model supports but does not replace physician's diagnostic process and can assist in taking effective clinical decisions by medical professionals. It can be used as a screening tool to assess and evaluate the utility of extracted knowledge for use in preliminary diagnosis by non-specialist medical professionals for effective decision-making. Overall, the most significant result of the work is an improvement in the diagnostic power of the whole diagnostic process.

The major bottleneck of this research work was that the clinical dataset had to be provided by expert pathologists; this caused long delays in data acquisition and a certain reluctance to accept the procedure in everyday practice.

This research can be extended with the application of the proposed approach on a large-scale real-world dataset. Further research can be carried out to test different combinations of machine learning techniques in CKD prediction. Additionally, new hybrid dimensionality reduction methods can be applied to get a broader perspective on the informative parameters related to CKD disease to enhance the prediction accuracy. Also, research can be further tested with deep learning methods by collecting higher dimensionality datasets.

References

[1] C. Nordqvist, "Symptoms, causes, and treatment of chronic kidney disease," https://www.medicalnewstoday.com/articles/172179.php. Accessed 14 Jan 2019.

[2] WebMed, "Kidney Disease," https://www.webmd.com/a-to-z-guides/understanding-kidney-disease-basic-information. Accessed 23 April 2020.

[3] P. Kathuria, and B. Wedro, "Chronic Kidney Disease," https://www.emedicinehealth.com/chronic_kidney_disease/article_em. Accessed 23 Feb 2019.

[4] Y. Kazemi and S. A. Mirroshandel, "A novel method for predicting kidney stone type using ensemble learning," Artificial Intelligence in Medicine, vol. 84, pp. 117–126, Jan. 2018.

[5] Kidney Care UK,2017, "An estimated 1 in 10 people worldwide have chronic kidney disease," https://www.kidneycareuk.org/news-and-campaigns/news/estimated-1-10-people-worldwide-have-chronic-kidney-disease/. Accessed 25 March, 2020.

[6] P. P. Varma, "Prevalence of chronic kidney disease in India - Where are we heading?," Indian Journal of Nephrology, vol. 25, no. 3. pp. 133–135, 2015.

[7] V.A Luyckx, M. Tonelli, J. W. Stanifer, "The global burden of kidney disease and the sustainable development goals" Bull World Health Organ, vol. 96. No. 6, pp. 414-422D, 2018. doi: 10.2471/BLT.17.206441.

[8] CDC, "National Chronic Kidney Disease Fact Sheet, 2017," https://www.cdc.gov/diabetes/pubs/pdf/kidney_factsheet.pdf. Accessed 25 March, 2019.

[9] NHS, "NHS Kidney Care," https://www.england.nhs.uk/improvement-hub/wp-content/uploads/sites/44/2017/11/Chronic-Kidney-Disease-in-England-The-Human-and-Financial-Cost.pdf. Accessed 25 March, 2019.

[10] National Kidney Foundation, "Chronic Kidney Disease (CKD) Symptoms and Causes," https://www.kidney.org/atoz/content/about-chronic-kidney-disease. Accessed 25 March, 2019.

[11] National Kidney Foundation, "Global Facts: About Kideny Disease," Retrieved from https://www.kidney.org/kidneydisease/global-facts-about-kidney-disease on 12th February, 2019.

[12] H. Polat, H. Danaei Mehr, and A. Cetin, "Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods," Journal of Medical Systems, vol. 41, no. 4, p. 55, Apr. 2017.

[13] P. Meesad and G. G. Yen, "Combined numerical and linguistic knowledge representation and its application to medical diagnosis," IEEE Transactions on Systems, Man, and Cybernetics. Part A Systems Humans., vol. 33, no. 2, pp. 206–222, 2003.

[14] E. Gürbüz and E. Kılıç, "A new adaptive support vector machine for diagnosis of diseases". Expert Systems, vol. 31, no. 5, pp. 389-397, 2014.

[15] M. Seera and C. P. Lim, "A hybrid intelligent system for medical data classification," *Expert Systems* with *Applications*, vol. 41, no. 5, pp. 2239–2249, 2014.

[16] N Liu,E.S Qi, M. Xu, M., B. Gao, B. and G.Q. Liu," A novel intelligent classification model for breast cancer diagnosis," Information Processing & Management, vol. 56,no. 3, pp. 609-623, 2019.

[17] I. Mandal and N. Sairam, "Accurate prediction of coronary artery disease using reliable diagnosis system," Journal of Medical Systems, vol. 36, no. 5, pp. 3353–3373, 2012.

[18] D. Jain and V. Singh, "Utilization of Data Mining Classification Approach for Disease Prediction: A Survey," International Journal of Education and Management Engineering, vol. 6, no. 6, pp. 45–52, 2016.

[19] D. Jain and V. Singh, "Feature selection and classification systems for chronic disease prediction: A review," Egyptian Informatics Journal, vol. 19, no. 3. pp. 179–189, 2018.

[20] C. Cortes, and V. Vapnik, "Support vector machine," Machine learning, vol. 20, no. 3, pp. 273-297, 1995.

[21] H. L. Chen, B. Yang, G. Wang, J. Liu, Y.D. Chen and D. Y. Liu, "A three-stage expert system based on support vector machines for thyroid disease diagnosis," Journal of medical systems, vol. 36, no. 3 pp. 1953-1963, 2012.

[22] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert Systems* with Applications, vol. 36, no. 2 PART 2, pp. 3240– 3247, 2009.

[23] Lin, S. L. and Liu, Z, "Parameter selection in SVM with RBF kernel function," Journal-Zhejiang University of Technology, vol. 35, no. 2, p. 163, 2007.

[24] Y. Wang and L. Feng, "Hybrid feature selection using component co-occurrence based feature relevance measurement," Expert Systems with Applications, vol. 102, pp. 83-99, 2018.

[25] L. Parisi, N. RaviChandran, and M. L. Manaog, "Feature-driven machine learning to improve early diagnosis of Parkinson's disease," Expert Systems with Applications, vol. 110, pp. 182–190, Nov. 2018.

[26] H. W. Park, D. Li, Y. Piao, and K. H. Ryu, "A hybrid feature selection method to classification and its application in hypertension diagnosis," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10443 LNCS, pp. 11– 19, 2017.

[27] H. Lu, J. Chen, K. Yan, Q. Jin, Y. Xue and Z. Gao, "A hybrid feature selection algorithm for gene expression data classification," Neurocomputing, vol. 256, pp. 56-62, 2017.

[28] J. Xie, J. Lei, W. Xie, Y. Shi, and X. Liu, "Two-stage hybrid feature selection algorithms for diagnosing erythemato-squamous diseases," Health Information Science and Systems, vol. 1, no. 1, p. 10, Dec. 2013.

[29] S. S. Kannan and N. Ramaraj, "A novel hybrid feature selection via Symmetrical Uncertainty ranking based local memetic search algorithm," *Knowledge-Based Systems*, vol. 23, no. 6, pp. 580–585, 2010.

[30] S. Ramya, and N. Radha, "Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms," International Journal of Innovative Research in Computer and Communication Engineering, vol. 4, no. 1 pp. 812-820, 2016.

[31] M. Kumar, "Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm," International Journal of Computer Science and Mobile Computing, vol. 5, no. 2 pp. 24-33, 2016.

[32] A. Dubey, "A Classification of CKD Cases Using MultiVariate K-Means Clustering," International Journal of Scientific and Research Publications, vol. 5, no. 8 pp. 1–5, 2015.

[33] L. J. Rubini and P. Eswaran, "Generating comparative analysis of early stage prediction of Chronic Kidney Disease," International OPEN ACCESS Journal of Modern Engineering Research, vol. 5, no. 7 pp. 49–55, 2015.

[34] E. M. Senan, M. H. Al-Adhaileh, F. W. Alsaade, T. H. Aldhyani, A. A. Alqarni, N. Alsharif, M. Y. Alzahrani, "Diagnosis of Chronic Kidney Disease Using Effective Classification Algorithms and Recursive Feature Elimination Techniques," Journal of Healthcare Engineering, pp. 1-10, 2021.

[35] S. Vijayarani, and S. Dharyanand, "Data mining classification algorithms for kidney disease prediction," International Journal on Cybernetics & Informatics, vol. 4, no. 4 pp. 13–25, 2015.

[36] N. A. Almansour, H. F. Syed, N. R. Khayat, R. K. Altheeb, R. E. Juri, J. Alhiyafi, and S.O. Olatunji, "Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study," Computers in biology and medicine, vol. 109, pp. 101-111, 2019.

[37] S. K. Sahu and A. K. Shrivas, "Comparative Study of Classification Models with Genetic Search Based Feature Selection Technique," International Journal of Applied Evolutionary Computation, vol. 9, no. 3, pp. 1-11, 2018.

[38] S. B. Akben, "Early Stage Chronic Kidney Disease Diagnosis by Applying Data Mining Methods to Urinalysis, Blood Analysis and Disease History," Irbm, vol. 39, no. 5, pp. 353–358, Nov. 2018.

[39] R. Misir, M. Mitra and R.K. Samanta, "A reduced set of features for chronic kidney disease prediction," Journal of pathology informatics, vol. 8, 2017.

[40] J. Norouzi, A. Yadollahpour, S. A. Mirbagheri, M. M. Mazdeh and S. A. Hosseini, "Predicting renal failure progression in chronic kidney disease using integrated intelligent fuzzy expert system," Computational and mathematical methods in medicine, 2016.

[41] A. A. Serpen, "Diagnosis Rule Extraction from Patient Data for Chronic Kidney Disease Using Machine Learning," International Journal of Biomedical and Clinical Engineering, vol. 5, no. 2, pp. 64–72, 2016.

[42] T. Li and S. Fong, "A Fast Feature Selection Method Based on Coefficient of Variation for Diabetics Prediction Using Machine Learning," International Journal of Extreme Automation and Connectivity in Healthcare, vol. 1, no. 1, pp. 55–65, 2018.

[43] A. K. Shukla, P. Singh, and M. Vardhan, "A two-stage gene selection method for biomarker discovery from microarray data for cancer classification," Chemometrics and Intelligent Laboratory Systems, vol. 183, pp. 47–58, Dec. 2018.

[44] A. Mert, N. Kılıç, and A. Akan, "An improved hybrid feature reduction for increased breast cancer diagnostic performance," Biomedical Engineering Letters, vol. 4, no. 3, pp. 285-29, 2014.

[45] D. Salas-Gonzalez, J. M. Górriz, J. Ramírez, M. López, I. Alvarez, F. Segovia, and C. G. Puntonet, "Computer-aided diagnosis of Alzheimer's disease using support vector machines and classification trees," Physics in Medicine and Biology, vol. 55, no. 10, pp. 2807, 2010.

[46] A.Y. Al-Hyari, A. M. Al-Taee, and M.A. Al-Taee, "Diagnosis and classification of chronic renal failure utilising intelligent data mining classifiers," International Journal of Information Technology and Web Engineering (IJITWE), vol. 9, no. 4, pp. 1-12, 2014.

[47] S.A. Mostafa, A. Mustapha, M.A. Mohammed, R.I. Hamed, N. Arunkumar, M.K.A. Ghani, and S.H. Khaleefah, "Examining multiple feature evaluation and classification methods for improving the diagnosis of Parkinson's disease," Cognitive Systems Research, vol. 54, pp. 90-99, 2019.

[48] S. Shilaskar and A. Ghatol, "Feature selection for medical diagnosis: Evaluation for cardiovascular diseases," Expert Systems with Applications, vol. 40, no. 10, pp. 4146-4153, 2013.

[49] D. Çalişir and E. Dogantekin, "A new intelligent hepatitis diagnosis system: PCA-LSSVM," Expert Systems with Applications, vol. 38, no. 8, pp. 10705–10708, 2011.

[50] I. Babaoğlu, O. Findik, and M. Bayrak, "Effects of principle component analysis on assessment of coronary artery diseases using support vector machine," Expert Systems with Applications, vol. 37, no. 3, pp. 2182–2185, 2010.

[51] D. Jain and V. Singh, "An Efficient Hybrid Feature Selection model for Dimensionality Reduction," in Procedia Computer Science, vol. 132, pp. 333–341, 2018.

[52] Z. Pang, D. Zhu, D. Chen, L. Li, and Y. Shao, "A Computer-Aided Diagnosis System for Dynamic Contrast-Enhanced MR Images Based on Level Set Segmentation and ReliefF Feature Selection," Computational and Mathematical Methods in Medicine, vol. 2015, pp. 1–10, 2015.

[53] H. Uğuz, "A hybrid system based on information gain and principal component analysis for the classification of transcranial Doppler signals," Computer Methods and Programs in Biomedicine, vol. 107, no. 3, pp. 598–609, 2012.

[54] H. L. Chen, D. Y. Liu, B. Yang, J. Liu, and G. Wang, "A new hybrid method based on local fisher discriminant analysis and support vector machines for hepatitis disease diagnosis," Expert Systems with Applications, vol. 38, no. 9, pp. 11796– 11803, 2011.

[55] M. S. Uzer, O. Inan, and N. Yilmaz, "A hybrid breast cancer detection system via neural network and feature selection based on SBS, SFS and PCA," Neural Computing & Applications, vol. 23, no. 3–4, pp. 719–728, 2013.

[56] C. Lu, Z. Zhu, and X. Gu, "An Intelligent System for Lung Cancer Diagnosis Using a New Genetic Algorithm Based Feature Selection," Journal of Medical Systems, vol. 38, no. 9, p. 97, Sep. 2014.

[57] G. T. Reddy and N. Khare, "An Efficient System for Heart Disease Prediction Using Hybrid OFBAT with Rule-Based Fuzzy Logic Model," Journal of Circuits, Systems and Computers, vol. 26, no. 04, p. 1750061, Apr. 2017.

[58] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," Telemat. Informatics, vol. 36, pp. 82–93, Mar. 2019.

[59] Parul Sinha and Poonam Sinha, "Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM," *International Journal* of *Engineering Research* & Technology, vol. V4, no. 12, Dec. 2015.

[60] K. Polat and S. Güneş, "Automatic determination of diseases related to lymph system from lymphography data using principles component analysis (PCA), fuzzy weighting pre-processing and ANFIS," Expert Systems with Applications, vol. 33, no. 3, pp. 636–641, 2007.

[61] M. Lichman, "UCI Machine Learning Repository," http://archive.ics.uci.edu/ml. Accessed 20 March 2020.

[62] D. Jain and V. Singh, "A two-phase hybrid approach using feature selection and Adaptive SVM for chronic disease classification," International Journal of Computers and Applications, vol. 43, no. 6, pp. 524-536, 2021.

[63] I. Kononenko, "Estimating attributes: Analysis and extensions of RELIEF," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 784 LNCS, pp. 171–182, 1994.

[64] E.H.A. Rady, and A.S. Anwar, "Prediction of kidney disease stages using data mining algorithms," Informatics in Medicine Unlocked, 2019.

[65] M. Friedman, "A Comparison of Alternative Tests of Significance for the Problem of m Rankings," The Annals of Mathematical Statistics., vol. 11, no. 1, pp. 86–92, 1940.

Dr. Vijendra Singh

Prof. Vijendra Singh received his Ph.D degree in Engineering and M. Tech degree in Computer Science and Engineering from Birla Institute of Technology, Mesra, India. He has 20 years of experience in research and teaching including IT industry. Dr. Singh major research concentration has been in the areas of data mining, image processing, big data, machine learning and deep learning. He has published more than 65 scientific papers in this domain. He has served as Editor in Chief, Procedia Computer Science, Vol 167, 2020, Elsevier; Editor in Chief, Procedia Computer Science, Vol 132, 2018, Elsevier; Editor in Chief, International Journal of Social Computing and Cyber-Physical Systems, Inderscience, UK; Editorial Board Member, International Journal of Information and Decision Sciences, Inderscience, UK. He has successfully organized several international events as a lead role including Elsevier International Conference on Computational Intelligence and Data Science (ICCIDS2019), 7-8 September 2019, Delhi-NCR, India; Elsevier International Conference on Computational Intelligence and Data Science, 6-7, April 2018, Delhi-NCR, India.

Dr. Divya Jain

Dr. Divya Jain is currently working as an Assistant Professor in the Department of CSE & IT. She holds a Doctorate in the area of Machine Learning with extensive experience in research and academics. She had completed her BTech (CSE) with Honors and MTech (CSE) with first division. She has many publications in reputed international journals and conferences including Elsevier, IGI Global and Taylor & Francis. She is an Ad-hoc Reviewer with various reputed journals and conferences. Her research areas include Data Mining, Machine Learning and Web Development.

HDDSS: An Enhanced Heart Disease Decision Support System Using RFE-ABGNB Algorithm

M. Dhilsath Fathima[1]*, S. Justin Samuel[2], S. P. Raja[3]

[1] Research Scholar, Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu (India)
2 Department of Computer Science and Engineering, PSN Engineering College, Tirunelveli, Tamil Nadu, (India)
3 School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, (India)

Received 20 April 2021 | Accepted 12 July 2021 | Published 20 October 2021

Abstract

Heart disease is the leading cause of mortality globally. Heart disease refers to a range of disorders that affect the heart and blood vessels. The risks of developing heart disease become minimized if heart disease is detected early. Previous studies have suggested many heart disease decision-support systems based on machine learning (ML) algorithms. However, the lower prediction accuracy is the main issue in these heart disease decision-support systems. The proposed work developed a heart disease decision-support system (HDDSS) that can predict whether or not a person has heart disease. The main goal of this research work is to use the RFE-ABGNB to improve HDDSS prediction accuracy. The Cleveland heart disease dataset is used for training and validating the proposed HDDSS. The two significant stages of HDDSS are the feature selection stage and the classification modeling stage. The recursive feature elimination (RFE) technique is used in the first stage of HDDSS to select the relevant features of the heart disease dataset. In the second stage of HDDSS, the proposed Adaptive boosted Gaussian Naïve Bayes (ABGNB) algorithm has been used to construct a classification model for training and validating a heart disease decision-support system. An output of HDDSS is analyzed using various classification output measures. According to the results obtained, our proposed method attained a predictive performance of 92.87 percent. This HDDSS model would perform well when compared to other heart disease decision-support systems found in the literature. According to our experimental analysis, the RFE-ABGNB focused heart disease decision-support system is more appropriate for a heart disease prediction.

Keywords

ABGNB Algorithm, Heart Disease Prediction, Machine Learning, Recursive Feature Elimination, UCI Heart Disease Dataset.

DOI: 10.9781/ijimai.2021.10.003

I. Introduction

Heart disease is the leading cause of the increasing mortality rate in humans. The risk of heart disease is more in patients with uncontrolled diabetes, low high-density lipoprotein (HDL), increased low-density lipoprotein (LDL), higher Body Mass Index (BMI), smoking, and high blood pressure [1], [2]. As a result, predicting the onset of heart disease at an early stage is essential for controlling risk factors and preventing heart disease. The main objective of this paper is to build a heart disease decision-support system (HDDSS) using the RFE-ABGNB model. This HDDSS can predict heart disease risk and can be used to diagnose and prevent heart disease at an early stage [3], [4].

The HDDSS uses the RFE-ABGNB, which is a machine learning approach, to predict heart disease. Recursive feature elimination (RFE) is a feature selection method used by HDDSS to select the relevant input features from a heart disease dataset. The Adaptive Boosted Gaussian Naive Bayes Algorithm (ABGNB) is a proposed ensemble classifier used to build an HDDSS that predicts heart disease in people by evaluating heart disease risk factors. The HDDSS utilizes the UCI heart disease dataset (UCI) for training and validating the proposed ABGNB classifier. The HDDSS model will predict the probability of developing heart disease using the patient heart disease risk factors as input. The HDDSS efficiency has been evaluated using various classification performance measures.

The remainder of the paper is structured as follows: Section II reviews relevant state-of-the-art research in the automated heart disease diagnosis system; Section III describes the proposed model; and Section IV illustrates and assesses the proposed model's experimental results. Section V depicts the conclusion of the proposed work.

II. Related State-of-art Work

Machine learning (ML) techniques are increasingly being utilized to predict heart disease. This section discusses the state-of-art approaches to develop a heart disease decision-support system using ML algorithms. The Cleveland heart disease dataset is used as an input in all examined literature to build a heart disease prediction model. Haq, Amin Ul et al. [5] developed a hybrid intelligent system

* Corresponding author.
E-mail address: dilsathveltech123@gmail.com

framework for predicting heart disease. This model uses three feature selection techniques such as Relief feature selection method, the minimal-redundancy-maximal relevance method, least absolute shrinkage and selection operator methods (LASSO) for selecting the best features of the input dataset. Logistic regression, K-nearest neighbor, Artificial neural network, Support vector machine (SVM), decision tree, and naive Bayes are among the ML classifiers used to classify the selected features. Various classifier performance measures have been used to test the proposed classifier results. According to the results of the performance analysis, logical regression and support vector machine outperforms other classifiers. Logistic regression got 84% accuracy before feature selection and achieved 89% after the relief feature selection algorithm. SVM obtained 86% accuracy before feature selection and 88% after the LASSO feature selection method. The results of this study show that integrating feature selection techniques into machine learning classifiers increases classifier accuracy.

This research paper [6] compared the performance of three machine learning algorithms such as BayesNet (BN), SVM, functional trees (FT) for effective diagnosis and monitoring of the consequences of heart disease. In this work, the BayesNet algorithm and SVM achieved 83.8% accuracy, and Functional trees achieved 81.5% accuracy. Then, the Best first selection algorithm is applied to select the best feature. The accuracy of the classifiers is improved by about 3% this time after they trained using the selected features. Thus, BayesNet's accuracy increased to 84.5 percent, SVM achieved 85.1 percent accuracy, and Functional trees achieved 84.5 percent accuracy.

Mohan et al. [7] proposed a hybrid machine learning model called Hybrid random forest with a linear model (HRFLM) for predicting cardiovascular disease. HRFLM-based heart disease prediction model gave a prediction accuracy level of 88.7% which is above other ML classifiers such as naïve bayes, generalized linear model, logistic regression, deep learning, decision tree, random forest, gradient boosted trees, SVM, VOTE classifier.

This research paper [8] used various machine learning classification algorithms such as SVM, k-nearest neighbor(K-NN), artificial neural network (ANN), naïve bayes (NB), logistic regression (LR), decision tree (DT) for the identification of heart disease. This model used a feature selection algorithm called the fast conditional mutual information algorithm (FCMIM) to improve classifier accuracy with improved classifier execution time. The performance of the FCMIM method has been compared with other feature selection algorithms like Relief, Minimal-redundancy-maximal-relevance (mRmR), Least-absolute-shrinkage-selection-operator algorithm (LASSO), Local learning-based features selection algorithms (LLBFS). This outputs analysis exhibits that the FCMIM outperforms other four feature selection method on the specified ML algorithms.

Chen et al. [9] developed a heart disease prediction system (HDPS) using Learning vector quantization (LVQ) which is a prototype-based classification algorithm that works based on Artificial intelligence network concepts. HDPS achieved an accuracy score of 80%, sensitivity of 85%, and specificity of 75%.

Hidayet takci [10] proposed an improved heart attack prediction system to decide the best machine learning approach and the best feature selection technique for predicting heart disease. This author has done a comparative analysis of ML algorithms such as c4.5 classifier, Classification-Regression Tree, SVM, Iterative Dichotomiser 3, K-NN, Multi-layer perceptron (MLP), Naïve bayes, Logistic regression models, and feature selection methods like reliefF, Forward-logit, Backward-logit, Fisher filtering. This model uses a Statlog heart disease dataset which is a publicly available dataset. A computer-aided heart disease diagnosis system is built using a combination of feature selection and classification algorithms in this model. Based on the comparative analysis outcomes, a SVM with a linear kernel is suggested as the best classification model when combined with the reliefF feature selection method. This model indicates that Linear kernel SVM with ReliefF feature selection algorithm is more efficient at predicting heart disease, with an accuracy of 84.81 percent.

Thippa Reddy et al. [11] developed an automated heart disease prediction model using a firefly and BAT swarm intelligence-based OFBAT-RBFL algorithm. This model focuses on three publicly accessible heart study datasets from the UCI machine learning repository: Hungarian, Cleveland, and Switzerland. The Fuzzy logic model is used to make a classification model by generating fuzzy system rules using the selected features. Then OFBAT algorithm is applied for selecting relevant fuzzy rules, enhance the performance of the prediction model, and optimizing the output rules of the fuzzy logic system. The outcome of the experiment indicates that the RBFL algorithm outperforms the existing ML model by achieving 78 percent accuracy.

This related study demonstrates how researchers have used feature selection approaches and machine classifiers to develop an automated heart disease diagnosis model. The motivation of this study is to enhance the accuracy of the heart disease prediction model using improved feature selection and machine learning classifier. The following is the contribution of this suggested work: (1) Recursive feature elimination algorithm is used to select a relevant feature of the input dataset, (2) For building a heart disease decision-support system, the ABGNB is proposed and used as a classifier.

III. Proposed Methodology

The proposed HDDSS has been developed using the RFE-ABGNB methodology, which combines the recursive feature elimination method (RFE) for identifying significant heart disease risk factors with the adaptive boosted Gaussian Naive bayes (ABGNB) algorithm for training and validating the HDDSS. The development of HDDSS consists of two main stages: In the first stage, the recursive feature elimination algorithm is applied to the UCI input dataset to determine the optimal heart disease input features. In the second step, the proposed ABGNB classifier train and validate the heart disease prediction model using selected inputs from the RFE algorithm. The results obtained from the ABGNB classification model are evaluated with other machine learning (ML) models such as Naïve bayes (NB), K-nearest neighbor (K-NN), SVM, and Decision tree (DT). For measuring the efficiency of the proposed model, different classification performance metrics [12] were used, namely, Classifier Accuracy, Misclassification rate, Sensitivity, Specificity, Precision, F-Score, Receiver operating characteristic curve. This proposed system's process flow diagram is shown in Fig. 1.

A. Dataset Description

This work uses a Cleveland heart disease dataset from the UCI repository; this is available online [13]. This dataset consists of data about 303 individuals (303 samples), 13 heart disease predictors, and one class attribute with binary outcomes as 1 (heart disease-Positive) and 0 (heart disease-Negative). Heart Disease-Positive indicates the patient has a heart disease problem, and heart disease-Negative implies the patient has no heart disease. The input dataset contains 164 samples of the positive class and 139 samples of the negative class. There are no missing values in this dataset.

B. Recursive Feature Elimination

Feature selection is one of the data preprocessing procedures for identifying and selecting the features most associated with the output variable. The feature selection step is necessary for this proposed

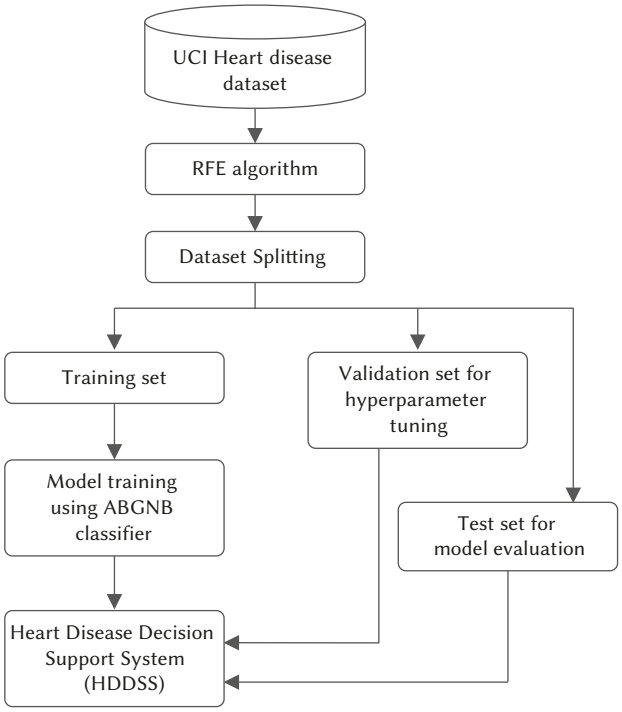


Fig. 1. The proposed HDDSS.

model for choosing relevant input features from the dataset by dropping irrelevant, redundant, noisy features [14], [15]. Keeping irrelevant features in our dataset may reduce the outcomes of the machine learning model.

For feature selection, the proposed work uses recursive feature elimination (RFE) [16]. The RFE algorithm is a recursive method to find out the statistical significance of the features. The statistical significance has calculated using criteria called hypothesis testing.

Hypothesis testing uses a p-value which is an observed significance value of input feature and it is a probabilistic measure to evaluate the hypothesis [17]. The statistical relationship exists between the input and output feature if the p-value of an input feature is less than the significance threshold (α). RFE uses 0.05 as the threshold value (α) [18]. The RFE algorithm begins with the full feature set D_s consisting of input features p_1, p_2, ... p_N, and then recursively prunes irrelevant features based on the hypothesis statement given in equations (1) & (2) at each iteration until the p-value of features is smaller than the threshold value (α). Fig. 2 depicts the steps of the RFE algorithm.

RFE uses two kinds of hypothesis called the null hypothesis and alternate hypothesis for selecting optimal features of input dataset D_s. Statements of a null hypothesis and alternate hypothesis are explained below:

Null Hypothesis: This hypothesis states that there is no relationship between the selected input feature and output feature when the p-value of the selected input feature is greater than or equal to the threshold value. According to this assumption, the input feature with a p-value greater than the threshold level is eliminated. In Equation (1), this null hypothesis statement is given as:

$$H_0: \mu \geq \alpha \tag{1}$$

Alternative Hypothesis: This hypothesis states that a strong relationship exists between the input feature and the output feature when the p-value of the input feature is less than the threshold values. According to this assumption, the input feature with a p-value lower than the threshold level is selected. In the following Equation (2), the alternative hypothesis is given as:

$$H_a: \mu < \alpha \tag{2}$$

Where H_0 is a Null Hypothesis, H_a is an Alternate Hypothesis, μ is a p-value of an input feature, and α is the threshold value. The RFE algorithm utilizes logistic regression to find the p-value of the input features to prove the alternative hypothesis claim by rejecting the null hypothesis statement. Logistic regression uses a logit function that is a form of statistical model [19] to determine the relationship between the selected input features and the output features by measuring the logarithm of odds as in Equation (3).

$$logit(p_i) = log\left(\frac{prob}{1-prob}\right) = \beta_0 + \beta_i p_i \tag{3}$$

Where *prob* is the probability of selected input features, p_i is the input feature, logistic regression parameters are β_0 and β_i. Algorithm 1 illustrates the RFE algorithm.

Algorithm 1. Recursive Feature Elimination

Input: data set D_s which consists of N training samples $D_s = ((p_1, q_1), (p_2, q_2), ..., (p_N, q_N))$ and $p_i \in P$; $q_i \in Q$ are the corresponding class labels of D_s associated with p_i. Value of $Q \in \{1, 0\}$.

1: Assign Threshold value = 0.05.

2: State the Null Hypothesis and Alternate Hypothesis.

3: Load the dataset D_s with all input features.

4: Calculate the *p*-value of each input feature using logit function.

5: Reject the alternative hypothesis if the *p*-value of the selected input feature is greater than or equal to α, and remove that feature from full feature set D_s.

6: Iterate the step 4-step 5 till getting the significant features with p-value lower than α.

Output: $D_{relevant}$ → Selected input features of D_s for training and validating the ABGNB classifier.

C. Classification Model Using Adaptive Boosted Gaussian Naïve Bayes Algorithm

Classification modeling is the next step in this proposed method. This process starts with a target dataset, which contains relevant input features obtained using the RFE algorithm. The dataset has been divided into three segments: 80% data for the training phase, 10% for the validation phase, and 10% for the testing phase. This classification modeling has two stages, which are the model training phase and the model validation phase. The ABGNB algorithm is needed to train the HDDSS. The training phase entails developing an HDDSS via learning the training algorithm (ABGNB) parameters and training dataset. Model validation is the second phase of this classification model; training results are evaluated during this phase using a validation dataset for tuning the AGBNB classifier hyper-parameters for improving efficiency and minimize the loss function of the ABGNB classifier. A test set is used to evaluate the final prediction model's working capacity using different classifier performance metrics.

1. Training Phase of ABGNB Classifier

During the training phase, an HDDSS has been developed using an ABGNB classifier. ABGNB is an ensemble of the Adaboost algorithm and Gaussian naive bayes, which outperforms conventional machine learning algorithms in prediction accuracy [20], [21]. The proposed ABGNB classifier utilizes the Adaboost algorithm to improve the prediction efficiency of the gaussian naive bayes classifier.

The Adaboost model has been trained with bootstrapped samples of the training dataset and the gaussian naive bayes algorithm. Bootstrap

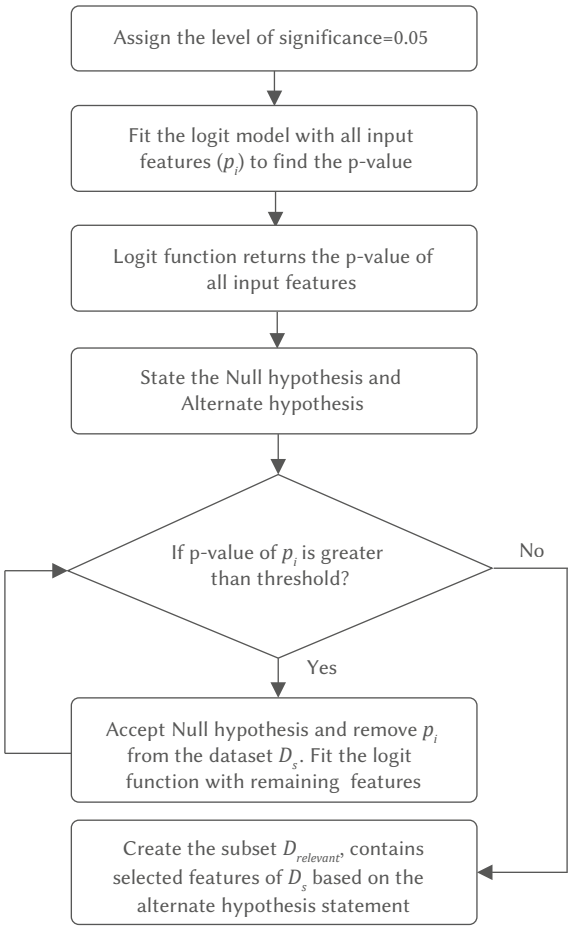


Fig. 2. Workflow of Recursive Feature Elimination Algorithm.

sampling avoids overfitting and improves the performance of training algorithms [22]. The Adaboost methodology aims to provide a correct prediction for each training instance in each iteration by training the gaussian naive Bayes classifier with differently weighted training samples.

During the first iteration of the training process, all samples in the training set are assigned the same weight, and the accuracy of the training phase will be measured after the first iteration. The weight of the misclassified samples is updated during the second iteration of the training phase to motivate the incorrect prediction in the training set, and the process continues for each iteration (s_i) of the training phase. Consequently, the Adaboost classifier constructs a linear hypothesis $h_t: P \rightarrow Q$ by the ensemble of weak hypothesis generated in iteration s_i. The resulting linear hypothesis h_t minimizes the misclassification rate by correctly classifying the given sample p_i according to the class label q_i Adaboost generates a final hypothesis by linearly combining the weak hypothesis $h_1, h_2, \ldots h_t$ for T steps and minimizing the weighted error of all training samples.

In the ABGNB framework, Gaussian Naïve Bayes (GNB) algorithm is used as a base estimator in the Adaboost classifier to calculate the class membership probability of an input sample ($Prob(p_i \mid q)$) using Gaussian probability density function which is given in the Equation (4).

$$GPDF(p_i, \mu_{classi}, \sigma_{classi}) = \frac{1}{\sigma\sqrt{2\pi}}\, exp^{-\frac{(p_i-\mu)^2}{2\sigma^2}} \tag{4}$$

So that $Prob(p_i \mid q)$ is calculated as in Equation (5)

$$Prob(p_i \mid q) = GPDF(p_i, \mu_{classi}, \sigma_{classi}) \tag{5}$$

Where $GPDF(p_i, \mu_{classi}, \sigma_{classi})$ is the gaussian probability density function of an input sample p_i; π is the mathematical constant value, σ and μ are the standard deviation and mean value of input features for each class label, exp is the mathematical constant. μ_{classi} is a mean value of input features of each class label that can be calculated using Equation (6).

$$\mu_{classi} = {}^1/_N * \sum_{i=1}^{N} p_i \tag{6}$$

Where N is the total training samples and p_i is the input value of training dataset D. σ_{classi} is a standard deviation value of each class label that can be calculated using Equation (7).

$$\sigma_{classi} = \sqrt{\frac{\sum(p_i - \mu)^2}{N}} \tag{7}$$

This ensemble of gaussian naïve bayes and Adaboost classifier (ABGNB) has more advantages as an increase in prediction accuracy and reducing overfitting problem over traditional ML algorithms. Fig.3 shows the graphical illustration of the ABGNB classifier.

2. Hyperparameter Tuning Phase of ABGNB Classifier

It is a validation phase used for tuning the hyperparameters of the ABGNB algorithm. The ABGNB hyperparameters are tuned using the grid search optimizer [23]. This grid search optimizer selects the best hyperparameter values from the hyperparameter search space. Hyperparameters are parameters used by machine learning classifiers to monitor and regulate the classifier's learning process. Tuning the hyperparameter of the classifier helps to improve the classifier prediction accuracy [24]. Grid search builds and evaluates a model for every combination of hyperparameters provided. Grid search finds the best ABGNB hyperparameter from the hyperparameter search space, and then the model is retrained with the new parameters. The validation dataset is used to measure the model's accuracy after the hyperparameters are tuned.

The generic hyperparameter statement of ML classifiers are defined below: Consider y as an ML algorithm with a M hyper-parameters (H). The hyperparameter search space of ML classifier is denoted as $H = \mathrm{H}\, H_1 \times H_2 \ldots . \times H_M$. A grid search method was used to optimize the ABGNB classifier's accuracy in this proposed HDDSS, as shown in Equation (8).

$$ABGNB\ Perf = argmax_p\, f(ABGNB, H, D_v) \tag{8}$$

Where *ABGNB Perf* returns the set of optimised hyperparameters which maximise ABGNB classifier efficiency, H denotes the hyperparameters of ABGNB classifier, D_v denotes the validation dataset, $argmax_p f$ is the grid search optimization function on *ABGNB*, H, D_v to maximise the accuracy score of training and hyperparameter tuning phase. Algorithm 2 describes the steps of the ABGNB algorithm in detail.

IV. Experimental Results and Discussions

The various experimental analysis is performed on the proposed HDDSS to demonstrate the efficiency of the Adaptive Boosted Gaussian Naïve Bayes Classifier with Recursive Feature Elimination.

A. Characteristics of Input Dataset

The proposed ABGNB and other ML classifiers are trained and validated using the UCI heart disease dataset. The characteristics of the input dataset are specified in Table I. The input features of the UCI dataset are mentioned in Table II.

Algorithm 2. Adaptive boosted Gaussian Naïve Bayes

ABGNB Input: Training data set D_s which consists of N training samples $D_s = ((p_1, q_1), (p_2, q_2), ..., (p_N, q_N))$ and $p_i \in P$; $q_i \in Q$ are the corresponding class labels of D_s associated with p_i. Value of $Q \in \{1,0\}$.

ABGNB Parameters:

- *base_learner*: Gaussian Naïve Bayes is used to train the Adaboost Classifier.
- $t = 1, 2, ..., T$ iterations
- *est_gnb*: Number of weak learners to be generated in each iteration
- *lr*: learning rate

ABGNB Output: $hyp_{op}(p)$ is a final hypothesis with improved classification performance.

1: Load $D_s = ((p_1, q_1), (p_2, q_2), ..., (p_N, q_N))$

2: Initialize weight of data sample (p_i):

$$w(p_i) = \frac{1}{N}$$

for all p_i, $i = 1, 2, ..., N$; N is a total training samples

3: for t = 1 to T do

4: for *est_gnb* =1 to T do

5: Generate a vector r with initial weight w_i

6: Apply bootstrap sampling on D_s to create a subset called s_m

7: Calculate likelihood of feature subset using Gaussian probability density function called $GPDF(s_m, r)$

8: Build weak hypothesis of gaussian naïve bayes model $hyp_{gnb}(p_i)$ using majority voting scheme

9: end *est_gnb*

10: Get Weak hypothesis $hyp_{gnb}(p_i) \rightarrow \{1, 0\}$ with error rate

$$e_w = \frac{sum\left(w_i * terror(i)\right)}{sum(w_i)}$$

where e_w is a weighted sum of an error rate, w_i is the weight for each training sample i, *terror* is the prediction error for training sample i

11: Update the weight of incorrect samples for $i = 1, 2, ..., N$ in each subset s_m

$$s_{m+1}(i) = \frac{w_i * exp\left(-a_w q_i hyp_{gnb}(p_i)\right)}{z_w}$$

where w_i is the weight of specific training sample, z_w is the normalization constant; a_w is the parameter to increase the generalization of abgnb classifier

12: end for t

13: Output the final hypothesis $hyp_{op}(p)$:

$$hyp_{op}(p) = \text{SIGN}\left(\sum_{1}^{T} a_w hyp_{gnb}(p)\right)$$

14: Calculate *training error* of ABGNB using D_v where D_v is a validation dataset;

15: Use grid search for selecting optimal hyperparameter of ABGNB from hyperparameter search space and retrain the model with optimized parameters.

16: If Validation error > Training error, Stop the retrain

17: Calculate performance of final model $hyp_{op}(p)$ using test dataset

TABLE I. Characteristics of UCI Heart Disease Dataset

Dataset	Number of input Attributes	Number of Classes in output attribute	Number of Samples
UCI heart disease dataset	13	2	303

TABLE II. Input Features of UCI Heart Disease Dataset

S.No	Feature Code	Description of features
1	AGE	The individuals' age
2	GEN	The gender of an individual
3	CP	The chest pain type of an individual
4	RBP	The resting blood pressure value
5	CHOL	The serum cholesterol
6	FBS	An individual's fasting blood sugar value
7	RESTECG	ECG resting value
8	MAXHR	Maximum heart rate achieved
9	EIA	Exercise included angina
10	OPK	Old Peak Value
11	PESS	Peak exercise ST segment
12	CF	Number of major vessels colored by fluoroscopy
13	THAL	The thalassemia

B. Performance Evaluation Measures of the Proposed Model

This HDDSS uses many classification performance metrics. Almost all evaluation measures of this proposed work are based on a Confusion matrix. This matrix assesses the classifier performance via four components named True Positive (*tp*), True Negative (*tn*), False Positive (*fp*), and False Negative (*fn*). True positive is a correctly labeled positive sample, True negative are the correctly labeled negative samples, False positive are falsely labeled negative samples, and False Negative is falsely labeled positive sample. The components of the confusion matrix (*cm*) have given in Equation (9) below.

$$cm = \begin{bmatrix} tp & fp \\ fn & tn \end{bmatrix} \tag{9}$$

Equations (10)-(17) define seven classification measures for evaluating the HDDSS model. Classifier Accuracy (Acc) is the overall effectiveness of the classifier. Misclassification rate (MCR) is the total number of incorrect predictions in the training sample. Sensitivity (Sen) refers to the number of positives that were supposed to be positive. Specificity (Spe) is the percentage of samples correctly labeled as a negative compared to the total negative samples. The number of true positives divided by the total number of true positives and false positives equals precision (Pre). F- Score (FS) is the weighted harmonic mean of the test's sensitivity and precision.

$$Acc = \frac{tp + tn}{tp + tn + fp + fn} \tag{10}$$

$$MCR = \frac{fp + fn}{tp + tn + fp + fn} \tag{11}$$

$$Sen = \frac{tp}{tp + fn} \tag{12}$$

$$Spe = \frac{tn}{tn + fp} \tag{13}$$

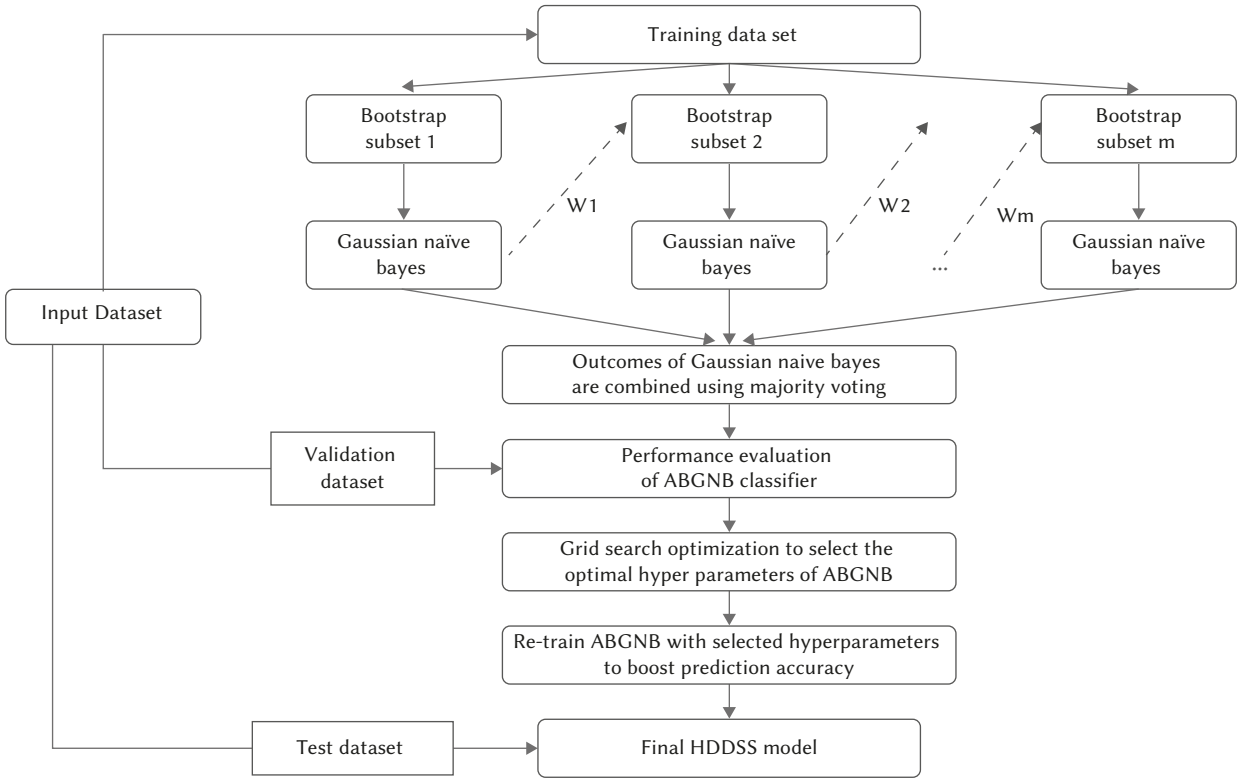


Fig. 3. Graphical Representation of Adaptive Boosted Gaussian Naïve Bayes Algorithm.

$$Pre = \frac{tp}{tp + fp} \tag{14}$$

$$FS = \frac{2}{\frac{1}{Recall} + \frac{1}{Pre}} \tag{15}$$

The recall is known as sensitivity in equation (15). The receiver operating characteristic curve (ROC) has been evaluated using the True positive rate (*tpr*) and false-positive rate (*fpr*). The best classifier shows a ROC value of 1, and the worst classifier shows a ROC value below 0.5. The equation of ROC has given below:

$$tpr = \frac{tp}{tp + fn} \tag{16}$$

$$fpr = \frac{fp}{fp + tn} \tag{17}$$

C. Experimental Result of Recursive Feature Elimination

HDDSS uses the Recursive Feature Elimination (RFE) for selecting the relevant input features for predicting heart disease. This algorithm uses a threshold value (alpha value) of 0.05 to choose the relevant input features of the heart disease dataset. The alpha value has been compared to the p-value of the input feature. If the p-value of the input attribute is less than the alpha value, then it is considered an optimal feature for predicting heart disease, so RFE selects it; otherwise, it will reject. Table III displays the p-values of the input features of the heart disease dataset. This table illustrates how the RFE algorithm selects optimal input features of the UCI heart disease dataset based on the p-value of the input features.

RFE selected seven input features from the heart disease dataset for classification modeling based on their p-values. Table IV displays the selected input features of the input dataset based on the RFE.

TABLE III. The RFE Algorithm Selects Optimal Input Features of the UCI Heart Disease Dataset Based on the P-Value of the Input Features

Feature Code	p-value	Compared with alpha -value (0.05)	Select / Reject the Feature
AGE	0.832	Greater than alpha	Rejected
GEN	0.000	Less than alpha	Selected
CP	0.000	Less than alpha	Selected
RBP	0.060	Greater than alpha	Rejected
CHOL	0.221	Greater than alpha	Rejected
FBS	0.947	Greater than alpha	Rejected
RESTECG	0.181	Greater than alpha	Rejected
MAXHR	0.026	Less than alpha	Selected
EIA	0.017	Less than alpha	Selected
OPK	0.012	Less than alpha	Selected
PESS	0.098	Greater than alpha	Rejected
CF	0.000	Less than alpha	Selected
THAL	0.002	Less than alpha	Selected

The efficiency of the proposed RFE feature selection algorithm has been compared to other feature selection methods, such as sequential forward selection (SFS) [25], sequential backward elimination (SBE) [25], univariate feature selection (UFS) [26]. Table V shows the feature selection parameters, such as the objective function and the number of features selected, for different feature selection methods like RFE, SFS, SBE, and UFS algorithms. All of the above feature selection methods are used for choosing the best feature from a heart disease dataset, and the selected attributes are utilized to build an HDDSS; Table VI displays an output assessment of the RFE with other ML feature selection methods. It is evident from Table VI that the suggested RFE performs better than the other ML feature selection algorithms for the UCI heart disease dataset.

TABLE IV. Optimal Input Features of a Heart Disease Dataset Based on the RFE Algorithm

Dataset	Input features selected by RFE for developing HDDSS
UCI heart disease dataset	GEN, CP, MAXHR, EIA, OPK, CF, THAL

TABLE V. Parameter Setting of Various Machine Learning Feature Selection Algorithms

Feature selection Algorithm	Parameter	Value
SFS	Objective function	Gini Index
		Entropy
	Number of features selected	7
SBE	Objective function	Gini Index
		Entropy
	Number of features selected	7
UFS	Objective function	Chi-square
	Number of features selected	7
Proposed RFE	Objective function	Logistic regression
	Number of features selected	7

TABLE VI. Output Assessment of the RFE With Other ML Feature Selection Methods for Heart Disease Dataset Utilizing the Proposed ABGNB Classifier

Feature Selection	Acc	MCR	Sen	Spe	Pre	FS	ROC
SFS + Gini Index	75.18	24.82	75.62	80.63	75.5	74.94	75.59
SBE + Gini Index	83.92	16.08	83.61	80.35	83.4	84.74	83.19
SFS + Entropy	80.49	19.51	80.3	83.46	80.63	79	80.32
SBE + Entropy	80.34	19.66	80.56	73.14	80.67	81.64	80.64
UFS + Chi-square	88.49	11.51	88.76	86.31	85.01	88.34	88.69
Proposed RFE	92.87	7.13	93.45	90.76	91.64	92.08	92.42

D. Experimental Result of Classification Modeling

In this phase, the selected input features of the heart disease dataset fed into the ABGNB classifier along with the target feature. Repeated 10 × 5 stratified cross-validation is applied during the validation process to build a generalized classifier on an independent dataset and avoiding over-fitting problems. This process is where the 10-fold cross-validation has been repeated five times, in which the data samples being shuffled during each repetition, providing a different split of the given data. The grid search optimizer then adjusts the ABGNB classifier hyperparameters to improve the efficiency of the trained model. The ABGNB classifier's hyperparameter range is described in Table VII.

The classification performance of the ABGNB classifier is compared with other Machine Learning models, namely, Naïve bayes (NB) [27], K-Nearest neighbor (KNN) [28], Support vector machine (SVM) [29], Decision tree (DT) [30]. Table VIII shows the output of the proposed ABGNB classifier with other conventional ML classifiers on the heart disease dataset before using the RFE algorithm.

Table IX shows the output of the proposed ABGNB classifier compared to other ML classifiers on the heart disease dataset after implementing the RFE feature selection method. It's worth noting that the proposed ABGNB classifier performs well on the heart disease dataset and has a high accuracy score of 92.87%.

TABLE VII. Hyper Parameter Search Space of ABGNB Classifier

Proposed Classifier	Hyperparameter	Hyperparameter configuration space	Selected Hyper parameters by grid search
ABGNB Classifier	No. of weak learner	[10, 50, 100, 500]	500
	Learning rate	[0.0001, 0.001, 0.01, 0.1, 1.0]	0.1
	Random state	[50,30,40]	40

TABLE VIII. Performance Comparison of ABGNB Classifier With Other ML Algorithms on Heart Disease Dataset Before Using the Suggested Recursive Feature Selection

Classifier	Acc	MCR	Sen	Spe	Pre	FS	ROC
NB	86.31	13.69	80.52	93.31	92.45	86.74	86.41
KNN	75.23	24.77	80.41	70.54	73.15	76.75	75.78
SVM	70.42	29.58	87.27	53.97	65.39	75.04	70.57
Decision Tree	80.49	19.51	83.65	76.48	78.17	81.37	80.37
Proposed ABGNB	90.12	9.88	87.56	93.14	93.45	90.67	90.47

TABLE IX. Performance Comparison of Proposed ABGNB Classifier With Other Typical ML Algorithms on Heart Disease Dataset After Applying the Suggested Recursive Feature Selection

Classifier	Acc	MCR	Sen	Spe	Pre	FS	ROC
NB	90.5	9.5	85.86	95.17	95.49	90.34	80.19
KNN	80.31	19.69	85.17	74.5	78.9	82.04	60.14
SVM	73.16	26.84	87.06	58.76	70.37	77	48.37
Decision Tree	79.24	20.76	70.46	86.27	85.31	77.52	57.06
Proposed ABGNB	92.87	7.13	93.45	90.76	91.64	92.08	84.67

Table X compares the performance of the proposed HDDSS to the output of other heart disease prediction models. The purpose of this analysis (Table X) is to show how the proposed classifier, the adaptive boosted Gaussian Naive Bayes classifier, outperforms previous studies in terms of prediction accuracy.

TABLE X. The Output of the Proposed HDDSS Compared to Other Heart Disease Prediction Models Output

Author(s)	Year	Method	Highest Accuracy (in %)
Haq et al. [5]	2018	Relief + Logistic Regression	89
Otoom et al. [6]	2015	Best first search + BayesNet	84.5
Mohan et al [7]	2019	Hybrid random forest with a linear model	88.7
Li, Jian Ping et al [8]	2020	FCMIM -SVM	92.37
Chen et al. [9]	2011	Learning vector quantization	80
Hidayet et al. [10]	2018	Linear kernel SVM + ReliefF	84.81
Thippa Reddy et al. [11]	2017	Rule Based Fuzzy Logic Model	78
David et al. [31]	2018	Random Forest	81
Das et al. [32]	2020	K-NN	86.84
Apurv Garg et al. [33]	2021	K-NN	86.88
Proposed Method	-	RFE + ABGNB	92.87

E. Discussions

The proposed research aimed to demonstrate that the RFE and ABGNB classifiers could reliably predict heart disease. The RFE is being used to select relevant input features from the heart disease dataset. RFE selects optimal features using the p-value of an input feature. Table III shows how the p-value of the input feature is utilized to choose the best input features. RFE selected seven relevant input features for building the proposed HDDSS, as shown in Table IV. The efficiency of the RFE has been compared to that of other feature selection methods such as sequential forward, sequential backward, and univariate feature selection in the proposed framework. The performance of these feature selection methods has been illustrated in Table VI. Table VI shows that the RFE is the best feature selection procedure for the UCI heart disease dataset, exceeding other feature selection strategies in identifying the best features from the input dataset. RFE outperformed other ML feature selection methods in terms of recall, precision, F-score, and ROC shows, resulting in improved classification performance. HDDSS has been developed using the proposed ABGNB classifier with the relevant input features and the target feature. The performance of the ABGNB has been compared to that of other machine learning models such as naive Bayes, KNN, SVM, and Decision tree classifiers. Tables VIII and IX illustrate that the proposed ABGNB classifier outperforms other ML classifiers before and after using the suggested RFE approach on the heart disease dataset. The prediction accuracy of the ABGNB algorithm was 90.12 percent before using the RFE algorithm and 92.87 percent after using the RFE algorithm, meaning that the RFE algorithm enhanced the ABGNB algorithm's prediction accuracy by 2 to 3 percentage points. Table IX confirms that the ABGNB classifier has the best accuracy of 92.87 percent, meaning that ABGNB can differentiate between positive and negative samples. Table IX also shows that the proposed ABGNB classifier outperforms other ML classifiers on different performance metrics for the heart disease dataset. Table X demonstrates that the proposed heart disease prediction model (HDDSS) outperforms most of the present literature for increasing prediction accuracy. According to this study, the proposed HDDSS outperformed other heart disease prediction models and is suitable for assessing the risk of heart disease in a patient.

V. Conclusion

This research aims to build an enhanced heart disease decision-support system for the prediction of heart disease. This automated diagnosis system has been experimented on the UCI heart disease dataset. This proposed HDDSS utilizes the Recursive feature elimination method for selecting the most relevant input features of the heart disease dataset. The ABGNB classifier ensembled with a grid search optimizer for enhancing the prediction accuracy of HDDSS. The experiment result illustrates that the RFE+ABGNB method gives better performance than other compared ML models on the heart disease dataset. This suggested method achieves 92.87% prediction accuracy with 93.45% sensitivity on the UCI heart disease dataset. The analysis of the proposed system showed that the proposed ABGNB with recursive feature elimination provides better heart disease prediction performance on the UCI heart disease dataset. The proposed model's efficiency can be improved even more using an automated regularization technique.

References

[1] Anand, Sonia S., Shofiqul Islam, Annika Rosengren, Maria Grazia Franzosi, Krisela Steyn, Afzal Hussein Yusufali, Matyas Keltai, Rafael Diaz, Sumathy Rangarajan, and Salim Yusuf, "Risk factors for myocardial infarction in women and men: insights from the INTERHEART study," European heart journal, vol. 29, no. 7, pp. 932-940, 2008.

[2] Frohlich, Edward D., and Patrick J. Quinlan, "Coronary heart disease risk factors: public impact of initial and later-announced risks," The Ochsner Journal, vol. 14, no. 4, 532-537, 2014.

[3] Wah, Teh Ying, Ram Gopal Raj, and Uzair Iqbal, "Automated diagnosis of coronary artery disease: a review and workflow," Cardiology research and practice, 2018.

[4] Ali, Liaqat, Atiqur Rahman, Aurangzeb Khan, Mingyi Zhou, Ashir Javeed, and Javed Ali Khan, "An Automated Diagnostic System for Heart Disease Prediction Based on χ^2 Statistical Model and Optimally Configured Deep Neural Network," IEEE Access, vol. 7, pp. 34938-34945, 2019.

[5] Haq, Amin Ul, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, and Ruinan Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," Mobile Information Systems, 2018.

[6] Otoom, Ahmed Fawzi, Emad E. Abdallah, Yousef Kilani, Ahmed Kefaye, and Mohammad Ashour, "Effective diagnosis and monitoring of heart disease," International Journal of Software Engineering and Its Applications, vol. 9, no. 1, pp. 143-156, 2015.

[7] Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, vol. 7, pp. 81542-81554, 2019.

[8] Li, Jian Ping, Amin Ul Haq, Salah Ud Din, Jalaluddin Khan, Asif Khan, and Abdus Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," IEEE Access, vol. 8, pp. 107562-107582, 2020.

[9] Chen, Austin H., Shu-Yi Huang, Pei-Shan Hong, Chieh-Hao Cheng, and En-Ju Lin. "HDPS: Heart disease prediction system." In 2011 computing in cardiology, IEEE, 2011, pp. 557-560.

[10] Takci, Hidayet, "Improvement of heart attack prediction by the feature selection methods," Turkish Journal of Electrical Engineering & Computer Sciences, vol. 26, no. 1, pp. 1-10, 2018.

[11] Reddy, G. Thippa, and Neelu Khare, "An efficient system for heart disease prediction using hybrid OFBAT with rule-based fuzzy logic model," Journal of Circuits, Systems and Computers, vol. 26, no. 04, 2017.

[12] Hossin, Mohammad, and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," International Journal of Data Mining & Knowledge Management Process, vol. 5, no. 2, 2015.

[13] Janosi A, Steinbrunn W, Pfisterer M, Detrano R. Heart Disease UCI. Available online: https://www.kaggle.com/ronitf/heart-disease-uci (Cited on 23 Jan 2021)

[14] Khalid, Samina, Tehmina Khalil, and Shamila Nasreen. "A survey of feature selection and feature extraction techniques in machine learning." In 2014 Science and Information Conference, IEEE, 2014, pp. 372-378.

[15] Miao, Jianyu, and Lingfeng Niu, "A survey on feature selection," Procedia Computer Science, vol. 91, pp. 919-926, 2016.

[16] Guyon, Isabelle, and André Elisseeff, "An introduction to variable and feature selection," Journal of machine learning research, vol. 3, no. Mar 2003, pp. 1157-1182, 2003.

[17] Massey, Adam, and Steven J. Miller. "Tests of hypotheses using statistics." Mathematics Department, Brown University, Providence, RI 2912 2006, pp. 1-32.

[18] Murphy, Kevin P. "A Probabilistic Perspective." Text book (2012).

[19] Hoffman, J. I. E. "Logistic regression." Basic Biostatistics for Medical and Biomedical Practitioners, 2019, pp. 581-589.

[20] Perez, Aritz, Pedro Larranaga, and Inaki Inza, "Supervised classification with conditional Gaussian networks: Increasing the structure complexity from naive Bayes," International Journal of Approximate Reasoning, vol. 43, no. 1, pp. 1-25, 2006.

[21] Freund, Yoav, and Robert E. Schapire. "Experiments with a new boosting algorithm." In icml, vol. 96, 1996, pp. 148-156.

[22] Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection." In Ijcai, vol. 14, no. 2, 1995, pp. 1137-1145.

[23] Syarif, Iwan, Adam Prugel-Bennett, and Gary Wills, "SVM parameter optimization using grid search and genetic algorithm to improve classification performance," Telkomnika, vol. 14, no. 4, pp. 1502-1509, 2016.

[24] Wu, Jia, Xiu-Yun Chen, Hao Zhang, Li-Dong Xiong, Hang Lei, and Si-Hao Deng, "Hyperparameter optimization for machine learning models based on Bayesian optimization," Journal of Electronic Science and Technology, vol. 17, no. 1, pp. 26-40, 2019.

[25] Kumar, Vipin, and Sonajharia Minz, "Feature selection: a literature review," SmartCR, vol. 4, no. 3, pp. 211-229, 2014.

[26] Jović, Alan, Karla Brkić, and Nikola Bogunović. "A review of feature selection methods with applications." In 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO), IEEE, 2015, pp. 1200-1205.

[27] Rish, Irina "An empirical study of the naive Bayes classifier." In IJCAI 2001 workshop on empirical methods in artificial intelligence, vol. 3, no. 22, 2001, pp. 41-46.

[28] Tan, Songbo, "An effective refinement strategy for KNN text classifier," Expert Systems with Applications, vol. 30, no. 2, pp. 290-298, 2006.

[29] Cortes, Corinna, and Vladimir Vapnik, "Support-vector networks," Machine learning, vol. 20, no. 3, pp. 273-297, 1995.

[30] Safavian, S. Rasoul, and David Landgrebe, "A survey of decision tree classifier methodology," IEEE transactions on systems, man, and cybernetics, vol. 21, no. 3, pp. 660-674, 1991.

[31] David, H. B. F., and Belcy, S. A., "Heart Disease Prediction Using Data Mining Techniques," ICTACT Journal on Soft Computing, vol. 9, no. 1, 2018.

[32] Das, Sumit, Manas Kumar Sanyal, and Sourav Kumar Upadhyay. "A Comparative Study for Prediction of Heart Diseases Using Machine Learning." Available at SSRN 3526776, 2020.

[33] Garg, A., Sharma, B. and Khan, R., "Heart disease prediction using machine learning techniques." In IOP Conference Series: Materials Science and Engineering, IOP Publishing, 2021: Vol. 1022, No. 1, p. 012046.

M. Dhilsath Fathima

She received her B.E degree in CSE from Anna University, Tamil Nadu, India in 2005, Master's degree in CSE from Sathyabama Institute of Science & Technology, Tamil Nadu, India in 2011. Now, pursuing Ph.D in Computer science and Engineering Department from Sathyabama Institute of Science & Technology, Tamil Nadu, India. Since 2007, she is working as an Assistant Professor in Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Tamil Nadu, India. Her major research areas are Machine learning, Data mining, Deep learning etc.

Dr. S. Justin Samuel

Dr. S.Justin Samuel received his Ph.D in Computer Science and Engineering from Sathyabama University, India. His area of interest includes Data Mining, Wireless Sensor networks, and Image processing. He has published more than 25 research papers in International & National Journals and Conferences. He is a professor at PSN Engineering college for Department of Science and Technology, India.

S. P. Raja

S. P. Raja was born in Sathankulam, Tuticorin District, Tamilnadu, India. He completed his schooling in Sacred Heart Higher Secondary School, Sathankulam, Tuticorin, Tamilnadu, India. He completed his B. Tech in Information Technology in the year 2007 from Dr. Sivanthi Aditanar College of Engineering, Tiruchendur. He completed his M.E. in Computer Science and Engineering in the year 2010 from Manonmaniam Sundaranar University, Tirunelveli. He completed his Ph.D. in the year 2016 in the area of Image processing from Manonmaniam Sundaranar University, Tirunelveli. His area of interest is image processing and cryptography. He has more than 14 years of teaching experience in engineering colleges. He currently works as an Associate Professor in the Department of Computer Science and Engineering at the Vellore Institute of Technology, Vellore. He has published 48 papers in International Journals, 24 in International conferences and 12 in national conferences. He is an Associate Editor of the International Journal of Interactive Multimedia and Artificial Intelligence, International Journal of Image and Graphics and International Journal of Biometrics.

ResNet18 Supported Inspection of Tuberculosis in Chest Radiographs With Integrated Deep, LBP, and DWT Features

Venkatesan Rajinikanth[1], Seifedine Kadry[2,3,4], Pablo Moreno Ger[5*]

[1] Department of Computer Science and Engineering, Division of Research and Innovation, Saveetha School of Engineering, SIMATS, Chennai 602105 (India)
[2] Department of Applied Data Science, Noroff University College, Kristiansand (Norway)
[3] Artificial Intelligence Research Center (AIRC), Ajman University, Ajman, 346 (United Arab Emirates)
[4] Department of Electrical and Computer Engineering, Lebanese American University, Byblos (Lebanon)
[5] School of Engineering and Technology, Universidad Internacional de La Rioja (UNIR), Logroño (Spain)

Received 1 September 2022 | Accepted 28 April 2023 | Published 19 May 2023

Abstract

The lung is a vital organ in human physiology and disease in lung causes various health issues. The acute disease in lung is a medical emergency and hence several methods are developed and implemented to detect the lung abnormality. Tuberculosis (TB) is one of the common lung disease and premature diagnosis and treatment is necessary to cure the disease with appropriate medication. Clinical level assessment of TB is commonly performed with chest radiographs (X-ray) and the recorded images are then examined to identify TB and its harshness. This research proposes a TB detection framework using integrated optimal deep and handcrafted features. The different stages of this work include (i) X-ray collection and processing, (ii) Pre-trained Deep-Learning (PDL) scheme-based feature mining, (iii) Feature extraction with Local Binary Pattern (LBP) and Discrete Wavelet Transform (DWT), (iv) Feature optimization with Firefly-Algorithm, (v) Feature ranking and serial concatenation, and (vi) Classification by means of a 5-fold cross confirmation. The result of this study validates that, the ResNet18 scheme helps to achieve a better accuracy with SoftMax (95.2%) classifier and Decision Tree Classifier (99%) with deep and concatenated features, respectively. Further, overall performance of Decision Tree is better compared to other classifiers.

Keywords

Algorithms, Classification, Deep Learning, Radiographs Tuberculosis.

DOI: 10.9781/ijimai.2023.05.004

I. Introduction

LUNGS are one of the chief internal organs in human physiology and infectious and acute disease in the lung will always lead to a medical emergency. Even though several preventive actions are taken, the occurrence rate of the infectious disease is steadily increasing due to different causes and appropriate recognition and treatment will help in reducing the impact of the syndrome and its spread.

An infection in the lungs will cause a severe respiratory problemand untreated lung disease causes increased morbidity and mortality. Tuberculosis (TB) is a communicable disease in the lungs, which causes a severe diagnostic burden globally. The occurrence rate of TB in a person is closely associated with the immune system and a weaker immune system will increase the chance of having TB [1] - [4].

TB is a communicable illness, which is initiated by a bacteria named Mycobacterium-Tuberculosis and the bacteria will spread through the air from an infected person. The discussion by Zaman [5] points to TB as one of the earliest diseases in mankind and along with the lung, it will affect sensitive organs, such as the brain, kidneys, spine, and intestines. TB easily affects people having a lower immune system and the happening rate of TB is progressively raising in low- and middle-income countries and causes a severe diagnostic burden. The report of World-Health-Organization (WHO) lists TB among one of the top 10 causes of death and hence several awareness and vaccination programs are regularly conducted to reduce TB. This report points out that, around 10 million global populations are infected with TB and this infection is small in children (12%) compared to adults and the elderly. Further, TB is found more in men (56%) compared to women (32%) and appropriate prevention and medication are necessary to control the occurrence rate and spread of TB [6].

The infection level of TB in a person is primarily tested using; (i) TB skin test and (ii) TB blood test [7], [8]. When the above tests confirm the risk of TB, the person is then diagnosed with bio-medical imaging procedures, such as Computed-Tomography (CT) scan and Chest Radiograph (X-ray) scan. Chest X-ray is widely preferred in detecting TB infection compared to CT because of its simplicity, reputation, and

* Corresponding author.
E-mail address: pablo.moreno@unir.net

cost. The recorded X-ray slices are then inspected by an experienced pulmonologist and based on the detected infection score; the doctor will plan for necessary medication/treatment to cure the patient. In recent years computerized diagnostic procedures are widely employed to support doctors in detection, decision making, and treatment planning processes. Further, the employment of Artificial Intelligence (AI) and Deep-Learning (DL) schemes helped in increasing the diagnostic accuracy and speed of computerized methods employed in hospitals [9]-[11].

The recent literature confirms the implementation of pre-trained/customary Deep-Learning-Procedure (DLP) to detect disease in various bio-medical image modalities [12]-[14]. Pre-trained schemes are widely employed due to their simplicity, readiness, and adaptability and a number of X-ray detection with pre-trained DLP confirmed its merit in accurately detecting the disease with appropriate clinical significance.

This research aims to develop a TB detection framework to inspect X-ray slices with improved accuracy. The proposed scheme consists of the following phases: (i) Image collection and initial processing, (ii) Employment of DLP to mine the deep features, (iii) Image enhancement and feature extraction using weighted Local Binary Pattern (LBP) and Discrete Wavelet Transform (DWT), (iv) Feature optimization using Firefly-Algorithm (FA), (v) Feature ranking and serial feature concatenation and (vi) Binary classification and validation.

This framework considered pre-trained DLP, such as AlexNet, VGG16, VGG19, ResNet18, ResNet50 and ResNet101 for the demonstration [15]-[18]. An experimental investigation is performed using X-ray images considered in the research work of Rahman et al. [19] and this experimental investigation confirms that ResNet18 with SoftMax helped to get better TB detection accuracy (95.2%) with deep features. Further, the ResNet18 scheme executed with a Decision Tree (DT) classifier helped to achieve a TB detection accuracy of 99% with integrated deep, LBP and DWT features. This research confirms that the pre-trained ResNet18 along with the proposed features helped to acquire an improved TB screening on the adopted image datasets.

The main contributions of this research include:

(i) Assessment of pre-trained DLP on chosen image database

(ii) Firefly algorithm supported feature optimization

(iii) Validating the significance of the proposed technique with similar existing works.

The remaining sections of this study are structured as follows: Section II presents the motivation behind this research, Sections III and IV present the context and methodology. Section V presents experimental results and discussions, and the conclusion of our research is presented in Section VI.

II. Motivation

In the literature, a considerable amount of TB screening measures is presented using pre-trained/customary DLP and every method helps to get an appropriate result.

The proposed research is motivated by the recent works of Rahman et al. [19] and Rajakumar et al. [20] who invented a pre-trained DLP to examine the TB in X-ray images and attained considerable accuracy. Particularly, the work of Rahman et al. [19] presented two schemes; (i) Without segmentation and (ii) With segmentation and the overall detection performance of the TB detection framework is claimed to be superior compared to alternatives.

Employment of Convolutional Neural Network (CNN) segmentation and classification is a challenging as well as computationally complex task even though the pre-trained CNN schemes are adopted. The recent literature claims that the integration of the deep and handcrafted features helps to achieve better disease detection irrespective of the image modalities [21].

Hence, in this research, a Disease Detection Framework (DDF) is proposed to improve TB detection accuracy using pre-trained DLP with integrated features.

III. Related Work

The development of clinically significant DDF is necessary for accurate recognition and analysis of TB from X-ray images. In the literature, a number of DDF are presented using various AI and DLP techniques to recognize TB.

TB detection with a chosen AI/DLP involves in; clinical grade image collection and processing, DDF implementation for disease diagnosis, sharing the findings to the pulmonologist for decision making, treatment implementation and curing the patient. In this work, DDF development and implementation play a major role in automatic TB detection systems and the findings of this scheme can also be preserved to create disease models to improve the TB diagnosis process. Some chosen earlier DDFs proposed to identify the TB in X-ray are summarized in Table I.

TABLE I. Outline of Recent TB Detection Methods Using X-ray

Reference	Procedure	Database	Findings
Afzali et al. [22]	Contour-supported shape descriptor method to distinguish the TB in X-ray	Montgomery dataset [23]	Accuracy 92.86%
Hijazi et al. [24]	Integrated Canny edge recognition and DLP is discussed to detect TB.	Montgomery and Shenzhen datasets [23]	Accuracy 89.77%
Hooda et al. [25]	DLP with seven convolutional and three fully connected layers is employed to detect TB and achieved 94.73% overall accuracy.	Montgomery and Shenzhen datasets [23]	Accuracy 94.73%
Rahman et al. [19]	Implementation of DLP with and without lung segmentation is discussed.	IEEE-Dataport [26]	Accuracy 96.47%
Rajakumar et al. [20]	A dual deep scheme with integrated VGG16 and VGG19 is presented to detect the TB in X-ray pictures.	Montgomery and Shenzhen datasets [23]	Accuracy 98.60% (with CNN segmentation)
Rohilla et al. [27]	Detection of TB using AlexNet and VGG16 are presented.	Montgomery and Shenzhen datasets [23]	Accuracy 97.25%
Kadry et al. [28]	VGG-UNet based joint segmentation and classification using deep and handcrafted features optimized with spotted hyena algorithm.	IEEE-Dataport [26]	Accuracy 99.22%
Ramya et al. [29]	VGG19 based TB detection using seagull-algorithm.	IEEE-Dataport [26]	Accuracy 98.62%

Earlier works in the literature implemented customary/pre-trained DLP to inspect the X-ray imagery related to TB using Montgomery dataset [23], Shenzhen dataset [23] and IEEE-Dataport [26] and every method discussed its merit on the considered database. The work of Kadry et al. [28] implements VGG-UNet scheme and achieves a detection accuracy of 99.22%. This approach is complex and time consuming compared to the proposed work, since it implemented segmentation and the classification task. The proposed research initially implemented the pre-trained DLP and tested its performance

using SoftMax with a 5-fold cross-validation and confirmed that the ResNet18 provides better accuracy on the considered database.

After confirming the performance with deep features, the handcrafted features like LBP and DWT are then extracted from the test images and are then reduced using the FA. The reduced deep, LBP and DWT features are then successively concatenated, and the obtained feature vector is then considered to test the performance of the proposed DDF with SoftMax and other binary classifiers found in the literature [30]-[33].

The experimental outcome of this research confirms that ResNet18 with DT offers better TB detection compared to other methods.

IV. Methodology

Computerized disease screening is widely preferred in hospitals to decrease the investigative burden of doctors [34]-[36]. In most cases, computerized approaches are operated by skilled workers under the supervision of the doctors and the result (report) generated by the computer algorithm is shared with the doctor along with the patient information [37]-[39]. The doctor verifies the report and plans the necessary treatment procedure for the patient.

This work proposes a DLP-based DDF to distinguish TB in X-ray images and to improve TB detection accuracy, the image enhancement methods, such as LBP and DWT are adopted and integrated with the DLP. Fig. 1 depicts the basic framework proposed to examine the X-ray images of normal and TB class using the pre-trained ResNet18 scheme. Implementation of the pre-trained DLP-based TB detection is implemented using Python® software and other works, such as feature extraction (LBP and DWT) and FA-based optimization are implemented in Matlab® and the outcomes are then evaluated in Python's workspace.

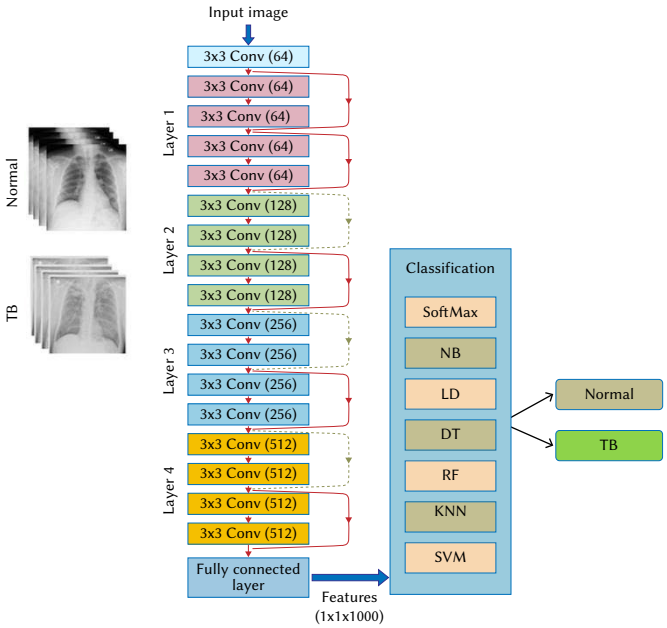


Fig. 1. TB detection framework with ResNet18.

In this work, the required test images of Normal/TB class are collected from the benchmark datasets, and then every image is resized to 230x230x3 pixels. After the resizing, the test image database is separated into two sections, forming training and validation sets. The training set is initially considered to train the pre-trained models with a chosen image dimension (227x227x3 pixels for AlexNet and 224x224x3 pixels for the remaining DLP).

After the training, the performance of DDF is then verified using the validation images with a 5-fold cross-validation process and the necessary metrics, such as Accuracy (ACC), Misclassification (MIC), Precision (PRE), Sensitivity (SEN), Specificity (SPE), F1-Score (F1S) and Negative Predictive Value (NPV). Initially, the performance of the proposed scheme is tested using SoftMax and then, other well-known classifiers, such as Naïve-Bayes (NB), Linear-Discriminant analysis (LD), Decision-Tree (DT), Random-Forest (RF), K-Nearest Neighbor (KNN) and linear kernel Support-Vector-Machine (SVM) are considered.

This procedure is initially implemented using only the deep features and then repeated using FA optimized Deep, LBP and DWT feature set and the results are presented and discussed.

A. Image Database

The necessary test images for this research are acquired from the dataset by Rahman et al. [19] found in [26]. It consists of 7000 images (3500 TB and 3500 normal) of high clinical grade.

In this work, the test images which are having cropped sections are ignored and only 5000 images (2500 TB and 2500 normal) alone are considered for experimental investigation.

Before using the test images, every image is resized into 230x230x3 pixels, in which 90% (2250 images) images are considered for training the DLP and 10% (250 images) are utilized for validation. Further, the combined training and validation images are considered to test the performance of this scheme with 5-fold cross validation to prevent overfitting. Some sample test images of the dataset considered in this work are presented in Fig. 2.

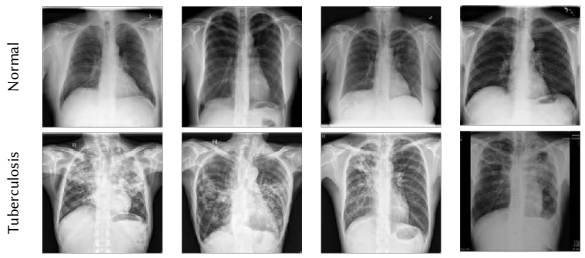


Fig. 2. Sample test images representing the normal and TB class X-ray.

B. Deep-Learning Procedure

Implementation of pre-trained/customary DLP is widely adopted to examine the clinical images recorded with a range of modalities and the DLP developed for a particular modality will work well on other modalities when it is trained with the new database of interest.

Developing and implementing a customary DLP is computationally complex, and it needs a considerable number of images to test and validate the performance. Hence, pre-trained DLPs are widely adopted to screen the medical images recorded with varied modalities. The conventional pre-trained DLPs are trained with the ImageNet database and a transfer-learning methodology helps to implement these schemes towards the new medical dataset which is to be examined.

In the proposed research, the pre-trained schemes, such as AlexNet, VGG16, VGG19, ResNet18, ResNet50 and ResNet101 are considered to categorize the considered X-ray image dataset into Normal/TB class. Initially, the performance of these DLPs is tested using SoftMax classifiers and the attained outcome confirms that the ResNet18 helped to get a better accuracy compared to other methods.

Table II presents the information regarding the DLPs and their network parameters considered in this study. To implement a fair assessment, every network is assigned with similar parameters and the classification accuracy alone is adopted as the performance evaluation metric. In every image case, data augmentation is employed using;

TABLE II. The Initial Parameters for the Considered DLP

Parameter	AlexNet	VGG16	VGG19	ResNet18	ResNet50	ResNet101
Initial weights	Imagenet	Imagenet	Imagenet	Imagenet	Imagenet	Imagenet
Epochs	100	100	100	100	100	100
Optimizer	Adam	Adam	Adam	Adam	Adam	Adam
Pooling	Average	Average	Average	Average	Average	Average
Hidden-layer activation	Relu	Relu	Relu	Relu	Relu	Relu
Classifier activation	Sigmoid	Sigmoid	Sigmoid	Sigmoid	Sigmoid	Sigmoid
Training data	2250	2250	2250	2250	2250	2250
Validation data	250	250	250	250	250	250
Total features	1000	1000	1000	1000	1000	1000

horizontal flip, vertical flip, rotation=20°, zoom=0.2, width shift=0.2, height shift=0.2, and shear range=0.1.

These images help the DLP to understand the information in the test pictures appropriately.

The ResNet18 has a simple structure compared to its variants and in the proposed work, this scheme helped to get better classification accuracy and therefore this architecture was considered to detect TB using deep and integrated deep, LBP and DWT features.

C. Handcrafted Features

The performance of automatic medical image examination is monitored by the features extracted from the test images considered for the assessment.

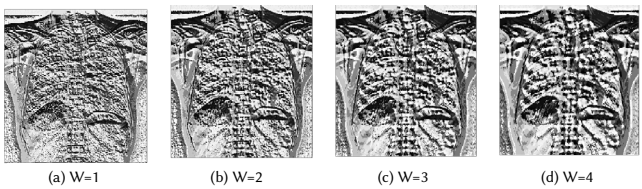
These features represent the technical information existing in the images and play a vital role in automatic segmentation and classification tasks in machine learning as well as DLP. The choice of a particular feature extraction procedure depends on; (i) Image modality, (ii) Methodology and (iii) Insight obtained from earlier literature. The previous works in the literature verify the use of LBP and DWT features to examine a class of images [40],[28],[29].

1. Weighted Local Binary Pattern

If LBP is an approved technique to mine the pixel level information from grayscale images and every LBP-treated image will help to get a one-dimensional feature vector of size 1x1x59. Recent procedures considered the LBP technique to mine the features from the normal/ disease class images and in this work, the LBP of varied weights (W= 1 to 4) discussed in Gudigar et al. [40] is considered.

Fig. 3 presents the weighted LBP enhanced TB class X-ray image and Fig 3(a) to (d) presents the images enhanced with various weights. Every X-ray picture is treated with this technique with chosen weights and the necessary features of size 1x1x236 are mined.

For the normal class X-ray images, approximately similar results are obtained. Other information related to the LBP features can be found in [16].



(a) W=1 (b) W=2 (c) W=3 (d) W=4

Fig. 3. Weighted LBP treated TB class X-ray pictures for W=1 to 4.

The LBP features of this study can be found below:

$$LBP_{W_1\,(1\times59)} = W_{1_{(1,1)}}, W_{1_{(1,2)}}, \ldots, W_{1_{(1,59)}} \quad (1)$$

$$LBP_{W_2\,(1\times59)} = W_{2_{(1,1)}}, W_{2_{(1,2)}}, \ldots, W_{2_{(1,59)}} \quad (2)$$

$$LBP_{W_3\,(1\times59)} = W_{3_{(1,1)}}, W_{3_{(1,2)}}, \ldots, W_{3_{(1,59)}} \quad (3)$$

$$LBP_{W_4\,(1\times59)} = W_{4_{(1,1)}}, W_{4_{(1,2)}}, \ldots, W_{4_{(1,59)}} \quad (4)$$

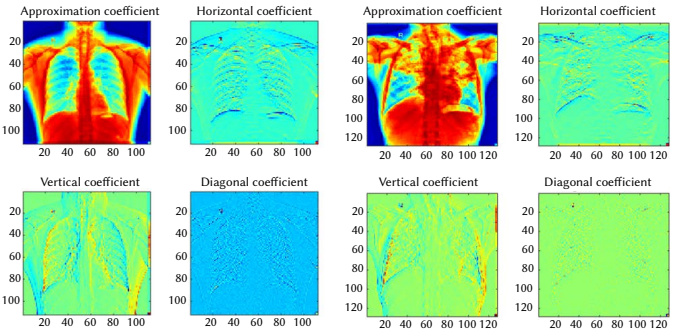
$$LBP_{Total\,(1\times236)} = LBP_{W_1\,(1\times59)} + LBP_{W_2\,(1\times59)} + LBP_{W_3\,(1\times59)} + LBP_{W_4\,(1\times59)} \quad (5)$$

Eqns. (1) to (4) present the features extracted for W=1 to 4 respectively and Eqn. (5) presents the arithmetic sum of total features extracted from all four images

2. Discrete Wavelet Transform

The concept of DWT-based feature extraction is one of the most widely adopted techniques and in this method, the chosen test image (RGB/grayscale) is divided into 4 sections, such as approximate- (LL), vertical- (LH), horizontal- (HL) and diagonal-coefficients (HH) and from every picture, necessary features are extracted as discussed in the work of Mirniaharikandehei et al. [41].

This work pointed out the essential 13 features to be extracted from the DWT treated images (contrast, correlation, energy, homogeneity, mean, standard-deviation, entropy, root-mean-square-level, variance, smoothness, kurtosis, skewness, and inverse-difference-moment) as discussed in the earlier research.



(a) Normal (b) TB

Fig. 4. Normal/TB class X-ray images enhanced with DWT approach.

Fig. 4 presents the DWT enhanced X-ray images and in this work, Jet colour map is considered to differentiate the normal and TB class pictures. Fig 4(a) presents the normal image and Fig 4(b) denotes the image with TB infection. The test image of size 224x224x3 pixels is divided into 4 equal sub-sections with a dimension of 112x112x3 pixels which gives a total feature vector of 1x1x52 (13 features x 4 sub-sections= 52 features).

The DWT features obtained from all 4 image sections are presented below;

$$DWT_{LL\,(1\times13)} = LL_{(1,1)}, LL_{(1,2)}, \ldots, LL_{(1,13)} \quad (6)$$

$$DWT_{LH\,(1\times13)} = LH_{(1,1)}, LH_{(1,2)}, \ldots, LH_{(1,13)} \quad (7)$$

$$DWT_{HL\,(1\times13)} = HL_{(1,1)}, HL_{(1,2)}, \ldots, HL_{(1,13)} \quad (8)$$

$$DWT_{HH\,(1\times13)} = HH_{(1,1)}, HH_{(1,2)}, \ldots, HH_{(1,13)} \quad (9)$$

$$DWT_{Total\,(1\times52)} = LL_{(1\times13)} + LH_{(1\times13)} + HL_{(1\times13)} + HH_{(1\times13)} \quad (10)$$

Eqns. (6) to (9) present the features extracted in LL, LH, HL and HH respectively, and Eqn. (10) presents the arithmetic sum of total features.

D. Firefly Algorithm-Based Feature Optimization

Feature reduction with a preferred statistical/heuristic scheme is commonly adopted in automatic data assessment to reduce the over-fitting problem and implementation of conventional statistical procedure (Student's t-test) is commonly adopted in the literature to reduce the features in medical data assessment [15]. Along with the statistical method, heuristic algorithm-supported feature selection is also widely considered by researchers to select the optimal feature set.

In this work, the Brownian-walk Firefly Algorithm (BFA) is employed to select optimal values of deep, LBP and DWT features and after the selection, serial integration of features is employed to form a new 1D feature vector of reduced size.

BFA is an improved form of FA, in which the search process is monitored by a slow and steady Brownian-walk operator [42]. When this scheme is implemented, the agents (artificial fireflies) are allowed to slowly explore the normal/TB class features and select the features whose Hamming-Distance (HD) is large.

Fig. 5 presents the feature selection procedure using the BFA and this technique helps to reduce the original features (F & F') of Normal/TB class image into a reduced set (S) using HD'_{max} (F & F') as the objective value. This procedure helped to reduce the deep, LBP and DWT features to an optimal value and the workflow of this technique is presented in Fig. 6 shows the implementation of the proposed scheme to optimize the Deep/LBP/DWT features.

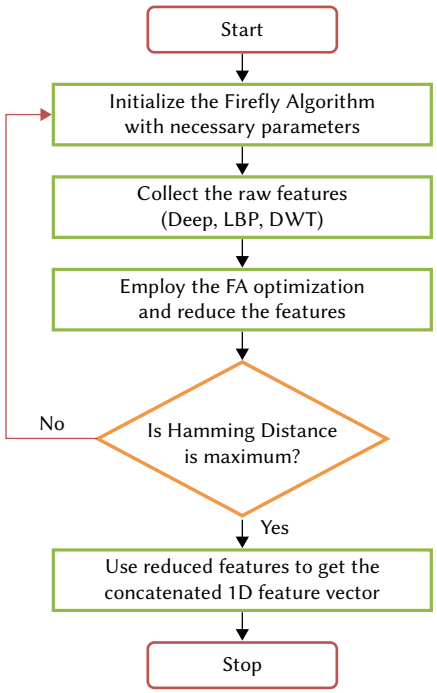


Fig. 5. Feature selection with BFA – Flowchart.

In this work, the initial parameters for the BFA are allocated as; number of agents=30, search operator=Brownian motion, monitoring parameter=maximization of HD and number of iterations ($Iter_{max}$) = 1000. The search will be stopped after finding the necessary feature vector from the chosen features and in this work, the optimized values of the features of each case are presented in Eqns. (11) to (13).

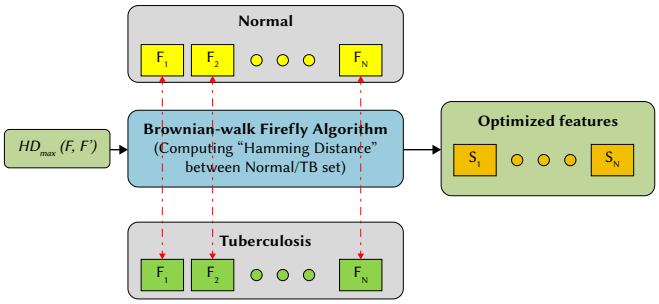


Fig. 6. Feature optimization with BFA - Flowchart.

$$s\,Net18_{Deep\,(1\times414)} = Deep_{(1,1)}, Deep_{(1,2)}, \ldots, Deep_{(1,414)} \quad (11)$$

$$LBP_{(1\times83)} = LBP_{(1,1)}, LBP_{(1,2)}, \ldots, LBPp_{(1,83)} \quad (12)$$

$$DWT_{HH\,(1\times13)} = HH_{(1,1)}, HH_{(1,2)}, \ldots, HH_{(1,13)} \quad (13)$$

$$Concatenated\ Features_{(1\times533)} = s\,Net18_{Deep(1\times414)} + LBP_{(1\times83)} + DWT_{(1\times36)} \quad (14)$$

Eqns. (11) to (13) present the BFA selected features of Deep, LBP and DWT and all these features are then serially combined to get a new feature vector as presented in Eqn. (14). This feature vector is then adopted to train and confirm the binary classifiers.

E. Performance Validation

The merit of the proposed TB detection framework is tested and authorized using the well-known binary classifiers, such as SoftMax, NB, LD, DT, RF, KNN ad SVM and the necessary information about these techniques can be obtained from [17]-[20]. During the performance evaluation, the necessary metrics, such as ACC, MIC, PRE, SEN, SPE, F1S and NPV are computed and based on these values; the merit of the proposed DDF is verified.

The mathematical expressions of these metrics are presented in Eqns. (15) to (21) [16], [4], [43]:

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (15)$$

$$MIC = 1 - ACC \quad (16)$$

$$PRE = \frac{TP}{TP+FP} \quad (17)$$

$$SEN = \frac{TP}{TP+FN} \quad (18)$$

$$SPE = \frac{TN}{TN+FP} \quad (19)$$

$$F1S = \frac{2TP}{2TP+FN+FP} \quad (20)$$

$$NPV = \frac{TN}{TN+FN} \quad (21)$$

where FP, FN, TP, and TN represents false-positive, false-negative, true-positive, and true-negative, respectively.

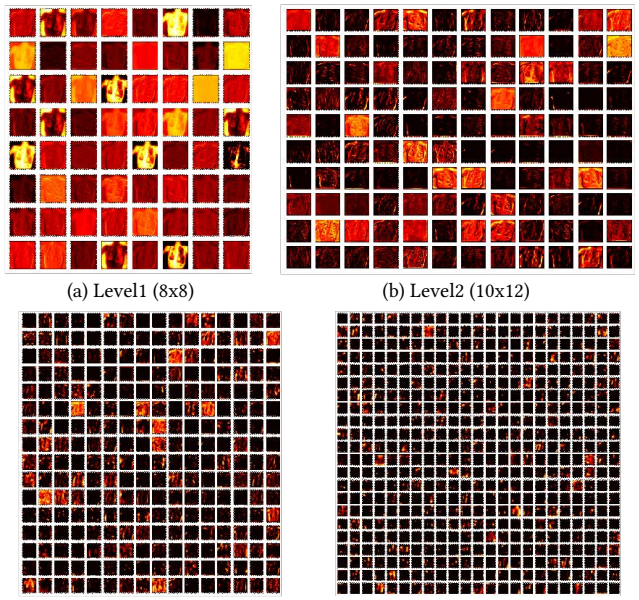
V. Experimental Results and Discussion

This part of the work reveals the investigational result achieved by the proposed DDF and this scheme is implemented using a workstation of; Intel i5 2.5 GHz CPU with 16GB RAM and 4GB VRAM and equipped with Python®.

In this work, 2500 images of Normal/TB class images are considered for the assessment, in which 90% of images (2250) are considered for the training and 10% images (250) are used for the validation process. During the deep-feature supported classification with a 5-fold cross-validation, SoftMax based binary-classification is implemented and the attained performance metrics are evaluated.

The experimental outcome confirms that ResNet18 combined with the SoftMax helps to achieve a classification accuracy of >95%, which is better compared to other DLP adopted in this study. Due to its performance, the ResNet18 scheme is then considered to construct the necessary DDF to examine the X-ray pictures.

ResNet18 consists of four separate levels as presented in Fig. 1 and each layer learns based on the information available in the images. The learned pixel information obtained from all these four levels is presented in Fig. 7, in which Fig. 7(a) to (d) represents the outcomes of levels 1 to 4, respectively. The final level result is then connected to the fully connected layer, which then presents a 1D learned (deep) features with size 1×1×1000.

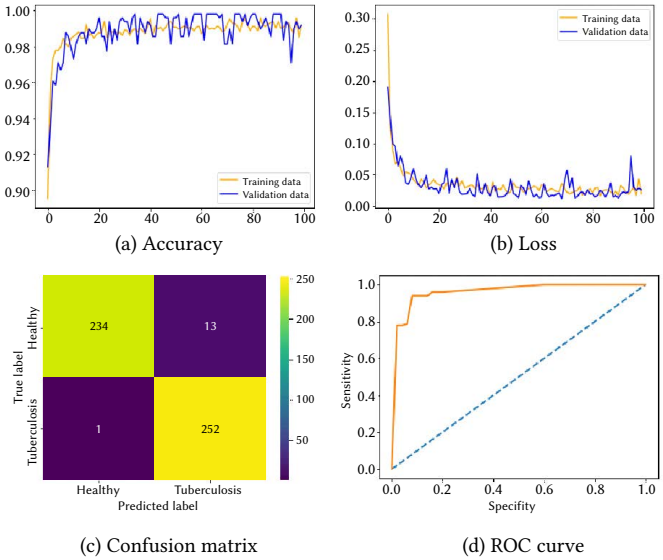


(a) Level1 (8x8) (b) Level2 (10x12)

(c) Level3 (16x16) (d) Level4 (22x23)

Fig. 7. Various convolutional-layer outcomes of ResNet18.

This feature is initially employed to train and authorize the SoftMax classifier existing in the DLP, and the attained results are presented in Fig. 8 and Table III. Fig. 8(a) and (b) present the accuracy and loss value achieved for the ResNet18 scheme and Fig. 7(c) and (d) depict the confusion matrix and the Receiver Operating Characteristic (ROC) curve (p-value >0.981) achieved with SoftMax.



(a) Accuracy (b) Loss

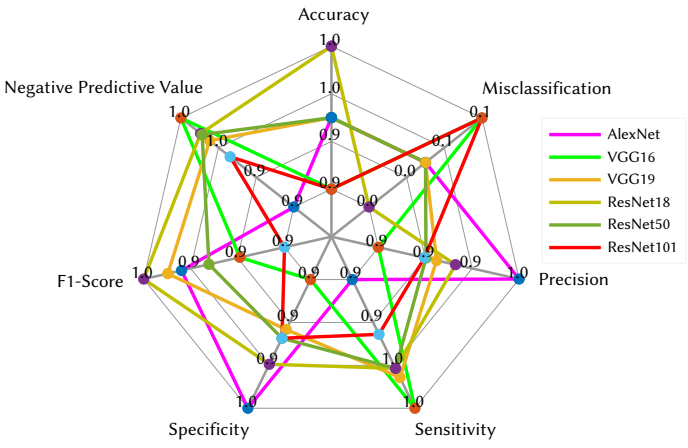
(c) Confusion matrix (d) ROC curve

Fig. 8. Classification results achieved with ResNet18 with SoftMax.

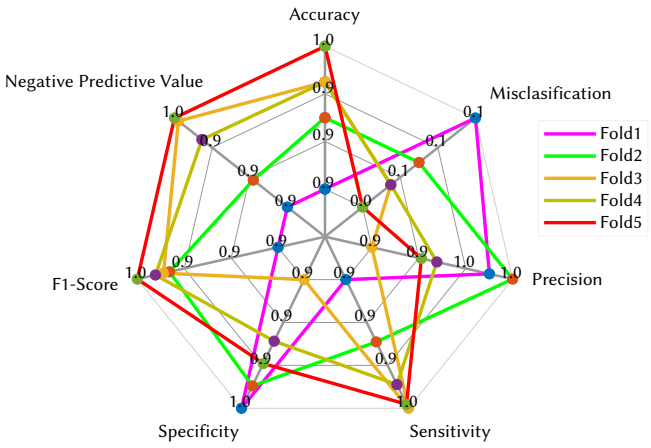
In this work, a 5-fold cross-validation is considered, and the best result is selected as presented in Table IV. In this table, Fold5 presents the best accuracy compared to other folds. Table III and IV values are graphically compared as presented in Fig. 9 (spider-plot). Fig 9(a) confirms the merit of ResNet18 and Fig. 9(b) verifies the merit of considered validation (Fold5).

TABLE III. Experimental Cross-Validation Outcome Obtained for Various DLP With SoftMax

Method	TP	FN	TN	FP	ACC	MIC	PRE	SEN	SPE	F1S`	NPV
AlexNet	237	14	238	11	0.9500	0.0500	0.9556	0.9442	0.9558	0.9499	0.9444
VGG16	238	10	236	16	0.9480	0.0520	0.9370	0.9597	0.9365	0.9482	0.9593
VGG19	239	11	236	14	0.9500	0.0500	0.9447	0.9560	0.9440	0.9503	0.9555
ResNet18	233	11	243	13	0.9520	0.0480	0.9472	0.9549	0.9492	0.9510	0.9567
ResNet50	233	11	242	14	0.9500	0.0500	0.9433	0.9549	0.9453	0.9491	0.9565
ResNet101	232	12	242	14	0.9480	0.0520	0.9431	0.9508	0.9453	0.9469	0.9528



(a) Spider-Plot for pre-trained DLP

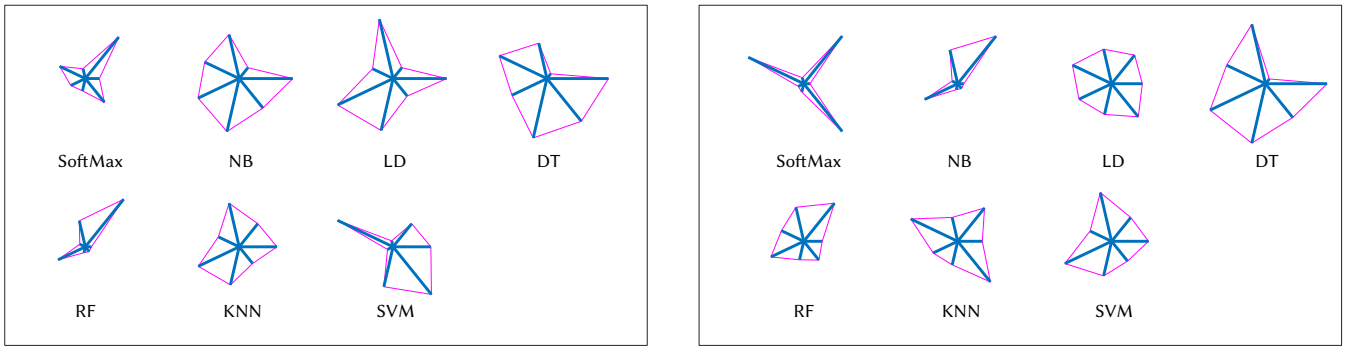


(b) Spider-Plot for 5-fold cross validation of ResNet18 with SoftMax

Fig. 9. Spider plot demonstration of various DLP outcomes with SoftMax.

TABLE V. 5-Fold Cross-Validation Result of ResNet18 With SoftMax

Features	Classifier	TP	FN	TN	FP	ACC	MIC	PRE	SEN	SPE	F1S`	NPV
Deep features	SoftMax	233	11	243	13	0.9520	0.0480	0.9472	0.9549	0.9492	0.9510	0.9567
	NB	240	10	241	9	0.9620	0.0380	0.9639	0.9600	0.9640	0.9619	0.9602
	LD	238	12	243	7	0.9620	0.0380	0.9714	0.9520	0.9720	0.9616	0.9529
	DT	238	8	244	10	0.9640	0.0360	0.9597	0.9675	0.9606	0.9636	0.9683
	RF	234	14	241	11	0.9500	0.0500	0.9551	0.9435	0.9563	0.9493	0.9451
	KNN	239	12	240	9	0.9580	0.0420	0.9637	0.9522	0.9639	0.9579	0.9524
	SVM	242	7	237	14	0.9580	0.0420	0.9453	0.9719	0.9442	0.9584	0.9713
Deep + LBP + DWT	SoftMax	252	1	234	13	0.9720	0.0280	0.9509	0.9960	0.9474	0.9730	0.9957
	NB	245	7	241	7	0.9720	0.0280	0.9722	0.9722	0.9718	0.9722	0.9718
	LD	251	3	239	7	0.9800	0.0200	0.9729	0.9882	0.9715	0.9805	0.9876
	DT	249	3	246	2	0.9900	0.0100	0.9920	0.9881	0.9919	0.9901	0.9880
	RF	248	5	240	7	0.9760	0.0240	0.9725	0.9802	0.9717	0.9764	0.9796
	KNN	246	2	243	9	0.9780	0.0220	0.9647	0.9919	0.9643	0.9781	0.9918
	SVM	243	5	248	4	0.9820	0.0180	0.9838	0.9798	0.9841	0.9818	0.9802



(a) Performance with deep-features

(b) Performance with FA optimized Deep+LBP+DWT

Fig. 10. Glyph-plot comparison of the performance of ResNet18 for various classifiers with deep and integrated features.

TABLE IV. 5-Fold Cross-Validation Result of ResNet18 With SoftMax

Method	TP	FN	TN	FP	ACC	MIC	PRE	SEN	SPE	F1S`	NPV
Fold1	237	14	238	11	0.9500	0.0500	0.9556	0.9442	0.9558	0.9499	0.9444
Fold2	238	10	236	16	0.9480	0.0520	0.9370	0.9597	0.9365	0.9482	0.9593
Fold3	239	11	236	14	0.9500	0.0500	0.9447	0.9560	0.9440	0.9503	0.9555
Fold4	233	11	243	13	0.9520	0.0480	0.9472	0.9549	0.9492	0.9510	0.9567
Fold5	233	11	242	14	0.9500	0.0500	0.9433	0.9549	0.9453	0.9491	0.9565
Fold1	232	12	242	14	0.9480	0.0520	0.9431	0.9508	0.9453	0.9469	0.9528

The achieved result of this experiment confirms that the deep feature-based classification helped to get an accuracy >on95% with SoftMax. To verify other classifiers' merit, this experiment is repeated using the deep features and the attained results are then compared and in Table V and Fig. 10. Fig 10(a) verifies that the binary classification with the DT helps to get better accuracy (>96%) on the considered X-ray picture dataset with deep feature.

In order to improve the accuracy further, the integrated deep, LBP and DWT features (Eqn. (14)) are considered, and the experiment is repeated with various classifiers using 5-fold cross-validation. The outcome of this experiment is also presented in Table V and Fig 10(b) verifies that the classification with DT helps to get a better result (99%) compared to other approaches.

The Glyph-plot of Fig 10(b) confirms that the texture formed by DT is big compared to the alternatives, which verifies the overall merit of the DT classifier. This experimental investigation confirms that the ResNet18 with DT classifier helps to get better TB detection using the X-ray images. Further, the proposed DDF is tested using deep-feature and integrated features and the outcome verifies that the integrated feature supported DDF works well compared to the alternatives.

This research work confirms that the proposed scheme helps to get a TB detection accuracy of 99%. To verify the merit of this scheme, the results of Rahman et al. [19] and Rajakumar et al. [20] are compared with the outcome of the proposed scheme along with other methods discussed in Table I.

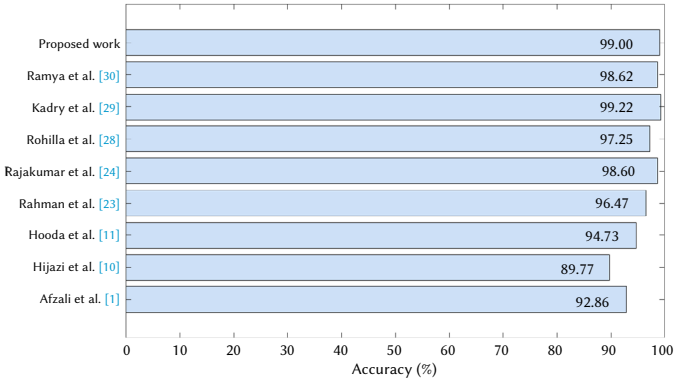


Fig. 11. Performance evaluation of the result of proposed work with recent existing works.

Fig. 11 presents the evaluation of the proposed scheme with the existing scheme and this comparison confirms that the classification accuracy of the proposed technique is better compared to the earlier works. Further, these results outperform the results of dual-deep features of Rajakumar et al. [20] and CNN segmentation combined classification of Rahman et al. [19] on the chosen image database. In the future, the performance of the REsNet18 supported DDF can be tested and validated on other existing TB datasets and clinically collected X-ray images.

VI. Conclusion

Tuberculosis (TB) causes severe lung problems and early screening, and treatment will reduce the impact. X-ray-supported TB detection is a common clinical scheme, and the recorded X-ray is then examined by radiologists and pulmonologists to detect the infection rate and plan the treatment.

Computerized X-ray detection is very common in hospitals and to support computerized screening of TB, this work proposed a DDF using the pre-trained DLP. This work initially executes a performance assessment of the DLPs such as AlexNet, VGG16, VGG19, ResNet18, ResNet50 and ResNet101 existing in the literature and finds that the ResNet18 along with SoftMax classifier helps to get a better classification result.

To improve the result of this scheme, the deep features are then optimized with FA and then combined with FA-optimized LBP and DWT features. This scheme helps to get a reduced 1D feature vector of dimension 1×1×544, which is then considered to train and validate the classifiers.

The DT classifier with a 5-fold cross-validation helped to get a classification accuracy of 99% with this feature vector and it is close to the earlier work implemented with CNN segmentation and classification. This outcome substantiates that the developed DDF helps to obtain an improved TB screening compared to the earlier methods. In the future, this scheme can be considered to evaluate the other X-ray image dataset in the literature and clinically collected X-ray images.

References

[1] S. Arunmozhi, , A. P. Kamath, V. Rajinikanth, . Detection of Tuberculosis in Chest X-Ray using Cancatinated Deep and Handcrafted Features. *In 2021 International Conference on System, Computation, Automation and Networking (ICSCAN) (pp. 1-4)*. IEEE, 2021.

[2] L. Ma, Y.Wang, L.Guo, Y.Zhang, P.Wang, X.Pei, et al.Developing and verifying automatic detection of active pulmonary tuberculosis from multi-slice spiral CT images based on deep learning. *Journal of X-Ray Science and Technology*, (Preprint), 1-13, 2020.

[3] M. Nijiati, Z. Zhang, A. Abulizi, H.Miao, A.Tuluhong, S.Quan, et al. Deep learning assistance for tuberculosis diagnosis with chest radiography in low-resource settings. *Journal of X-Ray Science and Technology*, (Preprint), 1-12, 2021.

[4] X. Wang, Z. Zhang, D. Chen, N. Peng, P. U. Thakker, M. Z. Schwartz, Y. Zhang, Challenges in the diagnosis of testicular infarction in the presence of prolonged epididymitis: Three cases report and literature review. *Journal of X-Ray Science and Technology*, 28(4), 809-819, 2020.

[5] K. Zaman, Tuberculosis: a global health problem. *Journal of health, population, and nutrition*, 28(2), 111, 2010.

[6] WHO (https://www.who.int/news-room/fact-sheets/detail/tuberculosis)

[7] E. Priya, Optimization-Based Tuberculosis Image Segmentation by Ant Colony Heuristic Method. *International Journal of Swarm Intelligence Research (IJSIR)*, 13(1), 1-24, 2022.

[8] E. Priya, S. Srinivasan, Automated decision support system for tuberculosis digital images using evolutionary learning machines. *European Journal of Biomedical Informatics*, 9(2), 2013.

[9] S. Arunmozhi, V. Rajinikanth, M. P. Rajakumar, . Deep-Learning based Automated Detection of Pneumonia in Chest Radiographs. *In 2021 International Conference on System, Computation, Automation and Networking (ICSCAN) (pp. 1-4). IEEE*, 2021.

[10] A. Aziz, M. Attique, U.Tariq, U., Y. Nam, M. Nazir, C. W. Jeong et al., . An Ensemble of Optimal Deep Learning Features for brain tumor classification. doi:10.32604/cmc.2021.018606, 2021

[11] A. Bhandary, G. A. Prabhu, V. Rajinikanth, VK. P. Thanaraj, , S. C. Satapathy, ., D. E. Robbins, Deep-learning framework to detect lung abnormality–A study with chest X-Ray and lung CT scan images. Pattern Recognition Letters, 129, 271-278, 2020.

[12] X. Chen, X. Wang, K. Zhang, K., R. Zhang, K. M. Fung, T. C. Thai, et al.. Recent advances and clinical applications of deep learning in medical image analysis. arXiv preprint arXiv:2105.13381, 2021.

[13] V. Chouhan, S. K. Singh, A. Khamparia, D. Gupta, P. Tiwari, C. Moreira, et al.A noveltransfer learning based approach for pneumonia detection in chest X-ray images. Applied Sciences, 10(2), 559, 2020.

[14] N. Dey, Y. D. Zhang, Y. D., V. Rajinikanth, R. Pugalenthi, N. S. M. Raja. Customized VGG19 architecture for pneumonia detection in chest X-rays. *Pattern Recognition Letters*, 143, 67-74, 2021.

[15] U. Raghavendra, A., Gudigar, T. N. Rao,V. Rajinikanth, E. J. Ciaccio, , C. H. Yeong, et al. Feature-versus deep learning-based approaches for the automated detection of brain tumor with magnetic resonance images: A comparative study. *International Journal of Imaging Systems and Technology*, 2021.

[16] V. Rajinikanth, S. M. Aslam, S. Kadry, Deep Learning Framework to Detect Ischemic Stroke Lesion in Brain MRI Slices of Flair/DW/T1 Modalities. *Symmetry*, 13(11), 2080, 2021.

[17] V. Rajinikanth, S. Kadry, Y. Nam, Convolutional-Neural-Network Assisted Segmentation and SVM Classification of Brain Tumor in Clinical MRI Slices. *Information Technology and Control*, 50(2), 342-356, 2021.

[18] M. Ramzan, M. Raza, M. Sharif, M. A. Khan, Y. Nam, Y. . Gastrointestinal Tract Infections Classification Using Deep Learning. *Cmc-Computers Materials & Continua* DOI:10.32604/cmc.2021.015920, 2021.

[19] T. Rahman, A. Khandakar, M. A. Kadir, K. R. Islam, K. F. Islam, R. Mazhar, et al. Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization. *IEEE Access*, 8(, 191586-191601, 2020.

[20] M. P. Rajakumar, R. Sonia, B. U. Maheswari, S. P. Karuppiah, Tuberculosis detection in chest X-ray using Mayfly-algorithm optimized dual-deep-learning features. *Journal of X-Ray Science and Technology*, 29, 961-974, 2021.

[21] M. Odusami, R. Maskeliunas, R. Damaševičius, S. Misra, Comparable Study of Pre-trained Model on Alzheimer Disease Classification. *In International Conference on Computational Science and Its Applications (pp. 63-74)*. Springer, Cham, 2021.

[22] A. Afzali, F. B. Mofrad, ., M. Pouladian, . Contour-based lung shape analysis in order to tuberculosis detection: modeling and feature description. *Medical & biological engineering & computing*, 58, 1965-1986. 2020.

[23] S. Jaeger, S. Candemir, S. Antani, Y.X.J. Wáng, P.X. Lu, P., & Thoma, Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6), 475, 2014.

[24] M. H. A. Hijazi, S. K. T. Hwa, A. Bade, R. Yaakob, M. S. Jeffree, Ensemble deep learning for tuberculosis detection using chest X-Ray and canny edge detected images, *IAES International Journal of Artificial Intelligence*, 8(2019), 429, 2019.

[25] R. Hooda, S.Sofat, S.Kaur, A.Mittal, F. Meriaudeau, F, Deep-learning: A potential method for tuberculosis detection using chest radiography, *In 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)* 497-502. IEEE, 2017.

[26] TB Data [https://ieee-dataport.org/documents/tuberculosis-tb-chest-x-ray-database]. DOI:10.21227/mps8-kb56.

[27] A. Rohilla, R. Hooda, A. Mittal,TB detection in chest radiograph using deep learning architecture, *ICETETSM-17, 136-147*, 2017.

[28] S. Kadry, G. Srivastava, V. Rajinikanth, S. Rho, Y. Kim, Tuberculosis detection in chest radiographs using spotted hyena algorithm optimized deep and handcrafted features. *Computational Intelligence and Neuroscience*, 2022. https://doi.org/10.1155/2022/9263379

[29] R. Mohan, S., Kadry, V. Rajinikanth, A. Majumdar, O. Thinnukool, Automatic Detection of Tuberculosis Using VGG19 with Seagull-

Algorithm. *Life*, 12(11), 1848, 2022.

[30] S. Kadry, V.Rajinikanth, R.González Crespo, E. Verdú, Automated detection of age-related macular degeneration using a pre-trained deep-learning scheme. *The Journal of Supercomputing*, 1-20, 2021.

[31] M. A. Khan, A.Majid, N.Hussain, M. Alhaisoni, Y. D.Zhang, S.Kadry, Y.Nam,Multiclass Stomach Diseases Classification Using Deep Learning Features Optimization., 2021

[32] M. A. Khan, V.Rajinikanth, S. C.Satapathy, D.Taniar, J. R. Mohanty, U.Tariq, R.Damaševičius, VGG19 Network Assisted Joint Segmentation and Classification of Lung Nodules in CT Images. *Diagnostics*, 11(12), 2208, 2021.

[33] M. A. Khan, M. Sharif, T.Akram, R.Damaševičius, R.Maskeliūnas, Skin lesion segmentation and multiclass classification using deep learning features and improved moth flame optimization. *Diagnostics*, 11(5), 811, 2021.

[34] S. Kaliyugarasan, A. Lundervold, A. S. Lundervold, Pulmonary nodule classification in lung cancer from 3D thoracic CT scans using fastai and MONAI. *International Journal of Interactive Multimedia and Artificial Intelligence*, 6(7), 83-89. DOI: 10.9781/ijimai.2021.05.002, 2021.

[35] V. Srivastava, S. Gupta, G.Chaudhary, G., A. Balodi, M. Khari, V. García-Díaz, An enhanced texture-based feature extraction approach for classification of biomedical images of CT-scan of lungs. *International Journal of Interactive Multimedia and Artificial Intelligence*, 6(7), 18-25. DOI: 10.9781/ijimai.2020.11.003, 2021.

[36] A. A. Rezaie, A. Habiboghli, Detection of lung nodules on medical images by the use of fractal segmentation. *International Journal of Interactive Multimedia and Artificial Intelligence*, 4(5), 15-19. DOI: 10.9781/ijimai.2017.452, 2017.

[37] V. C. Osamor, A. A. Azeta, O. O. Ajulo, Tuberculosis–Diagnostic Expert System: An architecture for translating patients information from the web for use in tuberculosis diagnosis. *Health informatics journal, 20(4), 275-287*, 2014.

[38] M. A. Khemchandani, S. M. Jadhav, B. R. Iyer, Brain tumor segmentation and identification using particle imperialist deep convolutional neural network in MRI images. *International Journal of Interactive Multimedia and Artificial Intelligence*, 7(7), 38-47. DOI: 10.9781/ijimai.2022.10.006, 2022.

[39] G. R. Vásquez-Morales, S. M. Martinez-Monterrubio, P. Moreno-Ger, J. A. Recio-Garcia, Explainable prediction of chronic renal disease in the Colombian population using neural networks and case-based reasoning. *IEEE Access*, 7, 152900-152910, 2019.

[40] A. Gudigar, U. Raghavendra, T. Devasia, K. Nayak, S. M. Danish, G. Kamath, et al. Global weighted LBP based entropy features for the assessment of pulmonary hypertension. *Pattern Recognition Letters*, 125, 35-41, 2019.

[41] S. Mirniaharikandehei, M. Heidari, G.Danala, S. Lakshmivarahan, B.A novel feature reduction method to improve performance of machine learning model. In Medical Imaging 2021: Computer-Aided Diagnosis (Vol. 11597, p. 1159726). *International Society for Optics and Photonics*, 2021.

[42] N. Sri Madhava Raja, V. Rajinikanth, K. Latha, K. (2014). Otsu based optimal multilevel image thresholding using firefly algorithm. *Modelling and Simulation in Engineering*, 2014.

[43] M. I. Waly, M. Y. Sikkandar, M. A. Aboamer, S. Kadry, O. Thinnukool, Optimal Deep Convolution Neural Network for Cervical Cancer Diagnosis Model. *Cmc-Computers Materials & Continua*, 70(2), 3295-3309, 2022.

V. Rajinikanth

He is a Professor in Department of Computer Science, Division of Research and Innovation, Saveetha School of Engineering, SIMATS, Chennai 602105, Tamilnadu, India. He has published more than 150 papers and authored/edited 8 books in the field of medical data assessment. His main research interests includes; Heuristic algorithm based optimization, Image thresholding, Machine learning and Deep learning.

Seifedine Kadry

Professor Seifedine Kadry has a Bachelor degree in 1999 from Lebanese University, MS degree in 2002 from Reims University (France) and EPFL (Lausanne), PhD in 2007 from Blaise Pascal University (France), HDR degree in 2017 from Rouen University. At present his research focuses on Data Science, education using technology, system prognostics, stochastic systems, and applied mathematics. He is an ABET program evaluator for computing, and ABET program evaluator for Engineering Tech. He is a Fellow of IET, Fellow of IETE, and Fellow of IACSIT. He is a distinguished speaker of IEEE Computer Society.

Pablo Moreno Ger

Dr. Moreno-Ger was born in Madrid in 1981. He finished his doctorate in Computer Engineering from Universidad Complutense de Madrid (UCM) in 2007 and was an Associate Professor in the Department of Software Engineering and Artificial Intelligence at UCM. Now he is with Universidad Internacional de La Rioja (UNIR), where he is currently the Vice-Rector for Research. Formerly, he was the Director of the School of Engineering and Technology at UNIR, as well as Vice-Dean for Innovation at the School of Computer Engineering at UCM. His main research interests are in technology-assisted teaching, artificial intelligence, learning analytics and serious games. He has published more than 150 academic works in these fields.

RGBeat: A Recoloring Algorithm for Deutan and Protan Dichromats

M. Madalena G. Ribeiro[1], Abel J. P. Gomes[2,3,4] *

[1] Polytechnic Institute of Castelo Branco, CIPEC, Castelo Branco (Portugal)
[2] Universidade da Beira Interior, Covilhã (Portugal)
[3] INESC-ID, Lisboa (Portugal)
[4] Instituto de Telecomunicações, Lisboa (Portugal)

Received 8 January 2021 | Accepted 23 November 2021 | Published 20 January 2022

unir
LA UNIVERSIDAD EN INTERNET

Abstract

Deutan and protan dichromats only see exactly two hues in the HSV color space, 240-blue (240°) and 60-yellow (60°). Consequently, they see both reds and greens as yellows; therefore, they cannot distinguish reds from greens very well. Thus, their color space is 2D and results from the intersection between the HSV color cone and the 60°-240° plane. The RGBeat recoloring algorithm's main contribution here is that it is the first recoloring algorithm that enhances the color perception of deutan and protan dichromats but without compromising the lifelong color perceptual learning. Also, as far as we know, this is the first HTML5-compliant web recoloring approach for dichromat people that considers both text and image recoloring in an integrated manner.

Keywords

Color Blindness, Contents Adaptation Dichromacy, Recoloring, Visual Accessibility.

DOI: 10.9781/ijimai.2022.01.003

I. Introduction

ABOUT 5% of the world population is affected by color vision deficiency (CVD), also called color blindness. This visual impairment hampers the color perception, ending up by limiting the overall perception of CVD people about the surrounding environment. A CVD individual may not distinguish between two different colors, which often originates confusion or a limited understanding of the reality, including web environments, whose web pages are plenty of media elements like text, still images, video, and sprites.

A. Color Vision

In the human eye, there are two types of cells in the retina: rods and cones. Rod cells only function in scotopic (dark) conditions so that they add up nothing to our perception of lightness and darkness in photopic (bright) conditions. That is, rods are responsible for our vision in light-absent environments (i.e., night vision), while cones are responsible for our perception of color.

There are three types of cone cells: L-cones (also known as red cones), M-cones (or green cones), and the S-cones (or blue cones), depending on their sensitivity to the type of wavelengths of light: long (L) wavelengths, medium (M) wavelengths, and short (S) wavelengths, respectively. The tristimulus theory tells us that the perception of color results from the combination of the light stimulation of those three types of cones [1].

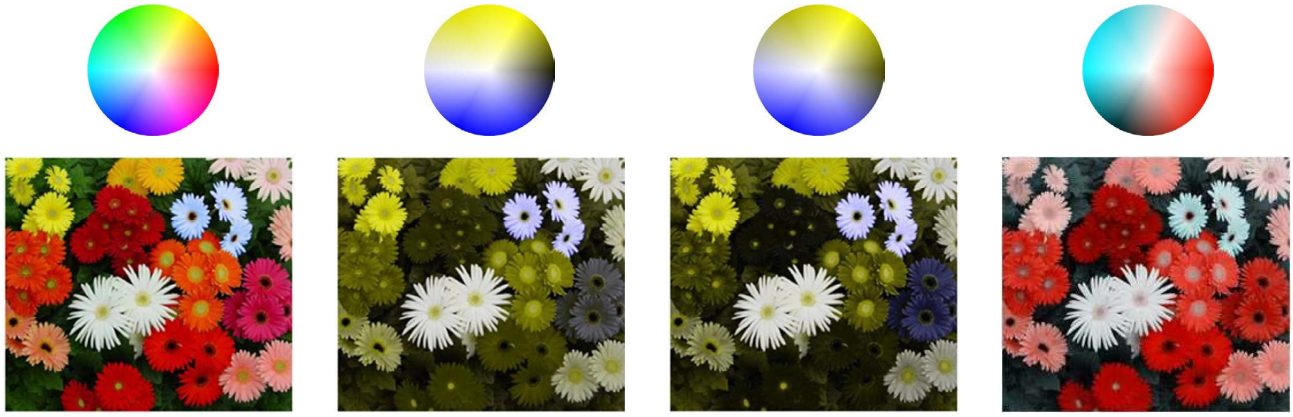
* Corresponding author.
E-mail addresses: mribeiro@ipcb.pt (M. M. G. Ribeiro), agomes@di.ubi.pt (A. J. P. Gomes).

B. Color Vision Anomalies

Typically, human beings are trichromats, i.e., they see the entire visible light spectrum because their three types of cones are working correctly, as illustrated in Fig. 1(a). CVD, sometimes called color blindness, is the result of the misfunctioning of some cones. Accordingly, CVD fits in one of the following categories: anomalous trichromacy, dichromacy, and monochromacy [2].

Anomalous trichromacy is the less severe CVD type and occurs when (at least) a kind of cone cell does not work correctly, either because they are not distributed regularly on the retina or because their sensitivity is weak [3] and [4]. Consequently, there exists a displacement of the sensitivity curve of the corresponding color channel, changing the way one perceives the color as a whole, i.e., in a distorted fashion. Vision anomalies depend on the type of affected cone cells [5]. Protanomaly, also known as red-weak vision, denotes the existence of anomalous L-cones. Deuteranomaly, also known as green-weak vision, indicates the existence of anomalous M-cones. Tritanomaly, also called blue-weak vision, shows the existence of anomalous S-cones. In conformity with the MPEG-21 standard [6], the degree of CVD severity for anomalous trichromats continuously varies in the interval [0.1, 0.9]. For example, those with a severity degree of 0.1 have a minor color vision distortion, while those with a severity degree of 0.9 almost see colors as dichromat people do because dichromacy corresponds to a severity degree of 1.0.

Dichromacy occurs when cones of a given sort do not work, mainly because those cones do not exist on the retina. So, the color perception is found on the other two channels, significantly reducing the color spectrum perceived by dichromat individuals. Depending on the type of inexistent cone cells, the anomaly calls protanopy, deuteranopy, and tritanopy, which accounts for the absence of the L-, M- and S-cone cells, respectively. Thus, all colors visible for trichromat people appear



(a) (b) (c) (d)

Fig. 1. Degrees of severity of color vision deficiency and its dichromat subtypes: (a) regular trichromat people; (b) deuteranope (or deutan) people; (c) protanope (or protan) people, and (d) tritanope (or tritan) people. Simulation of (b)-(c) using Brettel et al.'s algorithm [7] and (d) using Petrich's algorithm [8].

as two monochromatic hues for people with dichromacy: blues and yellows, both for people with deuteranopy and protanopy, and reds and bluish cyans for tritanope people [7], as illustrated in Fig. 1.

Monochromacy is the most severe CVD type due to the absence or non-functioning of two or three types of cone cells. This fact results in a rather severe reduction of the chromatic domain perceived by the individuals. There are two types of monochromacy: blue cone monochromacy, when only the S-cone cells are working [9], and rod monochromacy, when all types of cones are either missing or non-functioning for some reason. The blue-cone monochromacy leads to a gray-scale vision, yet some shades of blue may be noticeable [10]. Rod monochromacy, also known as achromatopsia, leads to a total lack of color experience and a low visual acuity, which is related to poor vision and high sensitivity to light [11] and [12].

Finally, let us mention that CVD has a prevalence of 5% in the Caucasian population on average, though its incidence is about 8% on men and 0.5% on women. Also, the prevalence of each CVD type decreases as the severity increases. Indeed, it is about 75% for anomalous trichromacy, 25% for dichromacy, and 0.00001% for monochromacy [13].

C. Contributions

RGBeat recoloring algorithm aims to help deutan and protan dichromat people. Like other recoloring algorithms, RGBeat aims at minimizing or even eliminating the likely confusion between reds and greens. Recall that deutan and protan dichromat people can distinguish some reds from greens, i.e., they can identify some reds because they have learned that such colors are reds. Nevertheless, dichromat people see reds as greenish colors. Indeed, lifelong color perceptual learning of each dichromat individual plays an important role to overcome part of ambiguity between reds and greens.

However, unlike other recoloring algorithms, RGBeat eliminates the color ambiguity as much as possible without compromising the lifelong color perceptual learning experienced by each dichromat individual. For this purpose, the following properties (or requirements) must be satisfied: color consistency, color naturalness, and color contrast. The challenge is how to increase the contrast between (confusing) colors and, at the same time, to maintain the consistency and naturalness of color.

The key contributions of the RGBeat recoloring algorithm are the following:

- It enhances the color perception of deutan and protan dichromats without undermining their lifelong color perceptual learning.
- It increases the contrast between confusing colors, though maintaining the color consistency and naturalness.
- It applies to HTML5-compliant web environments, including images, video, and text.
- It performs very fast because it only operates on the range of reds, making it feasible to recolor video in real-time.

The first contribution above concerns the research gap we have identified in the literature. Indeed, keeping the color perceptual learning of each dichromat person must be a priority for any recoloring algorithm.

D. Article Organization

The remainder of this article organizes itself as follows. Section II reviews prior recoloring algorithms against relevant requirements: color consistency, color naturalness, and color contrast. Section III details our recoloring algorithm, called RGBeat. Section IV approaches the text and background color adaptation, while Section V approaches color adaption for still images. Section VI presents the qualitative, quantitative, and performance results of RGBeat and competitor methods. Section VII describes usability testing and assessment of RGBeat and its competitor methods. Section VIII discusses the research work behind RGBeat. Section IX concludes the paper, pointing out some hints to future work in browser recoloring and adaptation, as needed for color-blind people.

II. Related Work

Dichromat people only see two distinct hues, although with different values of saturation and brightness. More specifically, deutan and protan dichromats see blues and yellows, respectively; in turn, tritan dichromats see reds and greenish blues. For example, as shown in Fig. 1, a deutan dichromat sees a weakly saturated yellow as a moss green. Besides, reducing the chromatic range to two hues may confuse what is seen in a given image (see Fig. 1).

As said above, the colorblind can distinguish some reds from greens because they have learned that some of such colors are reds. Indeed, the color perceptual learning of each dichromat individual is an essential tool to resolve the red-green ambiguity, even if it is partially. Consequently, in designing a new recoloring algorithm, it is of paramount importance to preserve the color perceptual learning of each dichromat individual. Thus, any contrast-based algorithm must apply minor contrast differences; otherwise, we end up undermining the color perceptual learning of each individual. Therefore, the challenge here lies in increasing the contrast between (confusing) colors and preserving the color consistency and naturalness. For more details, the reader is referred to Ribeiro and Gomes [14], a survey on recoloring methods based on these three concepts: color consistency, color naturalness, and color contrast.

Color consistency guarantees that if two colors are identical, the corresponding mapped colors will remain identical, independently of the set of colors subject to remapping. Note that the lack of color consistency creates bewilderment in the perception of the colorblind. Color consistency is an essential requirement that is gaining significance in recent recoloring methods [15]-[19], though others avoid approaching it [20] and [21]. Its importance stems from the fact that it prevents us from adding more color ambiguity to the typical ambiguity of reds and greens inherent to the colorblind. For example, let us consider that a pink hue maps to a magenta hue; one says that such color mapping is consistent when such pink always maps to the same magenta. Color consistency is vital for recoloring video; otherwise, it would be tough to maintain the temporal color consistency.

Color naturalness has to do with minimizing the perceptual difference between a color and its corresponding remapped color. This difference should be as low as possible so as not to break up the perceptual learning of the colorblind. Some works address this property [21]-[26]. However, others [27] and [28] do not show any concern about color naturalness.

Color contrast stems from increasing the perceptual difference of neighboring elements in an image to enhance the perception of the colorblind. In a way, color naturalness and contrast are contradictory requirements that make recoloring algorithms challenging to tune-up. For further details about color contrast in recoloring algorithms, the reader is referred to [23], [24] and [29]. However, this requirement has not been considered by other algorithms [21] and [22]. Interestingly, some algorithms create higher color contrast, although not referring to this goal explicitly [30].

Considering the three properties of color above, we developed a recoloring algorithm for dichromat people, RGBeat, which reduces the space for color ambiguity at the cost of a few colors that only slightly conflict with their perceptual learning. The reduction of color ambiguity between reds and greens occurs by increasing the color contrast between confusing colors. As will be seen ahead, we compare RGBeat to other two methods, which are due to Iaccarino et al. [24] and the Ching-Sabudin [30]. We selected these two methods among all those we may find in the literature because they satisfy the following requirements:

- They are color-consistent, a condition to prevent adding more confusion in the color perception.
- They preserve either color naturalness or increase color contrast.
- They apply to true-color images.

III. Recoloring Algorithm

RGBeat recoloring algorithm was designed for deuteranopy and protanopy simply because they are the most common types of dichromacy, and their color perception is quite similar.

A. Leading Idea

As shown in Fig. 2, deutan or protan dichromats only perceive two hues, yellow (60°) and blue (240°), yet with more or less luminance and saturation. We know that those people see greens as yellows, but they perceive them as unsaturated greens, i.e., they faintly perceive greens. Also, deutan or protan dichromats see reds as dark unsaturated yellows. Consequently, deutan and protan dichromats confuse reddish and greenish colors, though they can distinguish some reds from greens, which stems from their lifelong color perceptual learning.

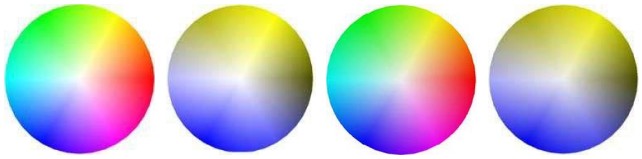


Fig. 2. HSV cone sections: (a) a section with value = 100%, as seen by trichromat people; (b) a section with value = 100%, as seen by deuteranope people; (c) a section with value = 75%, as seen by trichromat people; and (d) a section with value = 75%, as seen by deuteranope people. Deuteranope simulation based on Vienot et al.'s algorithm [31].

Thus, the leading idea of our algorithm is to reduce the confusion between reds and greens further so that only reds will be subject to a recoloring procedure. For that purpose, as mentioned in the previous section, it is crucial to preserve the color perceptual consistency and slightly increase the color contrast for reds to preserve the color naturalness as much as possible.

B. HSV Color Domain of Deuteranopy and Protanopy

In comparison to the color range seen by trichromat people, the color range seen by deutan or protan dichromats is quite limited. This turns up more evident when we consider HSV colors, as illustrated in Fig. 3. The HSV color space is a cone such that $H \in [0°,360°[$ stands for the hues (or color wheel), $S \in [0, 100]$ denotes the saturation range, which increases from the cone apex to base and perpendicularly to cone axis, and $V \in [0, 255]$ the brightness value, which increases from the apex to base of the cone [32]. In Fig. 3(a), we see the entire cone of colors seen by trichromat people, while the colors seen by deutan and protan dichromats map onto colors in the 60°/240° plane, which cuts the cone into two parts (cf. Fig. 3(b)).

As noted above, the deutan or protan dichromats only see yellows and blues. However, contrary to the general idea that such blues and yellows include multiple hues, protan and deutan people only see the yellow hue of 60° and the blue hue of 240°. In truth, they only see these two hues with more or less saturation and brightness. Our experiments confirmed this by using the simulation algorithm due to Brettel et al. [7], converting then the color to the HVS color space

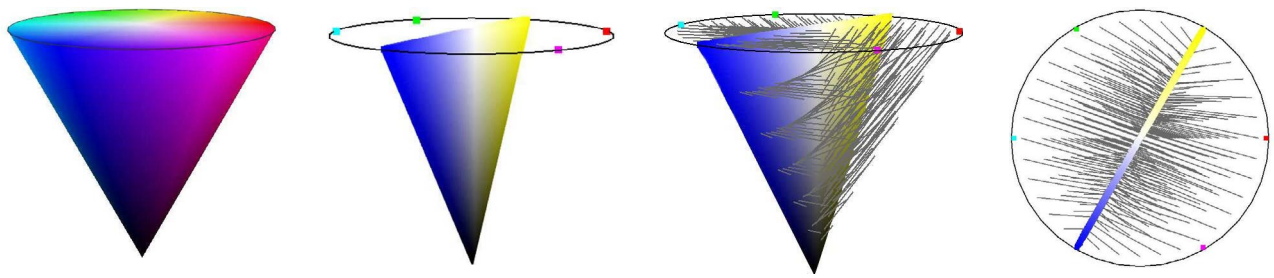


Fig. 3. The HSV color space when seen by: (a) trichromat people; (b) deuteranope people (deuteranope half planes); (c) and (d) deuteranope people, with the color projection lines, joining colors seen by trichromat and deuteranope people, being (c) a lateral view and (d) the top view. Deuteranope simulation based on Vienot et al.'s algorithm [12].

[33]. We constructed the deuteranope color gamut in the HSV color space by varying R, G, and B in the range [0, 255], converting RGB to RGB_{deutan} (i.e., using the method proposed by Vienot et al. [31]), and RGB_{deutan} to HSV afterward. This deuteranope color gamut reduces to two coplanar triangles, resulting from the intersection between a plane and the HSV cone (see Fig. 3). We can obtain a similar planar color gamut in the LMS color space [7] and CIE XYZ color space [34] for deuteranope people.

As shown in Fig. 3(c), the visual system of a deutan dichromat transforms the conical HVS color space (Fig. 3(a)) into a triangular color space (Fig. 3(b)). In other words, each hue of the HSV cone projects onto a triangle belonging to the plane defined by the apex, 60°-hue point and 240°-hue point. As illustrated in Fig. 3, such color projection entails a change of saturation and brightness (usually a loss). It is clear that colors projected onto the yellow part of the triangle]150°, 330°[noticeably lose value (brightness), while colors projected onto the blue part tend to keep value (brightness) unchanged. Regarding saturation, no significant changes occur (see Fig. 3(d)). When there is a change, the saturation may increase or decrease (although slightly). The behavior of the protan dichromat's projection is similar, though shades are slightly different.

C. RGBeat Recoloring Procedure

A glance at Fig. 2 shows us that deutan or protan dichromats tend to see reds as dark unsaturated greens, so that a pure red is seen as dark green, not to say black. As shown in Fig. 3, greens and reds are seen similarly by deutan or protan dichromats; hence, their well-known red-green confusion, also known as colorblindness. With this problem is mind, we decided not to change the saturation $S \in [0, 100]$ neither the brightness (value) $V \in [0, 255]$ in the HSV model; we only changed hues as follows:

- Hues in the range]0°, 60°[(range of reds and oranges) squeezed into the range]30°, 60°[.
- Hues in the range]300°, 360°] (range of reds, pinks, and magentas) squeezed into the range]300°, 330°].

The remaining hues in the range [60°, 300°] remain unchanged. This way, those deutan or protan dichromats end up having fewer dark hues (yellows and blues) in the range of reds (for trichromats). So, we ensure the fulfillment of the requirements of color consistency and naturalness. However, despite the soundness of this recoloring approach, we can ask ourselves about its actual help for colorblind. Indeed, it is convenient here to recall that:

- Reducing the hue range seems to us a good decision because we are just eliminating the subrange of confusing hues, i.e., the reds in this case. For example, in Figs. 4 and 5, we see that a deutan cannot distinguish a green flower from a red flower, but a deutan can easily distinguish them with our recoloring technique. In short, the question is not reducing the hue range but cutting off the subrange of reds after mapping reds to close hues as much as possible. The purpose is thus to be able to discriminate confusing hues.
- We must carry out the mapping of hues reducing at the same time the impact on the perceptual learning of the deutans as much as possible, i.e., without affecting the color naturalness that much. So, it makes sense to remap hues according to our technique, which keeps the color consistency while increases the contrast.

Before proceeding any further, let us recall that the hue compression of]0°, 60°[corresponds to the following non-linear interpolation formula:

$$H' = H + H.\left(\frac{60 - H}{60}\right) \tag{1}$$

while the hue compression of]300°, 360°[is given by

$$H' = H + (300 - H).\left(\frac{360 - H}{60}\right) \tag{2}$$

1. Mapping hues from]0°, 60°[: Let us consider the color (R, G, B), with $R, G, B \in [0, 255]$, being the corresponding normalized color (r, g, b) given by

$$r = R/255 \qquad g = G/255 \qquad b = B/255 \tag{3}$$

Let us also assume that the values of S and V remain unchanged. We know that by only changing the value of H in the range]0°, 60°[, we only change the value of G in the RGB color space. Indeed, considering $H \in]0°, 60°[$, the HSV-RGB conversion formula due to Smith [33] sets that

$$R = V \qquad G = V(1 - S(1 - 6H)) \qquad B = V(1 - S) \tag{4}$$

Since we stated that S and V remain unchanged in the HSV model, we conclude that G is the only RGB parameter that changes because of changing the parameter H in the HSV color space.

In these circumstances, we have $R = \max(R, G, B)$, $B = \min(R, G, B)$, and $G \in [B, R]$, being these equalities also valid for the (normalized) rgb colors. In other words, the values of R and B (resp., r and b) remain unchanged for $H \in]0°, 60°[$.

Now, recalling the RGB-HSV conversion formula due to Smith [33], and considering]0°, 60°[the domain of H, we get

$$H = 60.\left(\frac{g - b}{r - b}\right) \tag{5}$$

But we know that varying H in]0°, 60°[only provokes changes in the value of g. So, replacing the value of H given by Eq. (5) into Eq. (1), we have

$$60.\left(\frac{g - b'}{r - g}\right) = 60\left(\frac{g - b}{r - g}\right) + 60\left(\frac{g - b}{r - g}\right)\frac{60 - 60\left(\frac{g - b}{r - g}\right)}{60} \tag{6}$$

or, equivalently,

$$g' = g + (g - b)\left(\frac{r - g}{r - b}\right) \tag{7}$$

The novelty about the hue mapping of]0°, 60°[translates itself into a color mapping in the RGB color space given by Eq. (7). That is, there is no need to convert from RGB into HSV and *vice versa*. In short, we have only to remap the rgb reddish colors in conformity with Eq. (7), since $r' = r$ and $b' = b$. Note that in the range]0°, 60°[any hue H has more red than green ($r > g$); cf. line 1 of Algorithm 1.

2. Mapping hues from]300°, 60°]: Let us also assume that the values of S and V remain unchanged. By only changing the value of $H \in]300°, 360°]$, we end up only changing the value of B in the RGB color space. Accordingly, we have $R = \max(R, G, B)$, $G = \min(R, G, B)$, and $B \in [G, R]$, and the same applies to (normalized) RGB colors. Therefore, the values of R and G (resp., r and g) remain unchanged for $H \in [300°, 360°]$.

Now, taking into the account the RGB-HSV conversion formula due to Smith [33], with $H \in [0°, 360°]$, we get

$$H = 60.\left(\frac{g - b}{r - g}\right) + 360 \tag{8}$$

But we know that varying $H \in]300°, 360°]$ only provokes changes in the value of B. So, replacing the value of H given by Eq. (8) into Eq. (2), we obtain

$$b' = g - (g - b)\left(2 + \frac{g - b}{r - g}\right) \tag{9}$$

Note that in the range]300°, 360°[any hue H has more red than blue ($r > b$); cf. line 1 of Algorithm 1. In short, the recoloring procedure that builds upon Eqs. (7) and (9) translates itself into a color remapping in the RGB color space, as shown in Algorithm 1, not being necessary to make the RGB-HSV conversion, and *vice versa*, explicitly. This simplification constitutes a significant gain in processing time.

3. *Changes in perceived chroma and luminance*: The algorithm relies on —though it does not explicitly use— the HSV colour space. It maintains each color's saturation S and value V during recoloring. However, this does not mean that perceived chroma and perceived luminance hold. Indeed, as explained below, maintaining S and/or V in HSV space during recoloring does not maintain the perceived chroma nor perceived luminance of the color.

Regarding perceived chroma, its change is a result of changing the hues to close hues according to Eqs. (1) and (2); for example, an orange is mapped to a yellowish orange. Recall that we tried to reduce the color mapping to a minimum in order not to provoke significant perceived changes in the perceptual learning of deutan and protan dichromats.

As regards to changing the perceived luminance, it is not so obvious because the values of S and V remain unchanged. According to Poynton [35], the perceived luminance is given by

$$L = 0.299R + 0.587G + 0.114B \tag{10}$$

so that, taking also into consideration Eq. (4), we conclude that changes in G provokes changes in L; the values of R and B remain unchanged because the values of S and V do not change in the recoloring procedure. Thus, increasing the value of G results in an increase in the value of L; consequently, the contrast also increases, eliminating the confusion between reds and greens.

Algorithm 1: **RGBeat**
Input: r, g, b
Output: r, g', b'
1 **if** $(r > g) \wedge (r > b)$ **then** `// reddish hues`
2 **if** $(g > b)$ **then** `// reddish hues with` $g > b$
3 $g' \leftarrow g + (g - b)\,(r - g)\,/\,(r - b))$ (7)
4 **end**
5 **else** `// reddish hues with` $b > g$
6 $b' \leftarrow g - (g - b)(2 + (g - b)\,/\,(r - g))$ (9)
7 **end**
8 **end**

Algorithm 2: **TextRecoloring**
Input: HTMLDocument
1 *css*[ ] ← style sheets of HTMLDocument
2 n ← number of style sheets in *css*[ ]
3 **for** $i \leftarrow 1$ **to** n **do**
4 *css*[i] ←i-th style sheet
5 m ← number of CSS rules of *css*[i]
6 **for** $j \leftarrow 1$ **to** m **do**
7 *cssrule*[j] ←j-th CSS rule
8 **if** *cssrule*[j].color **then**
9 [R, G, B] ← *cssrule*[j].color
10 *cssrule*[j].color ← **RGBeat**([R, G, B])
11 **end**
12 **if** *cssrule*[j].backgroundcolor **then**
13 [R, G, B] ← *cssrule*[j].backgroundcolor
14 *cssrule*[j].backgroundcolor ← **RGBeat**([R, G, B])
15 **end**
15 **end**
16 **end**

IV. Text Recoloring For HTML Documents

Algorithm 2 was designed for recoloring text in HTML documents using RGBeat. It involves three main steps, namely:

1. Accessing to all cascading style sheets (CSS) associated with such web page and to all CSS rules associated with each style sheet.
2. Recoloring each text block by changing its corresponding CSS rules whenever necessary. This task performs by changing the `color` and `background-color` properties associated with each CSS rule.
3. Rendering the HTML web page with the modified CSS rules, which is an automatic process provided by any web browser.

Alg. 1 (RGBeat) is used in Alg. 2 to change both text and background. To dynamically access and update the content, structure, and style of an HTML document, we use the HTML Document Object Model (DOM) as a programming interface for Javascript. Indeed, tasks like accessing each style sheet, each rule defined in a style sheet, each value of the rule properties, as well as determining the number of style sheets associated with an HTML document, and the number of rules defined in each style sheet, are all done using DOM object methods.

Thus, recoloring text blocks of a web page is done by changing the CSS rule objects associated to text blocks of an HTML document. This recoloring procedure operates on the CSS rules instead of being on the text blocks themselves. The recoloring procedure of all text blocks of an HTML document corresponds to Alg. 2. Recall that an HTML document usually is tied to a set of CSSs, so that we need to retrieve this set of CSS using the Javascript statement `var css = document.styleSheets;` (cf. line 1 in Alg. 2).

V. Image Recoloring For HTML Documents

Recoloring a single $M \times N$ image is described in Alg. 3. This algorithm calls the RGBeat algorithm (cf. Algorithm 1) for every single pixel of the image. Alg. 3 applies to all images associated with a given web page, as described in Alg. 4.

Algorithm 3: **ImageRecoloring**
Input: *Image*, *width*, *height*
1 **for** $i \leftarrow 0$ **to** *width*-1 **do**
2 **for** $j \leftarrow 0$ **to** *height*-1 **do**
3 $R \leftarrow Image[i][j].R$
4 $G \leftarrow Image[i][j].G$
5 $B \leftarrow Image[i][j].B$
6 [R, G, B] ← **RGBeat**([R, G, B])
7 $Image[i][j].R \leftarrow R$
8 $Image[i][j].G \leftarrow G$
9 $Image[i][j].B \leftarrow B$
10 **end**
11 **end**

Algorithm 4: **HTMLDocumentImagesRecoloring**
Input: HTMLDocument
1 *Images*[] ← set of HTMLDocument's images
2 n ← number of images in *Images*[ ]
3 **for** $i \leftarrow 0$ **to** $n - 1$ **do**
4 *width* ← *Images*[i].width
5 *height* ← *Images*[i].height
6 **ImageRecoloring**(*Images*[i], *width*,*height*)
7 **end**

Recoloring starts with retrieving images from the HTML document (cf. line 1 in Alg. 4). The recoloring occurs in line 6, where one calls the procedure **ImageRecoloring** (i.e., Alg. 3), which in turn calls the

procedure **RGBeat** (i.e., Alg. 1) to adapt the color of each pixel. As seen above, **RGBeat** also applies to recoloring of text and backgrounds from the CSS style sheet rules. Recall that the HTML5 specification provides a 2D context (or even a 3D context) to a canvas to get a pixelized image.

VI. Qualitative, Quantitative, and Time Results

As argued above, we designed the RGBeat recoloring algorithm for deutan and protan dichromat people because of the following: first, they are the most common dichromat people; second, deutan and protan people perceive colors in a very similar manner. As shown above, the RGBeat algorithm applies to text, still images, and video, no matter whether they are on web pages or not.

A. Setup

Before proceeding any further, let us say that we obtained all experimental results using a laptop equipped with a 32-bit Microsoft Windows operating system running on an Intel Pentium Dual CPU T2330 1.60GHz, with 3G RAM. Besides, the algorithms described in this paper were coded in Javascript programming language for HTML5 web browsers, in particular for Chrome.

B. Methodology

For a fair comparison with the RGBeat, we selected the recoloring algorithms due to Iaccarino et a [24] and Ching and Sabudin [30] because they share several features, namely:

- They apply to both deuteranopy and protanopy.
- They use a phenomenal color space. Recall that a phenomenal color space uses hue, saturation, and brightness as classifying descriptors [32], [36] and [37].
- They likely are some of the fastest color mapping methods based on phenomenal color spaces.
- Iaccarino et al.'s method [24] tends to preserve color naturalness at the cost of not reinforcing too much contrast. On the other hand, Ching and Sabudin [30] reinforces the contrast but does not preserve color naturalness.

We also implemented the algorithms proposed by Vienot et al. [31]. These algorithms simulate how deutan and protan people see the colors (see Fig. 4(b) to (d)). Also, we have carried out two sorts of evaluation of the methods: qualitative and quantitative. The qualitative evaluation is visual and subjective, in the sense that we attempt to grasp which is the best method in recoloring process. The quantitative evaluation is objective because it is based on mathematical metrics or formulas and works here as a way of confirming our subjective, visual evaluation.

C. Qualitative Evaluation

As known, deutan and protan people only see some yellows and some blues, although with varying saturation and brightness. Thus, yellows and blues must remain unchanged after the recoloring a given image. Interestingly, greens look unsaturated greens, but indeed they are little saturated yellows. As shown in Fig. 4, the three algorithms leave the yellows unchanged somehow; the same applies to blues (Fig. 5). Note that the original image in Fig. 5 exhibits the primary colors: red, green, and blue.

The differences between those three algorithms become noticeable when recoloring adjacent image elements that people with dichromacy perceive as similar. Recall that deutan and protan people mistake reddish colors (including pinks and oranges) with greenish colors. A glance at Figs. 4-5 shows the following:

Iaccarino et al.'s: Iaccarino et al.'s recoloring technique [24] does not significantly improve the original images when seen by deutan or protan people, though the yellows seem less vivid.

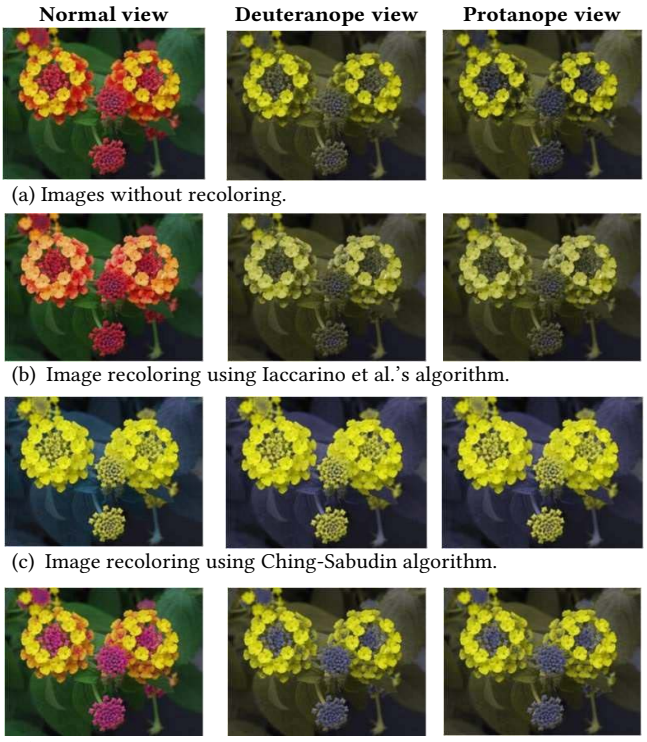
Ching and Sabudin's: On the other hand, Ching and Sabudin's technique [30] maps reds into yellows, which results in a loss of contrast between reds and yellows, i.e., between a primary color and a secondary color. Even worse it is the fact that greens map onto weak blues, so that deutan and protan people can no longer see greens (yet in the form of little saturated yellows).

RGBeat: Our recoloring technique preserves the colors seen by people with deuteranopy and protanopy, i.e., yellows and blues. Reds concerning hues (via RGB-HSV conversion) greater than zero (i.e., reds closer to oranges and yellows) map onto darkish yellows, while reds concerning hues less than 360 (i.e., reds close to pinks and magentas) are mapped to greyish blues.

D. Quantitative Evaluation

As mentioned in section I-B, RGBeat aims at eliminating the red-green ambiguity as much as possible. However, unlike other recoloring algorithms, RGBeat eliminates such color ambiguity without compromising each dichromat individual's lifelong color perceptual learning. Consequently, according to Ribeiro and Gomes [14], RGBeat must preserve the following properties: *consistency*, *naturalness*, and *contrast*.

1) Consistency-based Evaluation: The three benchmarking algorithms are all consistent in terms of recoloring. This means that the equally colored pixels of an image will exhibit the same color after applying the same recoloring procedure.



(a) Images without recoloring.

(b) Image recoloring using Iaccarino et al.'s algorithm.

(c) Image recoloring using Ching-Sabudin algorithm.

(d) Image recoloring using our RGBeat algorithm.

Fig. 4. (a) Images without recoloring as seen by trichromat people (left), deuteranope people (middle), and protanope people (right); (b) images recolored by Iaccarino et al.'s algorithm as seen by trichromat people (left), deuteranope people (middle), and protanope people (right); (c) images recolored by Ching-Sabudin algorithm as seen by trichromat people (left), deuteranope people (middle), and protanope people (right); (d) images recolored by the RGBeat algorithm as seen by trichromat people (left), deuteranope people (middle), and protanope people (right).

2) Naturalness-based Evaluation: As seen before in Section III, the color naturalness ensures that a mapped color will be close to the original color, as perceived by deutan and protan dichromats.

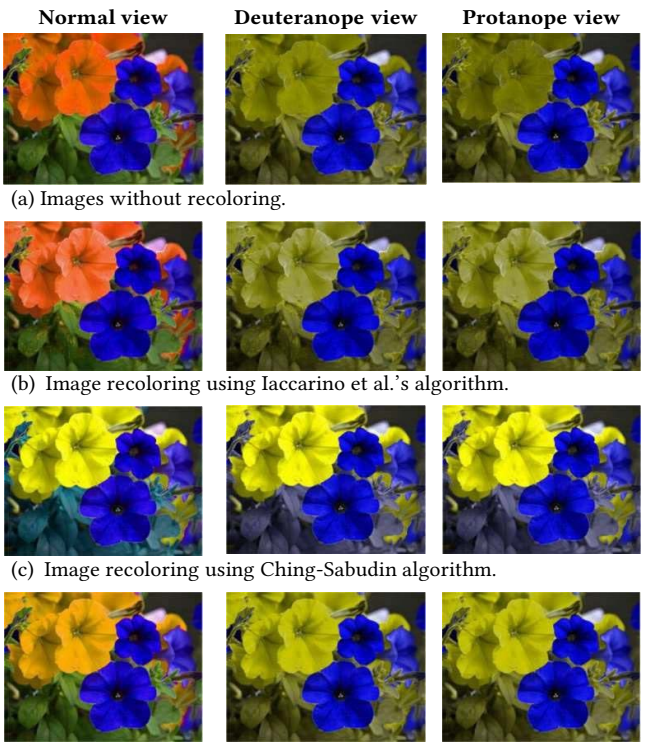
According to Flatla et al. [21], the naturalness of an image has the following formulation:

$$v = \frac{1}{n}\sum_{i=1}^{n} \Delta(P_i, P_i^*) \tag{11}$$

where $n = W \times H$ denotes the image resolution, P_i the color of the i-th pixel, and P_i^* the color of the i-th pixel after the recoloring, while $\Delta(P_i, P_i^*)$ denotes the color difference between P_i and P_i^* in conformity with the CIE76 color-difference formula expressed in CIE Lab space coordinates given by

$$\Delta = \sqrt{\left(L_i^* - L_i\right)^2 + (a_i^* - a_i)^2 + (b_i^* - b_i)^2} \tag{12}$$

where (L_i, a_i, b_i) and (L_i^*, a_i^*, b_i^*) represent the Lab colors of P_i and P_i^*, respectively. The smaller the value of v, the more natural is the recoloring procedure of each image.



(a) Images without recoloring.

(b) Image recoloring using Iaccarino et al.'s algorithm.

(c) Image recoloring using Ching-Sabudin algorithm.

(d) Image recoloring using our RGBeat algorithm.

Fig. 5. (a) Images without recoloring as seen by trichromat people (left), deuteranope people (middle), and protanope people (right); (b) images recolored by Iaccarino et al.'s algorithm as seen by trichromat people (left), deuteranope people (middle), and protanope people (right); (c) images recolored by Ching-Sabudin's algorithm as seen by trichromat people (left), deuteranope people (middle), and protanope people (right); (d) images recolored by our algorithm as seen by trichromat people (left), deuteranope people (middle), and protanope people (right).

Eq. (11) applies to trichromats. However, it also applies to deutan and protan dichromats since we consider that P_i and P_i^* are the colors seen by them before and after the recoloring procedure, respectively. The naturalness benchmarking of the three algorithms was carried out regarding a dataset of 100 still images (Fig. 6 shows 15 of them) possessing reddish colors (i.e., confusing colors for deutan and protan dichromats):

Iaccarino et al.'s: Regarding naturalness, Iaccarino *et al.*'s recoloring technique [24] ranks second, with a mean score of $\bar{v} = 8.6$ for the entire image dataset. We can explain this relatively high score as follows. First, even though Iaccarino *et al.*'s algorihm changes almost all colors, only a few of them change noticeably. Second, the hue rotation resulting from recoloring does not exceed 45° (albeit the cumulative adjustments in saturation and lightness). At the same time, the remaining colors suffer a less pronounced change (about 10% in saturation and lightness).

Ching and Sabudin's: The algorithm proposed in [30] ranks third in terms of naturalness, with a mean score of $\bar{v} = 20.76$ for the entire image dataset. The reasons behind this high score are twofold. First, this happens because 2/3 of colors are subject to recoloring. Second, the reddish hues in the range [-60°, 60°] suffer a rotation that may attain 120°, and the same applies to greenish colors.

RGBeat: Our technique ranks first regarding naturalness, with a mean score of $\bar{v} = 3.8$ for the entire image dataset. To explain this fact, recall that 2/3 of colors remain unchanged, mainly because only those satisfying the condition R > G, B end up being changed. Furthermore, the remapped colors are subject to a rotation with a maximum amplitude of 30°.

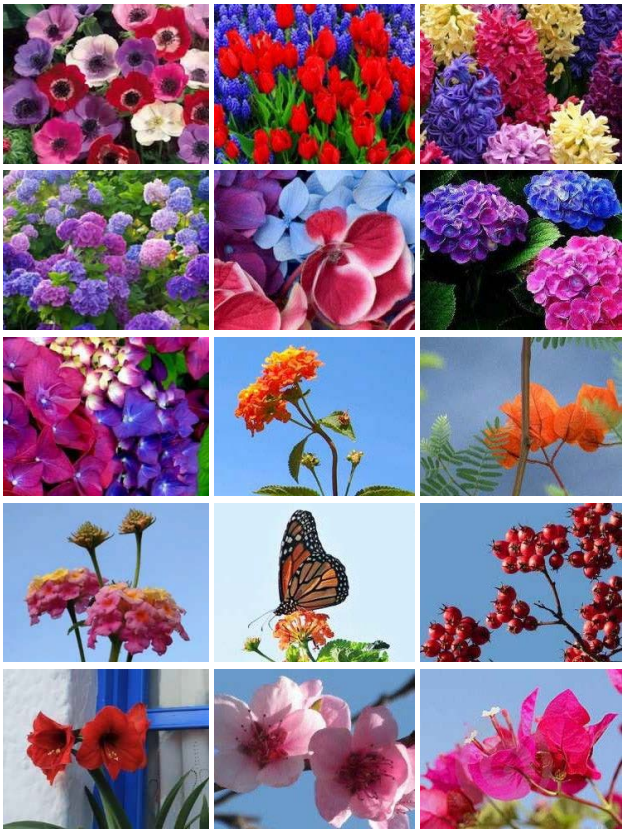


Fig. 6. Still images of a dataset with 100 images (350´270 resolution) that we used to study and benchmark algorithms in respect to the following properties: naturalness, contrast, and performance. All images possess reddish colors (i.e., confusing colors for deuteranope e protanope people), whose hues are in the range [-60°, +60°].

3) Contrast-based Evaluation: The contrast benchmarking of the three techniques was accomplished using the squared Laplacian (cf. [38]) as follows:

$$C = \frac{1}{W \times H}\sum_{x=1}^{W}\sum_{y=1}^{H} G(x,y)^2 \tag{13}$$

where $W \times H$ denotes the image resolution, while $G(x,y)$ is given by:

$$G(x,y) = \sum_{i=x-1}^{x+1} |I(x,y) - I(i,y)| + \sum_{j=y-1}^{y+1} |I(x,y) - I(x,j)| \tag{14}$$

where $I(x, y)$ is the intensity of the pixel (x, y), which in turn is given by (cf. [35]):

$$I(x,y) = 0.299\frac{R}{255} + 0.587\frac{G}{255} + 0.114\frac{B}{255} \quad (15)$$

The computation of contrast C through Eq. (13) applies to any color space. Using Eq. (13), we measured the contrast of the entire image dataset in the deutan (resp., protan) color space. We used the simulation algorithm due to Vienot et al. [31] to generate the entire image dataset as seen by deutan people. Therefore, our quantitative evaluation of contrast took place in the deutan color space. More specifically, in the deutan color space, we obtained a mean contrast value of C=0.0418 for the entire deutan image dataset before applying any recoloring procedure.

The contrast-based evaluation of the benchmarking algorithms output the following results:

Iaccarino et al.'s: Based on our testing, the algorithm due to Iaccarino et al. [24] does not show any contrast improvement because the mean contrast (in the deutan color space) after recoloring all dataset images is 0.0417, i.e., slightly lower than the mean contrast before recoloring. We explain this fact by the small changes in hue, saturation, and brightness inherent to the recoloring procedure of the Iaccarino et al. algorithm.

Ching and Sabudin's: Regarding Ching-Sabudin's technique [30], we noted its ability to enhance the contrast since the entire dataset of images scored 0.047 in mean contrast (in the deutan color space), featuring an increase of 12.4% relative to the mean contrast before recoloring. However, this contrast increase arises at the cost of some loss of naturalness. Indeed, the recoloring procedure remaps 2/3 of colors; but, more importantly, it is the fact that reddish hues in the range [-60°,60°] are all mapped to yellows, when it would be more natural to remap hues in the subrange [-60°,0°[to blues and hues in the subrange [0°,60°[to yellows.

RGBeat: Our technique also improves the contrast, but not so much as Ching-Sabudin's technique; it scored 0.045 in mean contrast (in the deutan color space), featuring an increase of 7.7% relative to the dataset images before recoloring. Recall that our recoloring technique only operates on the hue range [-60°,60°] (i.e., confusing hue range), so that hues in the subrange [-60°,0°[are mapped to blues, while hues in the subrange [0°,60°[are mapped to yellows.

Summing up, considering the color consistency, naturalness, and contrast requirements to maintain each dichromat individual's lifelong color perceptual learning, RGBeat performs better than Iaccarino *et al.*'s and Ching-Sabudin's algorithms. Indeed, they are all color-consistent, but RGBeat is the only one that is capable of enhancing the contrast with negligible loss of color naturalness.

E. Time Performance Evaluation

We encoded the three benchmarking algorithms in JavaScript. To assess their time performance, we used ten offline web pages containing a total of 160 images. Each page incorporates 16 images with identical resolution $n \times n$, but the resolution increases 100 pixels wide and 100 pixels high from page to page, i.e., we have images with resolutions of $n \times n$, with $n = 100, 200, ..., 1000$. Fig. 7 shows how the algorithms behave over images with increasingly higher resolutions. More specifically, Fig. 7(a) shows how much the average time spent by those algorithms depends on the resolution of the images, while Fig. 7(b) features the average time per pixel for each collection of images with identical resolution.

As expected, considering those three algorithms, the average time to recoloring images increases with the resolution (Fig. 7(a)). However, the average time per pixel decreases with the resolution, converging to a minimum that keeps constant when $n \to \infty$ (Fig. 7(b)). Therefore, the time complexity of the three algorithms is linear. To explain this fact, remind that not all the pixels are subject to recoloring. RGBeat is faster than Iaccarino *et al.*'s and Ching-Sabudin's algorithms because its time spent per pixel is shorter than for the other two algorithms. Indeed, the average time per pixel is about 1.4ms, 0.8ms, and 0.06675ms for the algorithms of Iaccarino *et al.*, Ching-Sabudin, and RGBeat, respectively, when $n \to \infty$.

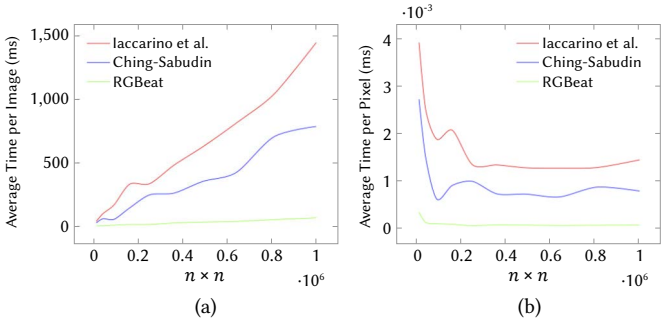


Fig. 7. Recoloring time performance: (a) average time per image against image resolution; (b) average time per pixel against image resolution.

VII. Usability Assessment

In section VI, we compared RGBeat with the other two methods from the algorithmic point of view. To make sure about the quality of the results produced by those algorithms, we proceeded to usability testing, i.e., testing from the user's point of view.

A. The Universe of CVD People

Testing involved a universe of thirteen CVD male volunteers, with ages between 17 and 69 years old. These individuals previously did the D-15 Color Arrangement Test [39]. The results of this test revealed the following distribution:

- 2 people with strong protanomaly (ages 32 and 49);
- 1 person with moderate protanomaly (age 27);
- 5 people with strong deuteranomaly (ages from 49 to 69);
- 4 people with moderate deuteranomaly (ages from 17 to 46);
- 1 person with deuteranopy (age 35).

We extended the usability testing to deutan and protan trichromat people because those with moderate or strong anomalies may have as poor color discrimination as dichromat people [5].

B. The Questionnaire

Although some authors have adopted the Law of Comparative Judgment (LCJ) of L.L.Thurstone for statistical studies [25] [26], we do not follow that pathway because LCJ does not allow the comparison of four alternatives simultaneously. Instead, we use descriptive statistics techniques to compare more than two methods simultaneously [40] and [41]. More specifically, we have used questionnaires as one of such descriptive statistics techniques.

In the makeup of the web questionnaire (see http://rgbeat.ipcb.pt), we considered five categories of images: visualization of information (InfoVis), indoor scenes (Indoor), outdoor scenes (Outdoor), visualization of scientific information (SciVis), and signage (Signage). We elected six images for each category, making up 30, precisely those depicted in Appendix. We selected these images regarding the importance of using a representative set of images (see Shaffer and Zhang [41]) to reduce the sampling error and, consequently, get a significant statistical confidence interval.

The questionnaire was designed for web browsers (via Google

	InfoVis				Indoor				Outdoor				SciVis				Signage				Overall			
Score	MWA	IACC	CHING	RGBeat	MWA	IACC	CHING	RGBeat	MWA	IACC	CHING	RGBeat	MWA	IACC	CHING	RGBeat	MWA	IACC	CHING	RGBeat	MWA	IACC	CHING	RGBeat
1	21	10	31	16	6	33	38	1	5	15	56	2	6	32	30	10	29	7	24	18	67	97	179	47
2	11	24	13	30	17	21	18	22	20	33	5	20	18	14	18	28	25	26	0	27	91	118	54	127
3	17	29	5	27	23	13	7	35	23	21	0	34	16	19	12	31	9	38	1	30	88	120	25	157
4	29	15	29	5	32	11	15	20	30	9	17	22	38	13	18	9	15	7	53	3	144	55	132	59
	(a)				(b)				(c)				(d)				(e)				(f)			

MWA: method without adaptation; IACC: Iaccarino et al.; CHING: Ching-Sabudin.

Fig. 8. Distribution of the preferences per image category and per recoloring method: a) *InfoVis*; b) *Indoor*; c) *Outdoor*; d) *SciVis*; e) *Signage*; and (f) *Overall*.

Forms) so that each category of images corresponds to a separate web page of the questionnaire. Each of these five pages displays 6x4 images, i.e., six rows of 4 optional images organized randomly in a horizontal manner. Each row includes the original image and more three recolored images produced by the above recoloring algorithms (Iaccarino et al., Ching-Sabudin, and RGBeat); these four images concerning the same scene appear randomly in each row. The questionnaire asks the volunteer for his/her preference order (i.e., from 1 to 4) among four images of each row, considering the criteria of naturalness and contrast. Note that this ranking scale is adequate to situations where the ranking involves a maximum of five alternatives [42]. Then, the image ranked first is assigned the score 4, the second the score 3, the third the score 2, and the fourth the score 1 (see [43] for further details about the design of questionnaires). Four specialists on visual representation and color vision deficiency researchers did validate the questionnaire.

C. Data Collecting

Fig. 8 shows the raw quantitative results of the questionnaire, where the CVD people's preferences are expressed relative to five categories of images mentioned in Section VIIB (see Figs. A1-A5 in Appendix A). Considering that we have a universe of 13 respondents and six images per category, the data sample size for each category is 78 responses (= 6´13). For example, for the InfoVis category in Fig. A1, the data sample of the RGBeat is 78, as a result of summing up 16 responses with score 1, 30 responses with score 2, 27 responses with score 3, and 5 responses with score 4. That is, we used scoring in the range [1,4], which has to do with the number of recoloring methods under benchmarking, those three above and the method without any adaptation (original images). Note that the overall data show up in Fig. 8(f).

D. Data Analysis

We carried out data analysis based on descriptive statistics. Specifically, our descriptive statistics-based analysis builds upon data collected from the questionnaire. For that purpose, we used two descriptive analysis tools: (i) box-and-whisker diagrams; (ii) coefficient of variation. The box-and-whisker diagram is a type of data visualization tool that allows us to display the distribution (and, inherently, the concentration) of preferences of the CVD people, while the coefficient of variation is a metric that quantifies such dispersion/concentration of preferences.

The box-and-whisker diagram is a data visualization tool to examine datasets graphically in a quick manner. Its central box at least features 50% of the preferences of the respondents relative to each method. This box consists of the second and third quartiles, separated by the median of the data sample (78 responses). A horizontal straight-line segment represents the median; a cross represents the arithmetic mean.

A glance at the diagrams depicted in Fig. 9, which represent the data listed in Fig. 8, shows us the following regarding each category of images:

- *Infovis*: The method without adaptation (MWA, for brevity) (i.e., original images without adaptation) and Iaccarino *et al.*'s method gathered more than 56% of preferences with the top scores 4 or 3 (see Fig. 8(a) and Fig. 9(a)). Looking at box-and-whisker diagrams of these methods, we observe that their averages are similar (cf. Table I). However, Iaccarino *et al.*'s method preferences are much less dispersed, as its box is smaller than the others. Indeed, Iaccarino *et al.*'s standard deviation is less than the MWA's (cf. Table I), and the same applies to the coefficient of variation. Thus, for InfoVis-type images, the best solution is to use Iaccarino *et al.*'s method.
- *Indoor*: In this case, either MWA or RGBeat reunites 70% of the preferences with scores 4 and 3. Besides, their arithmetic averages are similar (see Fig. 8(b) and Fig. 9(b)). Consequently, even looking at their box-and-whisker diagrams, it is difficult to say which is the top-ranked method of those two methods for Indoor-type images because they have similar visual dispersion (i.e., boxes of the same size). However, we see in Table I that RGBeat's standard deviation is less than the MWA's. Additionally, RGBeat' coefficient of variation is also less than the one of the methods without recoloring. Thus, we conclude that RGBeat ranks first for Indoor-type images.
- *Outdoor*: The results are similar to those obtained for the Indoor category. Indeed, MWA and RGBeat were scored with 4s and 3s in more than 68% of the preferences (see Fig. 8(c) and Fig. 9(c)). Note that their arithmetic averages and dispersion boxes are visually indistinguishable in their box-and-whisker diagrams, so we cannot draw any conclusion about the top-ranked method of those two for Outdoor-type images. However, from Table I, we observe that the RGBeat's coefficient of variation (and standard deviation) is less than the MWA's. Thus, RGBeat ranks first in this category.
- *SciVis*: MWA got the higher scored preferences, with approximately 50% of preferences, and scored 4 (see Fig. 8(d) and Fig. 9(d)). Besides, its arithmetic mean is much higher than for the other methods, which aligns with the fact that its coefficient of variation is smaller than for any other method (cf. Table I). Thus, in SciVis-type images, the best solution is to leave them as they stand, i.e., without color adaptation.
- *Signage*: Ching-Sabudin's method reached 68% of responses scored as 4. Therefore, its arithmetic mean is higher than other methods' one (see Fig. 8(e) and Fig. 9(e)), but its dispersion is significant. On the other hand, Iaccarino et al.'s method ranks second in terms of mean, but its coefficient of variation is smaller than any other method's one. Thus, in Signage-type images, Iaccarino et al.'s method ranks first.

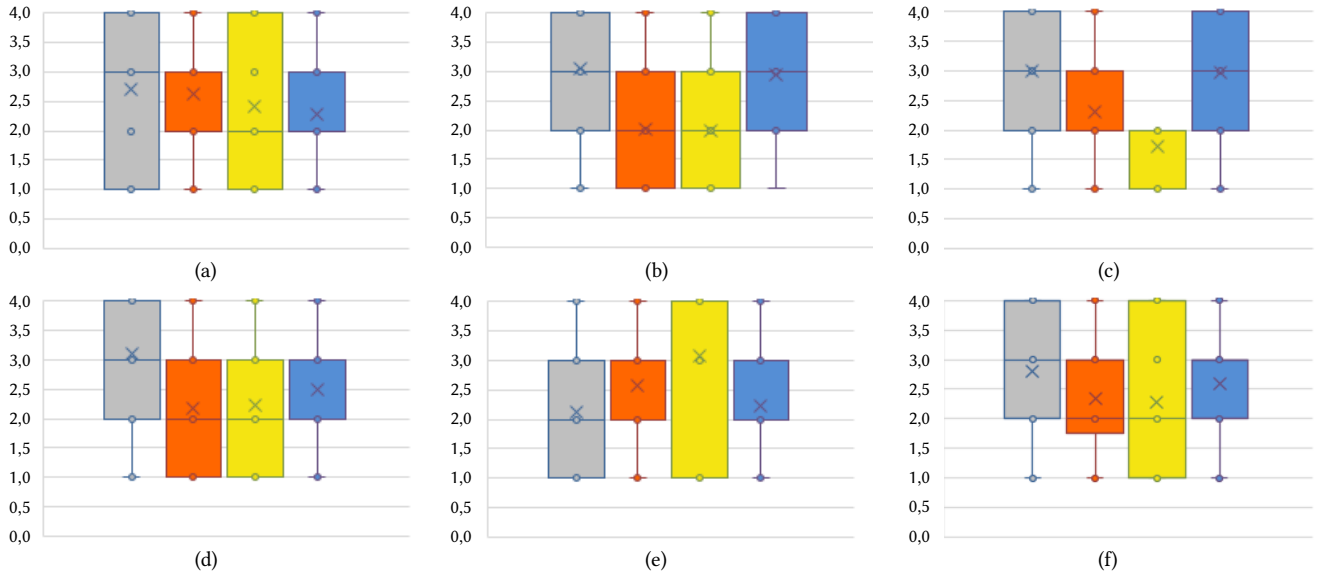


Fig. 9. Box-and-whisker diagrams of the distribution of the preferences per image category and per recoloring method: a) *InfoVis*; b) *Indoor*; c) *Outdoor*; d) *SciVis*; e) *Signage*; and (f) *Overall*. The mean is represented through a cross and the median by the horizontal line closer to the mean mark.

TABLE I. Statistical Results

Category	*Metric*	*MWA*	*Iaccarino et al.*	*Ching-Sabudin*	*RGBeat*
InfoVis	$\bar{x}$	2.69	2.63	2.41	2.27
	σ	1.231	0.941	1.343	0.863
	υ	46%	36%	56%	38%
Indoor	$\bar{x}$	3.04	2.03	1.99	2.95
	σ	0.973	1.081	1.168	0.771
	υ	32%	53%	59%	26%
Outdoor	$\bar{x}$	3.00	2.31	1.72	2.97
	σ	0.953	0.916	1.237	0.805
	υ	32%	40%	72%	27%
SciVis	$\bar{x}$	3.10	2.17	2.23	2.50
	σ	1.014	1.144	1.194	0.864
	υ	33%	53%	54%	35%
Signage	$\bar{x}$	2.13	2.58	3.06	2.23
	σ	1.121	0.782	1.390	0.852
	υ	53%	30%	45%	38%

Statistical metrics: $\bar{x}$: arithmetic mean; σ: standard deviation; υ: coefficient of variation.

In short, the RGBeat method ranks first in two categories, Indoor and Outdoor. Hence it preserves the naturalness more than any other method (see Section VI-D). On the other hand, Iaccarino et al.'s method gets on top in two other categories, InfoVis and signage, although it ranks second in terms of naturalness. However, its contrast varies only -0.24% compared to the contrast of the original images, while RGBeat's contrast is +7.7% (see Section VI-D).

Interestingly, the MWA ranks first in the SciVis category, i.e., there is no variation in naturalness and contrast for apparent reasons. These results show us that the recoloring process must not noticeably change the colors for SciVis-type images to preserve the individual's perceptual learning as much as possible. Recall that RGBeat only changes the red hues to close hues, whereas Iaccarino et al.'s method changes all hues to close hues. In contrast, the Ching-Sabudin method provokes significant color changes, i.e., the distance between an original hue and the mapped hue is greater than for the other two adaptation methods.

Overall, RGBeat ranks first because it ranks first in two categories (Indoor and Outdoor) and second in three categories (InfoVis, SciVis, and Signage), which explains why its coefficient of variation (34%) is lower than in any other method (see Table I). Moreover, it is the only method that outperforms the MWA, i.e., RGBeat images are perceptually better than original images.

VIII. Discussion

Now, we are in a position of highlighting important points of the RGBeat algorithm and its discussion, namely:

- *Color perceptual enhancement.* Adaptation methods are worthy of being investigated because they may enhance the perception of CVD people. Recall that RGBeat makes a noticeable enhancement relative to the original images (MWA). We have also learned that we cannot change the colors too much if we strike on preserving each individual's perceptual learning; otherwise, the naturalness and contrast may change significantly. We have demonstrated that it is possible to increase the color contrast without compromising the image naturalness and perceptual learning of CVD people, resulting in an augmented perception of CVD people. As shown in Table I, RGBeat is the only method that performs better than MWA because its coefficient of variation (34%) is less than the MWA's (40%).
- *Mathematical formulation.* RGBeat's mathematical formulation builds upon RGB and HSV color models, from which we have derived recoloring formulas that exclusively operate on the RGB color model (cf. Eqs. (7) and (10)).
- *Deuteranope and protanope's color space.* By using Brettel et al.'s simulation [7], we show that the deutan and protan color space is 2D, i.e., it consists of two coplanar half-planes (see Fig. 3), because deutan and protan individuals only see two different hues: 60° (yellow) and 240° (blue). A similar result was achieved by Brettel et al. for LMS color space and Meyer and Greenberg [34] for CIE XYZ color space.
- *Generality.* We have shown that RGBeat also applies to images and text in HTML documents. Supposedly, it also applies to video because a video is a sequence of frames. Considering RGBeat spends 0.06675ms per pixel (see Section VI-E) on average, we conclude that recoloring video in real-time is feasible for images with 623,220 pixels; for example, RGBeat can recolor youtube videos in the format 16:9 with resolutions 854×480 at most in real-time.

- *Time performance.* RGBeat attains real-time performance and thus outperforms the other two adaptation methods, primarily because it avoids the explicit conversion between the RGB and HSV color spaces.

Summing up, our method guarantees a balanced trade-off of four requirements, color consistency, naturalness maintenance, contrast improvement, and speed. At the same time, it is capable of augmenting the perception of CVD people.

IX. Conclusions

We have introduced a recoloring method, called RGBeat, that applies to HTML documents, including their images, videos, and text. The main novelty of this method is that it is the only one that produces images that are perceptually richer than the original images. RGBeat was capable of this accomplishment by stressing naturalness maintenance, which imposes limits to the increasing of contrast. Furthermore, RGBeat has revealed quite fast because it only operates on the range of reds, making it feasible to recolor video in real-time.

We have also developed an extension for the Chrome browser that automatically allows for online adaptation of web pages and their contents (e.g., text, still images, and video). Shortly, we intend to use some multi-threading or parallel processing tools (e.g., Web Workers API) to further speed up color-adaptive browsers in real-time.

Appendix A

As mentioned above, we used six images of each of the five datasets (*InfoVis*, *Indoor*, *Outdoor*, *SciVis*, and *Signage*) in the usability testing and assessment. Figs. A1-A5 show such images, as well as their corresponding images recolored by the three benchmarking algorithms.

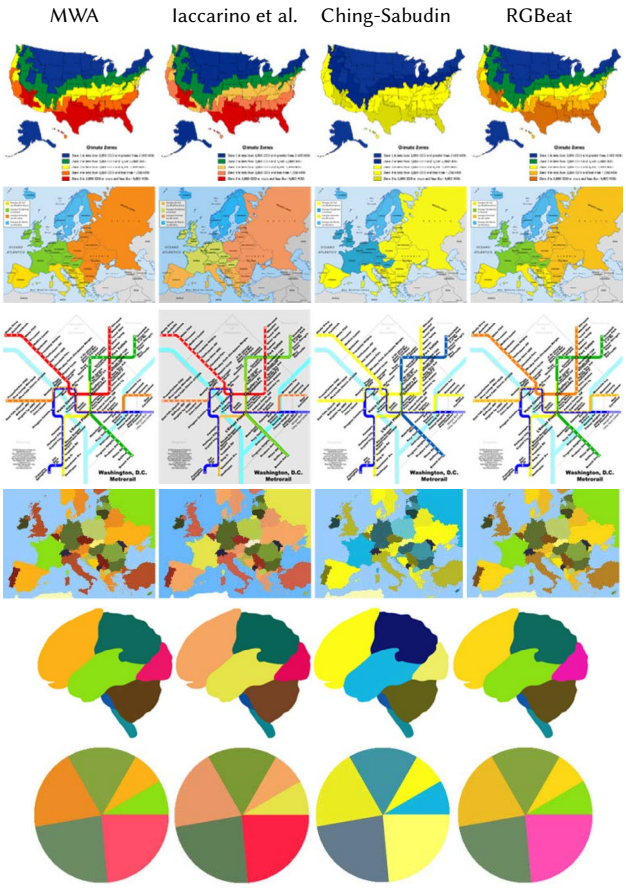


Fig. A1. Six images of the **InfoVis** dataset and recoloring algorithms.

Fig. A2. Six images of the **Indoor** dataset and recoloring algorithms.

Fig. A3. Six images of the **Outdoor** dataset and recoloring algorithms.

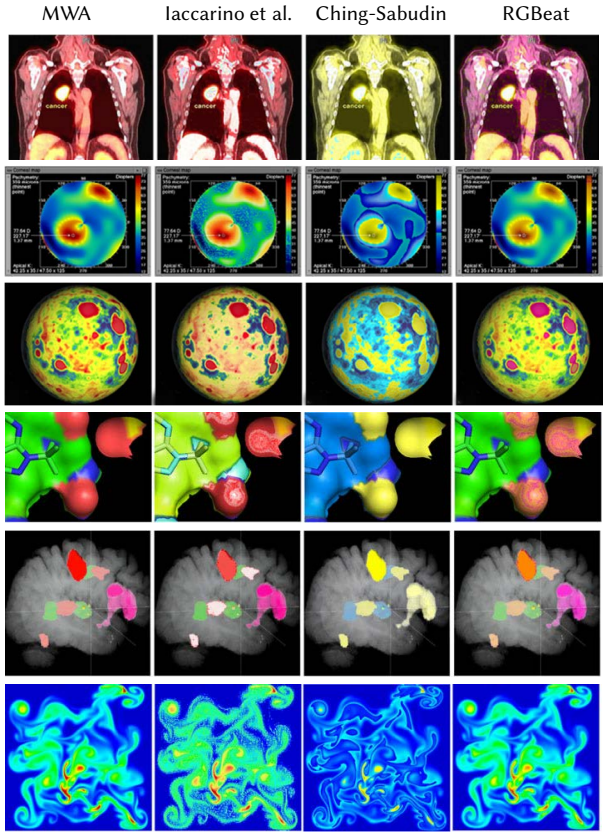


Fig. A4. Six images of the **SciVis** dataset and recoloring algorithms.

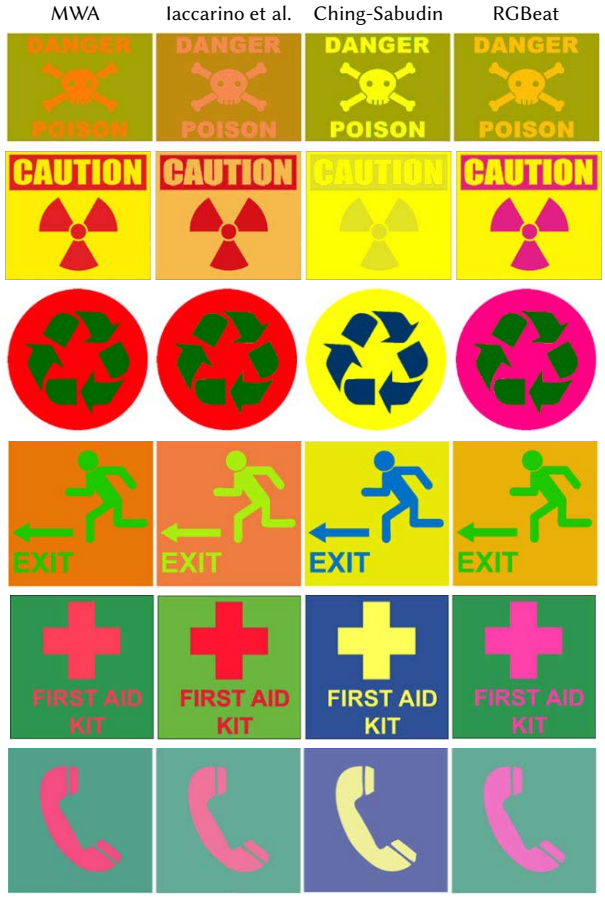


Fig. A5. Six images of the **Signage** dataset and recoloring algorithms.

Acknowledgment

The authors would like to thank Paulo Silveira and Carla S. Pedro for their support in the statistical analysis of the usability questionnaire, Marco Bernardo and Vasco Almeida for their criticism relative to color adaption techniques, as well as CVD participants for their help and time to answer the usability questionnaire.

This work has been partially funded by FCT/MCTES through national funds and when applicable co-funded EU funds under the project UIDB/50008/2020.

References

[1] E. Marieb and K. Hoehn, *Anatomy & Physiology*, Pearson, 2019.

[2] J. Birch, *Diagnosis of Defective Colour Vision*, Elsevier Science, 2001.

[3] J. Pokorny, V. C. Smith, and I. Katz, "Derivation of the photopigment absorption spectra in anomalous trichromats," *Journal of the Optical Society of America*, vol. 63, no. 2, pp. 232–237, 1973.

[4] D. McIntyre, *Colour Blindness: Causes and Effects*. Dalton Publishing, 2002.

[5] L. Sharpe, A. Stockman, H. Jagle, and J. Nathans, "Opsin genes, cone photopigments, color vision and color blindness," in *Color Vision*, K. Gegenfurtner and L. Sharpe (Eds.), Cambridge University Press, 1999.

[6] S. Yang, Y. M. Ro, J. Nam, J. Hong, S. Y. Choi, and J.-H. Lee, "Improving visual accessibility for color vision deficiency based on MPEG-21," *Electronics and Telecommunications Research Institute Journal*, vol. 26, no. 3, pp. 195–202, 2004.

[7] H. Brettel, F. Vienot, and J. D. Mollon, "Computerized simulation of color appearance for dichromats," *Journal of the Optical Society of America.– A: Optics Image Science and Vision*, vol. 14, no. 10, pp. 2647–2655, 1997.

[8] L. Petrich, "Color-blindness Simulators," https://cutt.ly/nboQhuP, accessed on 2021-04-27, 2021.

[9] J. C. Gardner, M. Michaelides, G. E. Holder, N. Kanuga, T. R. Webb, J. D. Mollon, A. T. Moore, and A. J. Hardcastle, "Blue cone monochromacy: causative mutations and associated phenotypes," *Molecular Vision*, vol. 15, pp. 876–884, 2009.

[10] A. Reitner, L. T. Sharpe, and E. Zrenner, "Is colour vision possible with only rods and blue-sensitive cones?," *Nature*, vol. 352, no. 6338, pp. 798–800, 1991.

[11] Y. Miyake, "Rod monochromacy," in *Electrodiagnosis of Retinal Dis-eases*. Springer Tokyo, 2006, pp. 136–137.

[12] OMIN (Online Mendelian Inheritance in Man), "An online catalog of human genes and genetic disorders," Johns Hopkins University, https://www.omim.org/ accessed on 2021-04-27, 2021.

[13] H. K. Kolb, E. F. Fernandez, and R. N. Nelson, "WebVision: The Organization of the Retina and Visual System," University of Utah Health Sciences Center, https://webvision.med.utah.edu/, accessed on 2021-04-27, 2021.

[14] M. Ribeiro and A. Gomes, "Recoloring Algorithms for Colorblind People: A Survey," *ACM Computing Sureys*, vol. 52, no. 4, Art. 71, pp.1-37, 2019.

[15] Y.-C. Chen and T.-S. Liao, "Hardware digital color enhancement for color vision deficiencies," *Electronics and Telecommunications Research Institute Journal*, vol. 33, no. 1, pp. 71–77, 2011.

[16] B.Liu, M.Wang, Y.Linjun, W.Xiuquing, and H.Xian-Sheng, "Efficient image and video re-coloring for colorblindness," in *Proc. IEEE Int. Conf. Multimedia and Expo* (ICME'09), 2009, pp. 906–909.

[17] G. M. Machado and M. M. Oliveira, "Real-time temporal-coherent color contrast enhancement for dichromats," *Computer Graphics Forum*, vol. 29, no. 3, pp. 933–942, 2010.

[18] C.-R. Huang, K.-C. Chiu, and C.-S. Chen, "Temporal color consistency-based video reproduction for dichromats," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 950–960, 2011.

[19] J.-Y. Jeong, H.-J. Kim, T.-S. Wang, Y.-J. Yoon, and S.-J. Ko, "An efficient re-coloring method with information preserving for the color- blind," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 4, pp. 1953–1960, 2011.

[20] J.-B. Huang, C.-S. Chen, T.-C. Jen, and S.-J. Wang, "Image recol- orization for the colorblind," in *Proc. IEEE Inter. Conf. Acoustics, Speech, and Signal Processing* (ICASSP'09), Vols 1- 8. IEEE Press, 2009, pp. 1161–1164.

[21] D. R. Flatla, K. Reinecke, C. Gutwin, and K. Z. Gajos, "SPRWeb: preserving subjective responses to website colour schemes through automatic recolouring," in *Proc. Conf. Human Factors in Computing Systems* (SIGCHI'13). ACM, 2013, pp. 2069–2078.

[22] M. Ichikawa, K. Tanaka, S. Kondo, K. Hiroshima, K. Ichikawa, S. Tanabe, and K. Fukami, "Web-page color modification for barrier-free color vision with genetic algorithm," in *Proc. Genetic and Evolutionary Computation* (GECCO'03). Lecture Notes in Computer Science, vol 2724. Springer, Berlin, Heidelberg, 2003, pp. 2134–2146.

[23] K. Wakita and K. Shimamura, "Smartcolor: Disambiguation framework for the colorblind," in *Proc. 7th Int. ACM SIGACCESS Conference on Computers and Accessibility* (ASSETS'05). ACM, 2005, pp.158–165.

[24] G. Iaccarino, D. Malandrino, M. Del Percio, and V. Scarano, "Efficient edge-services for colorblind users," in *Proc. 15th Int. Conf. World Wide Web* (WWW '06). ACM Press, 2006, pp.919–920.

[25] J.-B. Huang, Y.-C. Tseng, S.-I. Wu, and S.-J. Wang, "Information preserving color transformation for protanopia and deuteranopia," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 711–714, 2007.

[26] G. Kuhn, M. Oliveira, and L. Fernandes, "An efficient naturalness-preserving image-recoloring method for dichromats," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1747–1754, 2008.

[27] S. Oshima, R. Mochizuki, R. Lenz, and J. Chao, "Color-weakness compensation using Riemann normal coordinates," in *Proc. 2012 IEEE Int. Symp. Multimedia* (ISM'12). IEEE Press, 2012, pp. 175–178.

[28] T. Kojima, R. Mochizuki, R. Lenz, and J. Chao, "Riemann geometric color-weak compensation for individual observers," in *Proc. Int. Conf. Universal Access in Human-Computer Interaction* (UAHCI'14). Lecture Notes in Computer Science, vol 8514. Springer International Publishing, 2014, vol. 8514, pp. 121–131.

[29] C.Birtolo, P.Pagano, and L.Troiano,"Evolving colors in user interfaces by interactive genetic algorithm," in *Proc. World Congress on Nature & Biologically Inspired Computing* (NaBIC'09). IEEE Press, 2009, pp. 349–355.

[30] S.-L. Ching and M. Sabudin, "Website image colour transformation for the colour blind," in *Proc. 2nd Int. Conf. Computer Technology and Development* (ICCTD'10). IEEE Press, 2010, pp. 255–259.

[31] F. Vienot, H. Brettel, and J. D. Mollon, "Digital video colourmaps for checking the legibility of displays by dichromats," *Color Research and Application*, vol. 24, no. 4, pp. 243–252, 1999.

[32] A. Ford and A. Roberts, *Colour Space Conversions*. Westminster University, London, United Kingdom1998.

[33] A. R. Smith, "Color gamut transform pairs," *SIGGRAPH Computer Graphics*, vol. 12, no. 3, pp. 12-19, 1978.

[34] G. Meyer and D. Greenberg, "Color-defective vision and computer graphics displays," *IEEE Computer Graphics and Applications*, vol. 8, no. 5, pp. 28-40, 1988.

[35] C. Poynton, *Digital Video and HDTV: Algorithms and Interfaces*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003.

[36] S. D. Cotton, "Colour, colour spaces and the human visual system," University of Birmingham, United Kingdom, 1995.

[37] M. Tkalcic and J. F. Tasic, "Colour spaces: perceptual, historical and applicational background," in *Proc. IEEE Region 8 EUROCON 2003: Computer as a Tool*, vol. A. IEEE Press, 2003, pp. 304–308.

[38] X. Xu, Y. Wang, J. Tang, X. Zhang, and X. Liu, "Robust automatic focus algorithm for low contrast images using a new contrast measure," *Sensors*, vol. 11, no. 9, pp. 8281–8294, 2011.

[39] A. J. Vingrys and P. E. King-Smith, "A quantitative scoring technique for panel tests of color vision," *Investigative Ophthalmology and Visual Science*, vol. 29, no. 1, pp. 50–63, January 1988.

[40] J. W. Tukey, *Exploratory Data Analysis*. Addison-Wesley, 1977.

[41] D. S. Shafer and Z. Zhang, *Beginning Statistics*. Saylor Foundation, 2012.

[42] S.Abeyasekera,J.Lawson-Macdowell,andI.Wilson,"Converting ranks to scores for an ad hoc assessment of methods of communication available to farmers," DFID-funded work under the Farming Systems Integrated Pest Management Project, Malawi and DFID NRSP project R7033, Methodological Framework for Combining Qualitative and Quantitative Survey Methods., Tech. Rep., 2000.

[43] W. E. Saris and I. N. Gallhofer, *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. John Wiley & Sons, 2007.

M. Madalena G. Ribeiro

M. Madalena G. Ribeiro Adjunct Professor in Web Design, Interfaces Design and Usability and Programming at Polytechnic Institute of Castelo Branco, Castelo Branco, Portugal. PhD in Computer Science (from the University of Beira Interior, Portugal), the research topics includes human computer interfaces, design of interfaces, color, color accessibility, color adaptation and image processing. Her publications include articles in international journals, as well as chapters in books and papers at conference's proceeding. Among others, the scientific work includes articles reviewing. She is a researcher member of the Centro de Investigação em Património, Educação e Cultura.

Abel J. P. Gomes

Abel J. P. Gomes is an Associate Professor in Computer Graphics at the University of Beira Interior, Portugal. He obtained a PhD degree in geometric modeling at Brunel University, England, in 2000. He has over 100 publications, including journal and conference articles, and 1 book published by Springer-Verlag. He was Head of the Department of Computer Science and Engineering, University of Beira Interior, Portugal, and the leader of a research unit of Instituto de Telecomunicações, which is one of the biggest research centers in Portugal. He is also a licensed Professional Engineer and member of the IEEE, ACM, and Eurographics. His current research interests include color accessibility, computer graphics algorithms, molecular graphics, geometric computing, and implicit curves and surfaces.

Validity and Intra Rater Reliability of a New Device for Tongue Force Measurement

Marta Carlota Diaz-Saez[1,2], Hector Beltran-Alacreu[2,3]*, Javier Gil-Castillo[4], Gonzalo Navarro-Fernández[1,2], José Luis Cebrián Carretero[5], Alfonso Gil-Martínez[1,2,6]

[1] Physiotherapy Department, Centro Superior de Estudios Universitarios La Salle, Universidad Autónoma de Madrid, Madrid (Spain)
[2] CranioSPain Research Group. Centro Superior de Estudios Universitarios La Salle. 28023 Madrid (Spain)
[3] Toledo Physiotherapy Research Group (GIFTO), Faculty of Physical Therapy and Nursing, Universidad de Castilla-La Mancha. (Spain)
[4] Neuro-Rehabilitation Group. Instituto Cajal. Consejo Superior de Investigación Científicas (CSIC) (Spain)
[5] Maxilofacial and Oral Surgery Department, Hospital Universitario La Paz, Madrid (Spain)
[6] Unit of physiotherapy, Hospital La Paz Institute for Health Research (IdiPaz), Madrid (Spain)

Received 10 March 2021 | Accepted 23 December 2021 | Published 2 February 2022

Abstract

Background. The tongue is made up of multiple muscles both extrinsic and intrinsic. The hyoid, jaw and maxillary complex contain the tongue, which hangs between these structures forming an important biomechanical system. This organ has to work in coordination with craniofacial structures to ensure normal orofacial functioning. There are different devices on the market for tongue force measurement. However, they are not accessible for patients due to their size and very high prices. Likewise, other devices have not yet carried out validity and reliability studies. The purpose of this study was to validate a new device proving that it is accurate compared to the algometer. Moreover, the study wanted to determine the intra-rater reliability of a protocol to assess the maximum tongue force in asymptomatic subjects. *Material and methods.* This is an observational-longitudinal study with repeated measurements. A prototype device was developed specifically for this study to measure tongue force through force-sensitive resistor sensors. The prototype system was equipped with a device to perform and transmit the measurement and a C++ programming software in the computer to take data from the session. Different formulas were made to calibrate the system. For validity, the force measured by the prototype and the algometer was compared. For intra-rater reliability, 29 asymptomatic Spanish subjects were recruited, and a standardized protocol was carried out for the tests. *Results.* Experiments to assess validity showed a strong correlation (r>0.97) and an excellent reliability (ICC>0.90) between devices. On the other hand, the intra-rater reliability analysis showed an excellent ICC (0.93) with a 95% CI of 0.86 to 0.97 and a MDC_{90} of 6.26N. *Conclusion.* We demonstrated good validity values and high intra-rater reliability for the prototype device for the maximum tongue force.

Keywords

Device, Reliability, Temporomandibular Disorders, Tongue Force, Validity.

DOI: 10.9781/ijimai.2022.02.001

I. Introduction

The tongue is a muscle formed by extrinsic and intrinsic muscles which are responsible for tongue movements and tongue shape changes, respectively [1]. This organ belongs to the stomatognathic system and has to work in coordination with other craniofacial structures for the normal orofacial functions [2]. The hyoid, jaw and maxillary complex contain the tongue, which hangs between these structures forming an important biomechanical system [1], [2]. The main functions the tongue participates in are swallowing, mastication and speech, highlighting its important role during the oral and the pharyngeal phases of swallowing [3], [4]. In swallowing, the tongue has to generate a specific pressure to control and push the food bolus ensuring a well transportation from the oral cavity to the stomach. The applied force depends on the type of food bolus and must have to be coordinated with the jaw and hyoid movements [5]–[11]. During the tongue pressure against the anterior part of the hard palate, the hyolaryngeal complex has to elevate itself allowing a decline of the epiglottic cartilage to protect the airway. This coordinate process decreases the risk of suction increasing safety in the swallowing process. Therefore, any disturbance or pathology in one of these structures may result in a malfunctioning of the others. That is why, a reduction in tongue force could avoid the correct formation and placement of the food bolus. Moreover, a well propel and transport

* Corresponding author.
E-mail address: hector.beltran@uclm.es

could be compromised with a drop in the tongue force. Likewise, the amount of waste accumulated in the oral and pharyngeal cavity could be increased and other symptoms associated with dysphagia can also be generated [5]–[11].

Different studies have shown changes in tongue force in orofacial pathologies, such as swallowing disorders, motor speech disorders, oromotor deficits or temporomandibular disorders [2], [12], [13]. Interest in studying tongue force in pathologies where these organ functions are limited or affected has led to the design of measuring systems with different force or pressure sensors. Some of these systems are dynamometers, miniaturized pressure transducers and electromyographic monitoring techniques. Their purpose is to offer an objective characterization [14].

In this study, a lower-cost system available for use at home to assess and train tongue force with different dynamics is presented. The importance of this device's implementation lies in its ability to reduce social and health care costs and waiting lists of the health system, as well as promoting patient independence at home. Because this prototype system is portable, handy and has an intuitive interface, it becomes a tool not only for professionals, but also for patients. All these things are very important to ensure home treatments based on exercises that offer patients feedback about their measurements. Moreover, these home training exercises may allow patients to maintain long-term effects of the therapy and avoid the overcrowding of health care services [15].

At this time, there are similar devices in the market which research has demonstrated to be useful in measuring tongue force in several scenarios [16]–[18]. However, these devices present some limitations when compared to the prototype of this study (Table I). This is because certain systems, such as the Iowa Oral Performance Instrument (IOPI), involve an expense of around 800-2000€, whereas the complete product presented in this study costs 90-100€ (estimated price) [19]. In addition, non-portable devices such as the Kay Swallowing Workstation (KSW) decreases the possibility of being accessible for patients [18]. Furthermore, the limited durability of the system components of other devices is another shortcoming that we try to eliminate in our system by using force-sensitive resistor (FSR) sensors. These sensors do not present mobile parts, which reduces the risk of breaking or wearing out. Moreover, they are not significantly affected by the noise or vibration and can operate between -30° and 70° degrees [20]–[22]. In addition, allergies have to be considered since they are intraoral measurements. The current literature does not mention tongue force devices made with hypoallergenic materials. Some studies show devices made with latex materials [23], while other researches do not provide information about the intraoral system's materials [3], [18], [19]. For this reason, we developed a device which provides a single-use hypoallergenic protection covering the FSR sensor like a sleeve made of nitrile.

Based on the above, the aim of the study is to validate the new device comparing the measurements to the algometer. The research also aims to determine the intra-rater reliability of a protocol to assess the maximum tongue force in asymptomatic subjects.

II. Methods

This study is an observational-longitudinal study with repeated measurements. The research was approved by the ethical committee of the Centro Superior de Estudios Universitarios La Salle (CSEULS-PI-036/2019). For the intra-rater reliability, asymptomatic individuals were recruited from the Hospital Universitario La Paz according to the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) 2010 [24].

The research team was composed of a biomedical engineer and four experienced physical therapists. The biomedical engineer and two of the physical therapists (involved in the study) developed the measurement system for the validity, while the other physical therapists carried out the intra-rater reliability. Finally, all the data collected was analyzed by one of the assessors.

A. Study Description

1. Validity

Participants were not needed for the validity process. The biomedical engineer and two experienced physical therapists developed the measurements with the device and an algometer.

Wagner algometer (FPX25, Wagner Instruments, Greenwich CT, USA) was used in this study. The algometer is a device that can be used to standardize the intensity of palpation or to measure the degree of pressure used to evoke a painful response. It is a system that can be used to make kinesthetic measurements in an objective way. The algometer consists of a flat circular rubber tipped probe of 1 cm diameter attached to a soft-grip handle piston connected to a pressure sensitive strain gauge transducer, called a load cell. The force applied to the end of the probe is transmitted to the load cell and a voltage output is produced. This voltage output is directly related to the pressure exerted and allows for its calculation. Finally, the result is displayed on the LCD numerical display in lbf, Kgf, N or ozf. Furthermore, the device presents a calibration mechanism to measure and represent values in a graduate scale expressed as Newton or Kgf. These characteristics make the algometer suitable as comparison equipment.

TABLE I. Comparison Between the Most Common Tongue Devices in the Market

Item	New Device	IOPI	KSW	MOST
System	Computerized system with 1 FSR sensor	Silicon air-filled bulb	Computerized system with 3 air-filled sensors	Computerized system with 4-5 sensors inside 1 intraoral piece
Stability of the sensors	Good	Poor	Poor	Good
Size	Small	Small	Large	Medium
Portability	Portable	Portable	Non-portable	Portable
Price	Cheap (90-100€)	Expensive (1.190€)	Unknown	Expensive (800-34.000€)
Reliability	Yes	Yes	No	Yes
Validity	Yes	No	No	No
Visualization of signal	Excellent	Good	Excellent	Good
Data integrity	Good	Poor-movement artefacts	Good-with fixed sensors	Good
Patient usage	Hand held device	Hand held device	Fixed position	Fixed position

2. Reliability

A total of 32 asymptomatic Spanish subjects were recruited from the Hospital Universitario La Paz. Three subjects had to be excluded so a total of 29 asymptomatic participants were included in the study. The study was publicly advertised in different places so that people interested in the trial could participate. A non-randomized convenience sampling was performed for this study.

The whole procedure was developed from September 2019 to December 2019. The inclusion criteria were as follows: age between 18 and 65 years old; had not experienced any craniofacial/ temporomandibular/neck pain; had no facial palsy caused by a primary muscle disorder; and had no significant history of chronic pain disorder.

The exclusion criteria were: a surgery or history of traumatic injuries of the neck/head/face/tongue/teeth/jaw; cancer or an active infection of the neck/head/mouth; rheumatic disorders; neurological disorders; and pregnancy.

B. Instrumentation

For this study, a prototype device was specifically designed and developed to measure tongue force. The system consists of two main parts: device and software. The device is in charge of performing the measurement and transmission of the tongue force, while the software, located in the PC, allows taking data from the session (movement to be measured and time measured) and from patients (name, date of birth, sex and pathology) (Fig. 1). Moreover, the software shows tongue force in real time and stores the data corresponding to the session in an Excel database.

The device employs a piezo-resistive sensor to acquire force. This sensor, known as FSR sensor, is a type of technology which resists changes while a pressure is applied on the active surface. There is a negative correlation between resistance and force exerted. This means that the greater the force exerted, the lower the resistance that is generated [25]. For measurement of tongue force in the current study, a circular FSR sensor (Model: 402, Interlink Electronics, CA, USA) with 0.46 mm nominal thickness, an area of 12.7 mm diameter at the top and a measurement range of 0 to 10 kg was chosen [20]. This type of sensor has proven to be a good option compared to others on the market when dynamic measurements are required [22].

To measure the resistance of the FSR sensor, a voltage divider circuit was built with two resistors arranged in series, where one of them is the FSR [20]. The following formula was then applied to measure the FSR resistance (R_FSR):

$$R_{FSR} = \frac{R}{V_R} \cdot (V - V_R) \quad (1)$$

V: input voltage, V_R: resistance voltage, R: resistor. In our case, V = 5V and R = 10 kΩ

The resistance voltage was measured with a C++ program developed in Arduino UNO, which uses an analog to digital converter (ADC) module of a 10-bit Atmega328P microcontroller to transform the analog signal to a digital signal whose values are between 0 and 1023.

Since the variation of the resistance is not linear, it was decided to use conductance to calculate the force, as this allows interpretation on a linear scale. The conductance (C) is the inverse of the resistance, i.e.:

$$C_{FSR} = {}^{1}/_{R_{FSR}} \quad (2)$$

Using the information provided in the sensor's own datasheet [20], an approximate calculation of the force was made by relating it to the conductance and dividing it into two force ranges (from 0 to 1 kg and from 1 to 10 kg). For the small force range (0 to 1 kg) the following linearized approximation of the force-conductance relationship was used:

$$F_{FSR}[N] = {}^{C_{FSR}}/_{80} \quad (3)$$

Likewise, for the big force range (1 to 10 kg), the following linearized approximation of the force-conductance relationship was used:

$$F_{FSR}[N] = {}^{C_{FSR}}/_{30} + {}^{1000}/_{80} \quad (4)$$

This information is sent in real-time via USB communication to the PC, where it is displayed at the same time and it can be collected by the software developed in Java-based Processing. The interface has been designed and developed to be simple and intuitive for use in the clinical environment. This user-friendly interface, available for Windows, MacOS and Linux operating systems, facilitates the use of the system and the performance of specific measurements for a predetermined period of time by the professional for subsequent registration. In addition, it has been adapted for use on computers with different screen sizes. Currently, the movements enabled for recording in the interface are: lip to lip, tongue protrusion, tongue elevation, tongue depression, right tongue lateralization and left tongue lateralization. However, the system allows us to measure any type of movement, as long as the force is applied to the active surface of the sensor.

During measurements on subjects, a single-use hypoallergenic

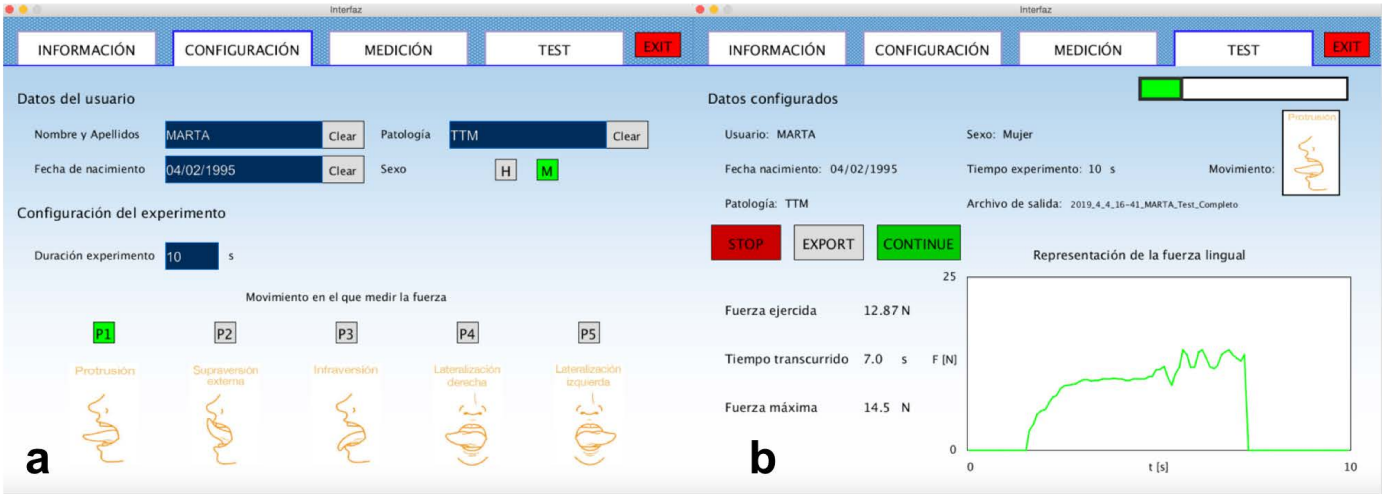


Fig. 1. Spanish prototype device interface view. a: Configuration screen view; b: Measurement screen view.

protection made by nitrile is used to isolate the system from the patient so that the same system and sensor can be used between patients. In addition, this protection allows us to isolate the sensor from the temperature, which can be added by inserting it in the mouth to make the corresponding measurement. Due to this insulation, the temperature does not seem to be a parameter that can influence the measurement made.

C. Procedure

1. Validity Procedure

For the validation of the system, conditions were set-up in order to reproduce the environment in which the measurement was going to be performed. In addition, an algometer was used to validate the measurements that the system collected. For this purpose, the system remained fixed on a flat horizontal surface to maintain the position and avoid possible variations that can be introduced by an uneven surface. The sensor was covered by the hypoallergenic protection, as is carried out in measurements with patients. The purpose was to imitate the actual measuring environment and limit the influence of temperature. With the help of some markers on the table, the algometer was placed so that it was always in contact with the same point of the sensor and supported by the sensor's entire active surface. The perpendicularity between algometer and FSR sensor was maintained during the experiment with the help of a level and two vertical supports to which the algometer was fitted. The aim of this was to maintain repeatability during the whole experiment.

Once all the experimental conditions were configured, the measurement process started connecting the device and opening the software for 5 minutes to stabilize the whole system. Measurements corresponding to 2N, 5N, 10N, 15N, 20N, 25N, 30N, 35N, 40N were obtained in the algometer for each FSR sensor. The adjustment to the desired value was made by gradually increasing the pressure, and independently for each measurement, for approximately 10 seconds. During the last 3 seconds a camera system was used to record the matching values of the measurement system and algometer and their subsequent analysis.

2. Reliability Procedure

Asymptomatic subjects were recruited for the reliability measurements. First of all, an informed consent was given to the participant. Secondly, they filled out some sociodemographic data and questions to know if they fulfilled the inclusion criteria. They were then asked about their initial subjective perception of fatigue before starting the measurements to ensure the subject started without it, and the assessor checked the proper functioning of the system. This variable was measured using a numeric verbal fatigue scale that consists on asking subjects quantify their perception of fatigue between 0 ("no fatigue") and 10 ("maximum fatigue conceivable") [26]. After that, participants sat down in front of the device and put a single-use hypoallergenic protection covering the sensor like a sleeve made of nitrile. They had to protrude the tongue pressing the sensor with the tip of their tongue against the anterior part of the hard palate (Fig. 2) for 10 seconds with a verbal feedback from the assessor. This allows to assess tongue fatigue. They rested for 5 minutes and were again asked about their subjective perception of fatigue to ensure the subject performed the measurement without it. Finally, they repeated the measurement again trying to replicate the way the first measurement was conducted. All measurements were collected in a specific results paper for each subject, which were kept by the principal investigator.

D. Sample Size Calculation

Sample size was calculated using the method based on the Intraclass Correlation Coefficient (ICC) [27], [28]. Based on the ICC obtained with the first 10 subjects of the present study, it was estimated that our ICC under the alternative hypothesis was 0.90. A sample of 26 with 2 observations per subject was needed to achieve 80% power to detect an ICC of 0.90, when the ICC under our null hypothesis was 0.75 using an F-test with a significance level of 0.05. The final sample size was 29 to allow for possible dropouts (10%).

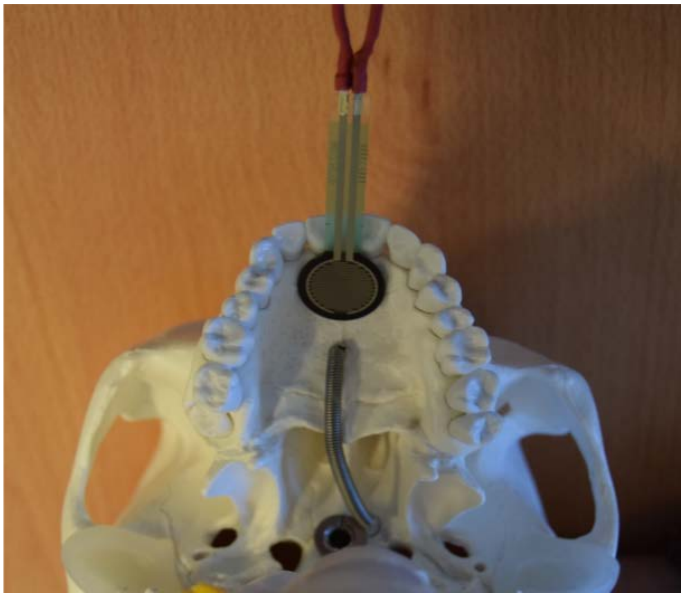


Fig. 2. Force pressing tongue against the superior part of the hard palate.

E. Statistical Analysis

1. Experiment to Assess the Validity of the Device

Some measures were made simultaneously to compare the algometer and our device with two different FSR sensors. Those measures were made in some specific values of the algometer: 2N, 5N, 10N, 15N, 20N, 25N, 30N, 35N and 40N. Thus, a total of 9 measurements for each FSR sensor were made. Those sensors will be named FSR-A and FSR-B to make it easier for the reader to understand the experiments. Two tests were made: the first one comparing algometry with FSR-A and the second one comparing algometry with FSR-B. Secondarily, a third comparison was made combining the data of test 1 and test 2. All tests included the Shapiro-Wilk test to assess normal distribution of the data, Pearson's Correlation test to analyze algometry-device correlation, the Student's t test to assess algometry-device mean differences, and the intraclass correlation coefficient ($ICC_{3,1}$) to assess agreement between devices. Pearson's correlation test values above 0.60 were considered a strong correlation [29]. Finally, another $ICC_{3,1}$ was made to assess reliability between FSR-A and FSR-B.

2. Reliability Analysis

A total of 5 $ICC_{3,1}$ were calculated using SPSS statistical software (SPSS, Inc, Chicago, IL, USA): 4 included in the validity experiment and 1 to assess intra-rater reliability with healthy volunteers. The $ICC_{3,1}$ was designated as the two-way analysis of variance mixed model for absolute agreement of single measures. Interpretation of the ICC was performed according to previously published categories: <0.50 is poor agreement, 0.50-0.75 is moderate agreement, and >0.75 is good to excellent agreement [30].

Bland-Altman plots were constructed only for intra-rater reliability using mean differences between measurements [31]. Limits of agreement (LOA) were calculated as mean differences ± (standard deviation multiplied by 1.96) [32]. Calculation of the occurrence of systematic or random changes in the data means that it was performed

TABLE II.A. Pearson's Correlation and Student's T Test for Validity Tests (Newtons)

Test 1	**Algometry**	**FSR A**	**Partial results**	**Total results**
	20.28 ± 13.37 SW: p=0.83	18.5 ± 15.1 SW: p=0.54	r=0.99 p<0.01 t=2.08 p=0.07	
Test 2	**Algometry**	**FSR B**	**Partial results**	
	20.25 ± 13.31 SW: p=0.83	20.8 ± 18.0 SW: p=0.25	r=0.97 p<0.01 t=-0.28 p=0.79	r=0.97 p<0.01 t=0.56 p=0.58
Total (test 1 + test 2)	20.27 ± 12.94 SW: p=0.25	19.7 ± 16.1 SW: p=0.09		

TABLE II.B. Between-Device and FSR A- FSR B Reliability Tests (Newtons)

Measure	**Mean difference**	**ICC (95% CI)**	**SEM**	**MDC90**
Algometry – FSR A	1.77	0.98 (0.88 to 1.00)	1.96	4.57
Algometry – FSR B	-0.55	0.94 (0.75 to 0.99)	3.76	8.78
Algometry – FSR total	0.61	0.95 (0.88 to 0.98)	3.23	7.53
FSR A – FSR B	0.45	0.97 (0.85 to 0.99)	2.80	6.53

Results are showed as mean ± standard deviation.
SW: Shapiro-Wilk test; ICC: intraclass correlation coefficient; CI: confidence interval
SEM: standard error of measurement; MDC: minimal detectable change; FSR: force-sensitive resistor.

TABLE III. Intra-Rater Reliability Analysis and Bland-Altman Plot (Newtons)

Reliability Analysis	
Measure 1	19.89 ± 9.59 [a]
Measure 2	19.44 ± 10.82 [a]
ICC (95% CI)	0.93 (0.86 to 0.97)
SEM	2.68
MDC90	6.26
Bland-Altman Plot	
Mean difference	0.45 ± 3.79 [a]
95% CI mean difference	-0.93 to 1.83 [b]
Limits of Agreement	-6.98 to 7.88

[a] data showed as mean ± standard deviation; [b] 0 is included in the CI (p >0.05)
ICC: intraclass correlation coefficient; CI: confidence interval; SEM: standard error of measurement; MDC: minimal detectable change.

through a calculation of 95% confidence intervals of the mean differences between the values of the measurements.

Measurement error was expressed as SEM. The SEM was calculated as $\sqrt{(RMS)}$ x $\sqrt{(1-ICC)}$, with RMS being the root mean square [30].

Responsiveness was determined with minimal detectable change at 90%, which was calculated as SEM x 1.65 x $\sqrt{2}$ [33], [34]. The MDC_{90} expresses the minimal change required to be 90% confident that the change observed between two measurements reflects a real change and not a measurement error.

III. Results

A. Experiments to Assess the Validity of the Device

All measures showed normal distribution in the Shapiro Wilk test (p>0.05). Both partial and total results showed a strong correlation between devices (r>0.97) and no statistically significant mean differences (p>0.05). In addition, between-device reliability analysis and reliability analysis for both FSR measures showed an ICC > 0.90. The results of all the experiments are summarized in Tables II.a and II.b.

B. Intra-rater Reliability

A total of 32 subjects were recruited for eligibility. Of those, 3 subjects did not meet the inclusion criteria because they had temporomandibular joint disorder diagnosed. Finally, 29 healthy subjects were included in the reliability analysis (17 men and 12 women aged 40.18 ± 14.99). The Shapiro-Wilk test showed normal distribution of the data (p>0.05). Descriptive data, $ICC_{3,1}$, SEM, MDC_{90} and the Bland-Altman analysis with 95% CI and LOA are summarized in Table III. The ICC obtained was excellent (ICC = 0.93), with a 95% CI of 0.86 to 0.97. Graphical representations of the Bland-Altman plot are shown in Fig. 3.

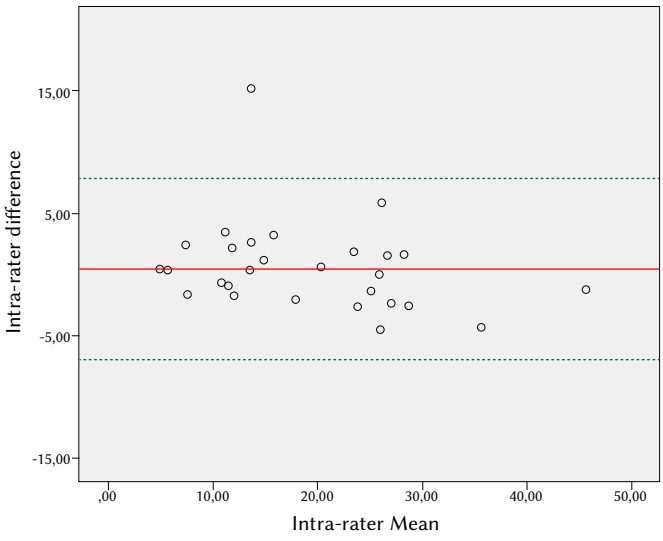


Fig. 3. Bland-Altman plot.

IV. Discussion

The main objective of the present study was to verify the validity and intra-rater reliability of a new device for the measurement of maximum tongue force using an FSR sensor. To the authors' knowledge, this is the first validity and intra-rater reliability study of a device that uses an FSR sensor to measure tongue force.

At present, there is no test or device that has become the gold standard for measuring tongue force [35]. Although the IOPI is the most studied and used device for assessing tongue force [16], [17], [36], [37], it still does not have a validity study. In addition to the IOPI, there are different devices to measure tongue force, such as the Madison Oral Strengthening Therapeutic (MOST) [38] and the KSW [39], but they have not conducted validity studies either. For this reason, we do not have enough data to compare our validity results with other tongue force measurement devices. However, we know that the Interlink 402 FSR sensors used in this research have demonstrated great precision in force measurements for dynamic exercises against other FSR sensors [22]. FSR sensors have shown a reliability of the 93% for bite force measure in several dentistry studies [40]. Moreover, Ahmed et al. observed an accuracy of over 95% in dynamic exercises for FSR sensors based on previous results [22], [25]. These sensors are characterized by their easy integration, their small size, their low weight and their low cost. All of this combined with the existing scientific literature supporting their reliability are the main reasons they are chosen for noninvasive force measurements [22], [41]–[43].

The Adam et al. study showed a good intra-rater reliability for the IOPI with high or very high ICC (0.77-0.90) in asymptomatic subjects (n=51) with an average age of 28.2 ± 9.3 [16]. Further, White et al. reported an excellent intra-rater reliability (ICC 0.92) for the KSW, with a 95% IC (0.85-0.97) in asymptomatic subjects with an average age of 64.7 ± 10.2 [3]. From our perspective, the device presented in this research seems to be a very good choice since the intra-rater reliability is slightly higher than the IOPI and very similar to the KSW. Its advantage over the KSW is its greater portability. There are few studies in which standard values of tongue pressure measurements were considered [44], [45]. In all those studies, the IOPI device or an IOPI-like design device was used. For example, Utanohara et al. showed an average maximum tongue pressure between 41.7 kPa (kilopascal) in young subjects to 31.9 kPa in old subjects, using a disposable tongue pressure measurement device with an IOPI-like design [44]. Also, Crow et al. carried out a similar experiment using IOPI and showed similar results (66.7 to 74.8 kPa) [45]. Note that both devices expressed results in pressure units (kPa) while our new device expressed results in force units (Newtons) so we cannot directly compare results. If our results were transformed to pressure units, it could be obtained, approximately, 155.5 kPa as maximum tongue force. This difference could be justified because the measurement protocols of the IOPI and our sensor are different. Crow et al. mentioned a protocol of measurement [46] in which it can be observed that raising the tongue and compressing the bulb of the IOPI onto the hard palate is not the same movement than making a protrusion of the tongue against the posterior side of superior incisor and palate with the tip of the tongue.

Regarding the SEM and MDC_{90} values, the results for the studied device were relatively low. It has to be considered that the way the SEM is calculated may differ between studies. Adams et al. used the standard deviation and our research used the RMS [16]. The RMS has been considered an alternative way to calculate the SEM, avoiding the uncertainties generated by the selection of the ICC type [30]. However, the low results for the SEM and the MDC_{90} indicates an agreement between studies in spite of the differences mentioned.

The available scientific evidence examines other devices that measure tongue force, such as the MOST or ad hoc devices, some of which have been compared with and formed partnerships with the IOPI and the KSW [18]. Nevertheless, there is an important limitation since these other devices have not yet carried out validity and reliability studies [18], [23], [47], [48]. For this reason, we cannot compare our results with these ad hoc devices. However, we can point out the existing differences between the devices on the market and the device used in this research. First of all, many of them have a limited durability because of the components used. Secondly, they have very high prices, poor stability systems or sensors that keep the mouth open, which affects the position of the different orofacial structures and hence tongue position [23], [48]–[50]. In contrast, the studied device is cheaper (90-100€), brings more stability due to the smooth sensor surface, and its small size avoids affecting the anatomical systems. Moreover, the device of this study uses hypoallergenic materials while the others do not, and this may be an important consideration when working with intraoral measurements [23], [48]–[50].

Finally, it should be highlighted that the main purpose of these devices is to quantify maximum tongue force of the different movements and tongue pressure. In addition, they may help in the diagnosis and strengthening of the tongue muscles. As a result, reliability studies like those developed by the IOPI, the KSW and the device proposed in this research are needed. Furthermore, validation of all devices is required in order to support their clinical use, as has been carried out in this study. These studies are needed in order to compare devices and select the best option.

A. Clinical Implications

From a clinical point of view, having validated instruments proves to be of great value. In this case, the instrument is a tongue force resistance device that could have an important positive impact on the treatment of patients with tongue disorders. Some of these primary tongue disorders could be orofacial cancer, swallowing disorders or disturbances associated with dysarthria. Other secondary tongue disorders may also benefit from this device, such as temporomandibular disorders. This is because some research has observed significant changes in tongue force in patients with this pathology [12], [13]. In this study, we assessed the protrusion movement of the tongue against anterior part of hard palate. That movement could be done by the action of the genioglossus, a midline muscle that can influence tongue shape and protract the tongue body [51]. This muscle has been identified as one of the core muscles of the tongue [52] and it is related with both functions: speech and swallowing [52], [53]. That is why, we consider that this research has important clinical implications because this new device can be used to assess objectively protraction movement of the tongue, but also it could be used as a training tool for improving speech and swallowing.

On the other hand, it should be noted that the 1 maximum repetition (1MR) system, a recognized system for assessing muscle strength, cannot be applied to an organ like the tongue. Moreover, this apparatus allows us to calculate the desired percentage through the force peak, based on the value established by it [36]. Although the subjects of this study did not see the screen while they were doing the force, it would also enable patients and professionals to develop force and resistance training with the visual biofeedback the system includes.

Moreover, future research lines should assess the utility and effectiveness of this device in tele rehabilitation processes. These are currently booming, and the device presented in this study could increase the quality of patient´s assessment and monitoring in a home treatment.

Finally, an important point to be considered is the need to make technology available to as wide a population as possible. Having access to the device will allow professionals to get objective measurements in the clinic for giving specific exercise doses. This also, will promote

patient independence at home controlling the training and observing the progression. For this reason, a key component of the healthcare system is to develop reliable and useful tools with lower prices like this device. Likewise, the portability and intuitive interface of this prototype system makes it a tool which could reduce social and health care costs and waiting lists of the health system.

Future studies need to be developed to know the inter-rater reliability and determine normality values in asymptomatic subjects and temporomandibular or speak disorder patients. The authors of this work also recommend planning quality clinical trials to verify the effects of different types of tongue exercises in patients with tongue disturbance pathologies.

B. Limitations

The present study has some limitations. First, convenience sampling was used due to the impossibility of accessing big databases for a simple random sampling.

Second, all the participants were asymptomatic, so statistical results have to be carefully considered when used in populations with pathology.

Third, this study has not considered the sensor deformation inside the mouth of the patients while tongue pressure is taking place. We recognize the need to work on future improvement of the device, adding a further element that eliminates the deformation of the sensor during tongue pressure. Moreover, this improvement needs to consider adding a sensitive feedback, as a grooved part, in order to facilitate a specific tongue pressure point for the patient.

Fourth, another current limitation is that the sensor can be damaged due to intra-oral misuse. Although sensor damage is complicated due to it resists millions of measurements without altering its operation, it could be damage if patients bite the sensor. That is why, work is underway to isolate and protect this sensor in order to solve this type of problem in the future. Nevertheless, the device has been designed in such a way that the replacement of these sensors can be done easily. In addition, they are very cheap sensors (4-7€).

V. Conclusion

The data obtained in this study suggest that this new device can be used by clinicians due to the good validity values and high reliability shown (ICC =0.93) for intra-rater evaluation of the maximum tongue force. It allows to obtain objective measurements of tongue force in clinical practice in order to help clinicians with the diagnosis process and treatment progression. This device also has a low economic cost, is easy to handle and has various utilities, making this tool a great alternative in the measurement of tongue force-endurance due to the few ones that are available in the market have these limitations.

Acknowledgements

The authors would like to thank the Centro Superior of Estudios Universitarios La Salle for funding the English language edition of this manuscript. Moreover, we thank the collaboration of Pablo Jordi Perea, the speech language therapist in the Functional Rehabilitation Institute of the Centro Superior of Estudios Universitarios La Salle. HBA would like to thank the European Regional Development Fund (2020/5154).

Funding

Centro Superior de Estudios Universitarios La Salle funded the English-language edition of this manuscript.

Beltran-Alacreu H is financed by the European Regional Development Fund (2020/5154).

Conflict of Interest

No conflict of interest was declared by the authors of this study.

Ethics

The experiments carried out in this study complied with ethics Committee of Centro Superior de Estudios Universitarios La Salle (CSEULS). Registration number: CSEULS-PI-036/2019.

References

[1] W. Wang, C. Di, S. Mona, L. Wang, and M. Hans, "Tongue Function: An Underrecognized Component in the Treatment of Obstructive Sleep Apnea with Mandibular Repositioning Appliance," *Canadian Respiratory Journal*, vol. 2018, pp. 1–7, Nov. 2018, doi: 10.1155/2018/2157974.

[2] G. Messina, "The Tongue, Mandible, Hyoid System," *European journal of translational myology*, vol. 27, no. 1, p. 6363, Feb. 2017, doi: https://dx.doi.org/10.4081/ejtm.2017.6363.

[3] R. White, S. M. Cotton, J. Hind, J. Robbins, and A. Perry, "A Comparison of the Reliability and Stability of Oro-lingual Swallowing Pressures in Patients with Head and Neck Cancer and Healthy Adults," *Dysphagia*, vol. 24, no. 2, pp. 137–144, Jun. 2009, doi: 10.1007/s00455-008-9181-0.

[4] H. Yamanashi *et al.*, "Validity of maximum isometric tongue pressure as a screening test for physical frailty: Cross-sectional study of Japanese community-dwelling older adults," *Geriatrics & Gerontology International*, vol. 18, no. 2, pp. 240–249, Feb. 2018, doi: 10.1111/ggi.13166.

[5] M. A. Nicosia *et al.*, "Age Effects on the Temporal Evolution of Isometric and Swallowing Pressure," *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, vol. 55, no. 11, pp. M634–M640, Nov. 2000, doi: 10.1093/gerona/55.11.M634.

[6] H. Taniguchi *et al.*, "Fluoroscopic Evaluation of Tongue and Jaw Movements During Mastication in Healthy Humans," *Dysphagia*, vol. 28, no. 3, pp. 419–427, Sep. 2013, doi: 10.1007/s00455-013-9453-1.

[7] J. Kieser, M. Farland, H. Jack, M. Farella, Y. Wang, and O. Rohrle, "The role of oral soft tissues in swallowing function: what can tongue pressure tell us?," *Australian Dental Journal*, vol. 59, pp. 155–161, Jun. 2014, doi: 10.1111/adj.12103.

[8] M. Peladeau-Pigeon and C. M. Steele, "Age-Related Variability in Tongue Pressure Patterns for Maximum Isometric and Saliva Swallowing Tasks," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 11, pp. 3177–3184, Nov. 2017, doi: 10.1044/2017_JSLHR-S-16-0356.

[9] J. Furuya, S. Nakamura, T. Ono, and T. Suzuki, "Tongue pressure production while swallowing water and pudding and during dry swallow using a sensor sheet system," *Journal of Oral Rehabilitation*, vol. 39, no. 9, pp. 684–691, Sep. 2012, doi: 10.1111/j.1365-2842.2012.02319.x.

[10] P. Bourdiol, A. Mishellany-Dutour, M.-A. Peyron, and A. Woda, "Tongue-mandible coupling movements during saliva swallowing," *Journal of Oral Rehabilitation*, vol. 41, no. 3, pp. 199–205, Mar. 2014, doi: 10.1111/joor.12135.

[11] K. Hori *et al.*, "Role of tongue pressure production in oropharyngeal swallow biomechanics," *Physiological Reports*, vol. 1, no. 6, p. e00167, Nov. 2013, doi: 10.1002/phy2.167.

[12] R. R. Rosa, M. da R. S. Bueno, R. R. Migliorucci, A. G. Brasolotto, K. F. Genaro, and G. Berretin-Felix, "Tongue function and swallowing in individuals with temporomandibular disorders," *Journal of Applied Oral Science*, vol. 28, 2020, doi: 10.1590/1678-7757-2019-0355.

[13] M. Corsalini, D. Di Venere, F. Pettini, D. Lauritano, and M. Petruzzi, "Temporomandibular Disorders in Burning Mouth Syndrome Patients: An Observational Study," *International Journal of Medical Sciences*, vol. 10, no. 12, pp. 1784–1789, 2013, doi: 10.7150/ijms.6327.

[14] S. N. Robinovitch, C. Hershler, and D. P. Romilly, "A tongue force measurement system for the assessment of oral-phase swallowing disorders," *Archives of physical medicine and rehabilitation*, vol. 72, no. 1, pp. 38–42, Jan. 1991.

[15] K. Frändin *et al.*, "Long-Term Effects of Individually Tailored Physical Training and Activity on Physical Function, Well-Being and Cognition in Scandinavian Nursing Home Residents: A Randomized Controlled Trial," *Gerontology*, vol. 62, no. 6, pp. 571–580, 2016, doi: 10.1159/000443611.

[16] V. Adams, B. Mathisen, S. Baines, C. Lazarus, and R. Callister, "Reliability of Measurements of Tongue and Hand Strength and Endurance Using the Iowa Oral Performance Instrument with Healthy Adults," *Dysphagia*, vol. 29, no. 1, pp. 83–95, Feb. 2014, doi: 10.1007/s00455-013-9486-5.

[17] V. Adams, B. Mathisen, S. Baines, C. Lazarus, and R. Callister, "A Systematic Review and Meta-analysis of Measurements of Tongue and Hand Strength and Endurance Using the Iowa Oral Performance Instrument (IOPI)," *Dysphagia*, vol. 28, no. 3, pp. 350–369, Sep. 2013, doi: 10.1007/s00455-013-9451-3.

[18] M. Yoshikawa, M. Yoshida, K. Tsuga, Y. Akagawa, and M. E. Groher, "Comparison of three types of tongue pressure measurement devices," *Dysphagia*, vol. 26, no. 3, pp. 232–7, Sep. 2011, doi: 10.1007/s00455-010-9291-3.

[19] "IOPI Medical LLC. Iowa Oral Performance Instrument: users manual," 2008.

[20] State-of-the-Art Pointing Solutions for the OEM, "Force Sensing Resistor Integration Guide and Evaluation Parts Catalog," *Interlinks Electronics*, 2002.

[21] "FSR™ Integration Guide Interlink Electronics FSR ™ Force Sensing Resistors ™ FSR® Integration Guide."

[22] S. Parmar, I. Khodasevych, and O. Troynikov, "Evaluation of Flexible Force Sensors for Pressure Monitoring in Treatment of Chronic Venous Disorders," *Sensors*, vol. 17, no. 8, p. 1923, Aug. 2017, doi: 10.3390/s17081923.

[23] Y. Utanohara, R. Hayashi, M. Yoshikawa, M. Yoshida, K. Tsuga, and Y. Akagawa, "Standard Values of Maximum Tongue Pressure Taken Using Newly Developed Disposable Tongue Pressure Measurement Device," *Dysphagia*, vol. 23, no. 3, pp. 286–290, Sep. 2008, doi: 10.1007/s00455-007-9142-z.

[24] J. Kottner *et al.*, "Guidelines for Reporting Reliability and Agreement Studies (GRRAS) Were Proposed," *Journal of clinical epidemiology*, vol. 64, no. 1, 2011, doi: 10.1016/J.JCLINEPI.2010.03.002.

[25] T. Ahmed, N. A. Rahman, and M. K. Alam, "Validation and reliability of a prototype orthodontic bracket debonding device equipped with force-sensitive resistor (FSR): a novel method of measuring orthodontic bracket debonding force in vivo," *Prog Orthod*, vol. 20, no. 1, p. 26, 2019, doi: 10.1186/s40510-019-0277-x.

[26] M. J. Hjermstad *et al.*, "Studies comparing Numerical Rating Scales, Verbal Rating Scales, and Visual Analogue Scales for assessment of pain intensity in adults: a systematic literature review," *Journal of pain and symptom management*, vol. 41, no. 6, pp. 1073–93, Jun. 2011, doi: 10.1016/j.jpainsymman.2010.08.016.

[27] S. D. Walter, M. Eliasziw, and A. Donner, "Sample size and optimal designs for reliability studies," *Statistics in Medicine*, vol. 17, no. 1, pp. 101–110, Jan. 1998, doi: 10.1002/(sici)1097-0258(19980115)17:1<101::aid-sim727>3.0.co;2-e.

[28] M. Adam Bujang and N. Baharum, "A simplified guide to determination of sample size requirements for estimating the value of intraclass correlation coefficient: a review," *Archives of Orofacial Sciences*, vol. 12, no. 1, pp. 1–11, 2017.

[29] D. E. Hinkle, W. Wiersma, and S. G. Jurs, *Applied statistics for the behavioral sciences*. Des Moines, Iowa : Houghton Mifflin, 1988.

[30] J. P. Weir, "Quantifying Test-Retest Reliability Using the Intraclass Correlation Coefficient and the SEM," *The Journal of Strength and Conditioning Research*, vol. 19, no. 1, p. 231, Feb. 2005, doi: 10.1519/15184.1.

[31] C. Bunce, "Correlation, Agreement, and Bland–Altman Analysis: Statistical Analysis of Method Comparison Studies," *American Journal of Ophthalmology*, vol. 148, no. 1, pp. 4–6, Jul. 2009, doi: 10.1016/j.ajo.2008.09.032.

[32] J. M. Bland and D. G. Altman, "Comparing methods of measurement: why plotting difference against standard method is misleading," *Lancet (London, England)*, vol. 346, no. 8982, pp. 1085–7, Oct. 1995, doi: 10.1016/s0140-6736(95)91748-9.

[33] S. M. Haley and M. A. Fragala-Pinkham, "Interpreting Change Scores of Tests and Measures Used in Physical Therapy," *Physical Therapy*, vol. 86, no. 5, pp. 735–743, May 2006, doi: 10.1093/ptj/86.5.735.

[34] K. W. Wyrwich, "Minimal Important Difference Thresholds and the Standard Error of Measurement: Is There a Connection?," *Journal of Biopharmaceutical Statistics*, vol. 14, no. 1, pp. 97–110, Dec. 2004, doi: 10.1081/BIP-120028508.

[35] E. Versi, "'Gold standard' is an appropriate term," *BMJ : British Medical Journal*, vol. 305, no. 6846, p. 187, 1992, doi: 10.1136/BMJ.305.6846.187-B.

[36] V. S. McKenna, B. Zhang, M. B. Haines, and L. N. Kelchner, "A Systematic Review of Isometric Lingual Strength-Training Programs in Adults With and Without Dysphagia," *American Journal of Speech-Language Pathology*, vol. 26, no. 2, pp. 524–539, May 2017, doi: 10.1044/2016_AJSLP-15-0051.

[37] N. P. Solomon, H. M. Clark, M. J. Makashay, and L. A. Newman, "Assessment of orofacial strength in patients with dysarthria," *Journal of Medical Speech-Language Pathology*, vol. 16, no. 4, pp. 251–258, 2008.

[38] J. Ulrich Sommer, R. Birk, K. Hörmann, and B. A. Stuck, "Evaluation of the maximum isometric tongue force of healthy volunteers," *European Archives of Oto-Rhino-Laryngology*, vol. 271, no. 11, pp. 3077–3084, Nov. 2014, doi: 10.1007/s00405-014-3103-6.

[39] R. K. Balasubramaniam, S. Babu, M. A. Anil, A. L. Varghese, Z. R. Hussain, and D. F. Dsouza, "Does Tongue - Hold Maneuver Affect Respiratory – Swallowing Coordination? Evidence from Healthy Adults," pp. 68–71, 2019, doi: 10.4103/jnsbm.JNSBM.

[40] C. Fernandes, P. Glantz, S. Svensson, and A. Bergmark, "A Novel Sensor for Bite Force Determinations," *Dental materials : official publication of the Academy of Dental Materials*, vol. 19, no. 2, 2003, doi: 10.1016/S0109-5641(02)00020-9.

[41] J. S. Schofield, K. R. Evans, J. S. Hebert, P. D. Marasco, and J. P. Carey, "The effect of biomechanical variables on force sensitive resistor error: Implications for calibration and improved accuracy," *Journal of Biomechanics*, vol. 49, no. 5, pp. 786–792, Mar. 2016, doi: 10.1016/j.jbiomech.2016.01.022.

[42] L. Paredes-Madrid, A. Matute, A. F. Cruz-Pacheco, C. A. Parra-Vargas, & Elkin, and I. Gutiérrez-Velásquez, "Experimental characterization, modeling and compensation of hysteresis in force sensing resistors," *Revista DYNA*, vol. 85, no. 205, pp. 191–198, 2018, doi: 10.15446/dyna.v85n205.66432.

[43] R. Tarricone, A. Torbica, and M. Drummond, "Challenges in the Assessment of Medical Devices: The MedtecHTA Project," *Health Economics*, vol. 26, pp. 5–12, Feb. 2017, doi: 10.1002/hec.3469.

[44] Y. Utanohara, R. Hayashi, M. Yoshikawa, M. Yoshida, K. Tsuga, and Y. Akagawa, "Standard values of maximum tongue pressure taken using newly developed disposable tongue pressure measurement device," *Dysphagia*, vol. 23, no. 3, pp. 286–90, Sep. 2008, doi: 10.1007/s00455-007-9142-z.

[45] H. C. Crow and J. A. Ship, "Tongue strength and endurance in different aged individuals," *The journals of gerontology. Series A, Biological sciences and medical sciences*, vol. 51, no. 5, pp. M247-50, Sep. 1996, doi: 10.1093/gerona/51a.5.m247.

[46] D. A. Robin, A. Goel, L. B. Somodi, and E. S. Luschei, "Tongue strength and endurance: Relation to highly skilled movements," *Journal of Speech and Hearing Research*, vol. 35, no. 6, pp. 1239–1245, 1992, doi: 10.1044/jshr.3506.1239.

[47] R. Hayashi, K. Tsuga, R. Hosokawa, M. Yoshida, Y. Sato, and Y. Akagawa, "A Novel Handy Probe for Tongue Pressure Measurement," *Int J Prosthodont*, vol. 15, no. 4, pp. 385–8, 2002.

[48] A. Hewitt *et al.*, "Standardized Instrument for Lingual Pressure Measurement," *Dysphagia*, vol. 23, no. 1, pp. 16–25, Mar. 2008, doi: 10.1007/s00455-007-9089-0.

[49] K. Tsuga, M. Maruyama, M. Yoshikawa, M. Yoshida, and Y. Akagawa, "Manometric Evaluation of Oral Function With a Hand-Held Balloon Probe," *Journal of oral rehabilitation*, vol. 38, no. 9, 2011, doi: 10.1111/J.1365-2842.2011.02202.X.

[50] S. N. Robinovitch, C. Hershler, and D. Romilly, "A Tongue Force Measurement System for the Assessment of Oral-Phase Swallowing Disorders," *Arch Phys Med Rehabil*, vol. 72, no. 1, pp. 38–42, 1991.

[51] M. Stone and A. Lundberg, "Three-dimensional tongue surface shapes of English consonants and vowels," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3728–37, Jun. 1996, doi: 10.1121/1.414969.

[52] H. Takemoto, "Morphological Analyses of the Human Tongue Musculature for Three-Dimensional Modeling," *Journal of Speech,*

Language, and Hearing Research, vol. 44, no. 1, pp. 95–107, Feb. 2001, doi: 10.1044/1092-4388(2001/009).

[53] K. M. Hiiemae and J. B. Palmer, "Tongue movements in feeding and speech," *Critical reviews in oral biology and medicine: an official publication of the American Association of Oral Biologists*, vol. 14, no. 6, pp. 413–29, 2003, doi: 10.1177/154411130301400604.

Marta Carlota Díaz Sáez

Marta Carlota Díaz Sáez got her Physical Therapy Degree at the Centro Superior de Estudios Universitarios La Salle (CSEULS), Universidad Autónoma de Madrid (UAM) in 2016. She is a current PhD student by the same university. She also developed a postgraduate study in Physical Therapy for Myofascial Pain, Dry Needling at the CSEULS (UAM) in 2016. She studied an Official University Master of Advanced Physical Therapy in the Treatment of Pain at the CSEULS (UAM) in 2016-2017. She got her expert diploma of Physical Therapy for Temporomandibular Disorders and Headaches at the CSEULS (UAM) in 2020. She is Orthopaedic Physical Therapist in the Instituto de Rehabilitación Funcional de La Salle (IRF La Salle). She is a full professor in Dry Needling postgraduate course at the CSEULS (UAM). She also has been a full professor in Physical Therapy degree in the Bases of Physical Therapy and Electrotherapy subject and in the International Academic Activities (ACAI) program for Erasmus students at the CSEULS (UAM). She is author of 2 scientific articles and her main research area is Craniofacial Pain. She belongs to CranioSpain Research Group at the CSEULS (UAM).

Héctor Beltrán Alacreu

Héctor Beltrán Alacreu got his Physical Therapy Degree at the Universidad Europea de Madrid in 2010. He also studied an Official Master in Orthopedic Manual Therapy at the Universidad Europea de Madrid in 2011 and an Official Master in Study and Treatment of Pain at the Universidad Rey Juan Carlos in 2016. He is a PhD at the Universidad Rey Juan Carlos in 2015. He has 10 years of experience as an assistant physiotherapy. He is a professor and researcher at the Universidad de Castilla-La Mancha in the Faculty of Physiotherapy and Nursing. Finally, he has published more than 30 original scientific articles His main research areas are in Physical Rehabilitation and Musculoskeletal Pain.

Javier Gil Castillo

Javier Gil Castillo got his BCs degree in Biomedical Engineering in Technical University of Madrid (UPM) in 2016. His Bachelor Thesis at the Biomedical Technology Center (CTB) was focused on the techniques of magnetic hyperthermia and magnetometry. He completed his internship in the innovation unit of the San Carlos Clinical Hospital in Madrid. He also got his MSc degree in Biomedical Engineering in University of Navarra, TECNUN, in 2018. His Master's thesis was developed in the Neural Rehabilitation Group (Cajal Institute, CSIC) and focused on the design, the integration and the implementation of a functional electrical stimulation (FES) assistance system during the gait for patients with incomplete spinal cord injury. He is currently working as a researcher in the Neural Rehabilitation Group since 2018. During the period at CSIC he has published an article in the Journal of NeuroEngineering and Rehabilitation (JNER) and two publications in congresses (IROS and ICNR) and has participated in projects at national and international level as REASISTE and TAILOR. He is currently working on his doctoral thesis related to gait rehabilitation using electrostimulation and wearable robots.

Gonzalo Navarro Fernández

Gonzalo Navarro Fernández got his Physical Therapy Degree at the Centro Superior de Estudios Universitarios La Salle (CSEULS), Universidad Autónoma de Madrid (UAM) in 2015. He also developed different postgraduate studies in Health Sciences Applied Statistics, in Advanced Methods in Applied Statistics and in Systematic Reviews and Metanalysis at the Universidad Nacional de Estudios a Distancia (UNED). He studied and Official University Master of Advanced Physical Therapy in the treatment of Pain at the CSEULS (UAM) in 2015-2016. He obtained his expert diploma of Physical Therapy in the Management of temporomandibular disorders and headaches at the CSEULS (UAM) in 2020. He is currently working in his PhD at the Universidad Rey Juan Carlos (URJC). He works as orthopaedic physical therapist in the Instituto de Rehabilitación Funcional y Ciencias Aplicadas al Deporte La Salle (IRF La Salle). He also works as full professor in Physical Therapy Degree at CSEULS in different subjects such as Orthopaedic Manual Therapy I, Biomechanics, Physical Therapy assessment and evaluation; and in 2 Physical Therapy postdegrees: Professional Development Course in Invasive Physical Therapy and Expert Course in Physical Therapy in the Management of Temporomandibular Disorders and Headaches. Finally, he belongs to CranioSPain Research Group at the CSEULS (UAM) and is author of 9 scientific articles in his main research area (Cranio-cervico-facial pain) and other physiotherapy areas.

José Luis Cebrián Carretero

José Luis Cebrián Carretero got his Medical Degree at the Universidad Autónoma de Madrid (UAM) in 1993. He is PhD by the same university in 2005. He also studied Oral and Maxillofacial training H.U. La Paz from 1995 to 2000. He got is Dental Degree in the Universidad Europea de Madrid (UEM) in 2009. He is fellow of the European Board of Oral and Maxillofacial Surgery, Specialist in Oral and Maxillofacial Surgery, La Paz University Hospital, Madrid and Chief of Division in Oral and Maxillofacial Surgery, La Paz University Hospital, Madrid, Hospital Sur de Madrid and Clínica La Luz. Finally, he is Vice-President of the SECOM (Spanish Society of Oral and Maxillofacial Surgery). He is author of 2 books, 15 chapters and more than 50 scientific articles. His main research areas are Implantology, Head and Neck Oncologic Surgery and Orthognatic and Oral Surgery.

Alfonso Gil Martínez

Alfonso Gil Martínez got his Physical Therapy Degree at the Hospital Militar Gómez Ulla in 2003. He also studied an Official Master in Neuromotor Control at the Universidad Rey Juan Carlos in 2011-2012. He is a PhD by the same university in 2017. He is an assistant physiotherapist with 20 years of experience. He is a professor and researcher at the Hospital Universitario La Paz and Centro Superior de Estudios Universitarios La Salle. Finally, he has published more than 40 original scientific articles in the Pain area, several book chapters and has raised more than 1 million euros in competitive funding for healthcare research and innovation projects.

Deep Learning Assisted Medical Insurance Data Analytics With Multimedia System

Cheng Zhang[1]*, B. Vinodhini[2], Bala Anand Muthu[3]

[1] School of Law, Jiangnan University, Wuxi 214122 (China)
[2] Assistant Professor, SNS College of Technology, Coimbatore - 641035 (India)
[3] Associate Professor, Department of Computer Science, Adhiyamaan College of Engineering (India)

Received 14 July 2020 | Accepted 18 April 2021 | Published 27 January 2023

Abstract

Big Data presents considerable challenges to deep learning for transforming complex, high-dimensional, and heterogeneous biomedical data into health care data. Various kinds of data are analyzed in recent biomedical research that includes e-health records, medical imaging, text, and IoT sensor data, which are complex, badly labeled, heterogeneous, and usually unstructured. Conventional statistical learning and data mining methods usually require first to extract features to acquire more robust and effective variables from those data. These features help build clustering or prediction models. New useful paradigms are provided by the latest advancements based on deep learning technologies for obtaining end-to-end learning techniques from complex data. The abstractions of data are represented using the multiple layers of deep learning for building computational models. Clinician performance is augmented by the prospective of deep learning models in medical imaging interpretation, and automated segmentation is used to reduce the time for the diagnosis. This work presents a convolution neural network-based deep learning infrastructure that performs medical imaging data analysis in various pipeline stages, including data-loading, data-augmentation, network architectures, loss functions, and evaluation metrics. Our proposed deep learning approach supports both 2D as well as 3D medical image analysis. We evaluate the proposed system's performance using metrics like sensitivity, specificity, accuracy, and precision over the clinical data with and without augmentation.

Keywords

Convolution Neural Network, Deep Learning, Image Regression, Medical Image Analysis, Segmentation.

DOI: 10.9781/ijimai.2023.01.009

I. Introduction

According to recent reports, the present society produces data more quickly than in any other decade, making numerous doors open for various prediction strategies and making it difficult for analysts [1]. Multiple industries become progressively dependent on excellent data quality, and the interest in the sound factual examination of these data is rising in like manner. In the insurance sector, the provision of information has always been believed to be relevant. The insurance provider shall be entitled to the cases arising from this Arrangement in the form of an agreement to provide a client and shall retain reserves to cover any future obligations. For all possibilities taken into account, the insurance premium has to be paid before the real costs are identified. It is referred to as the reversal of the creation cycle. It infers that the exercises of reserving and pricing are firmly interconnected in actuarial practice. From one perspective, statisticians need to decide a reasonable cost for the insurance items they need to pay. Setting the excellent levels charged to the insured's is done in a data-driven way where prediction models are fundamental and essential.

Hazard based estimating is essential in a serious and well-working insurance market. An insurance organization must then protect its solvency and hold money to satisfy outstanding liabilities [2]. Therefore, holding statisticians should foresee, with most extreme exactness, the aggregate sum expected to pay guarantees that the insurance provider has legitimately conceded to cover for. These reserves structure the primary thing on the risk side of the insurance organization's financial record and accordingly have a significant economic effect. This exploration aspires to advance new, precise, and accurate models for the actuarial work field. Non-life (for example, engine, fire, obligation), life, and well-being backup plans are continually faced with the difficulties of quickly expanding facilities for information assortment and data collection, stockpiling, and investigation. Anyway, utilizing their best-in-class approaches for the insurance business won't have the option to plan a satisfactory reaction to these difficulties and associations with the controls of measurements and considerable information investigation.

Besides, the expanded spotlight on inward hazard and the changing administrative rules encourages the importance of improved apparatuses for actuarial modeling of the prediction models. Specifically, the European Solvency II Directive1 forces new necessities to upgrade policyholder security. With these new administrative rules' ongoing presence, the estimation of future incomes and their vulnerability turns out to be progressively significant [3]. Simultaneously, actuarial prediction models need to consent to existing and pending guidelines. Throughout strategies to survive with complex problems, conventional

* Corresponding author.
E-mail address: sunday@jiangnan.edu.cn

machine learning approaches are not adequate. High performance deep learning computing offers the ability to manage massive medical image data for precise and efficient diagnosis. Deep learning helps to pick and extract characteristics and create new ones. It diagnoses the illness and the predictive objective and offers actionable prediction models to effectively support doctors. The Gender Directive2 has denied the utilization of sex as a hazard factor in insurance estimating, and antidiscrimination laws may advance sooner rather than later, further constraining the legally binding opportunity of insurance organizations.

Demonstrating guarantee losses – alternatively called guarantee sizes or severities – is urgent when evaluating insurance items, deciding capital necessities, or overseeing dangers inside money related establishments. For example, different essential circulations, the gamma or lognormal, have been utilized to demonstrate nonnegative losses [4]. These parametric conveyances are not frequently suitable for actuarial information, which might be multimodal or substantial. Besides, while building aggregate hazard models or joining actuarial dangers from numerous business lines, these serious appropriations don't prompt a systematic structure for the relating total loss circulation. While numerical or recreation calculations are accessible, it is considered advantageous to use routine procedures whenever the situation allows. There is continuously a tradeoff between scientific effortlessness from one perspective and practical demonstrating modeling on the other. Fig. 1. shows the essential features involved in modeling insurance data. It offers the basic features involved in insurance data collection, analysis, and modeling the same.

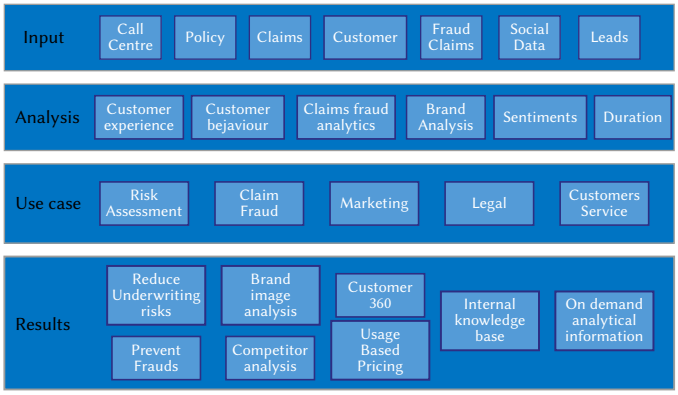


Fig. 1. Insurance Data Modeling.

These days, insurance organizations monitor all kinds of data for every individual case. Rich information sources record, for instance, the event occurrence date, the announcement of a reporting delay, the date and measure of every payment loss, and the settlement date. Many recent forensics advancements [5] monitor the insurance company databases for data theft and attacks over insurance data stored in the cloud [6]. The current strategies for claims holding are intended for totaled information, yet numerous valuable data is lost through this information compression. For considering all the useful information for analytics and establish new relations, deep learning models are considered. In any case, data accessible to play out the prediction work originates from two sources; a group of client characteristics (e.g., client socioeconomics and insurance enlistment data) and the historical backdrop of client emergency clinic affirmation claims. Big data can potentially expose connections, secret trends and other observations by analyzing massive data sets. And data from the learning of machine healthcare led to the discovery of human genomes or medicines for life threatening diseases such as cancer.

The last can be considered as a period of inconsistent occasions. Besides, every hospitalization occasion would itself be communicated as a succession of emergency clinic administrations used by the patient during that specific stay. While it is clear to straightforwardly introduce the principal kind of data as features to a characterization calculation, data mining of time arrangement and point forms, for example, the hospitalization occasions depicted above, is yet an ongoing research territory [7] [8]. It is somewhat testing to remove irrelevant data in a valuable way. The general planning of recorded medical clinic visits for prediction appears to contain unrelated yet significant data that may prompt progressively exact forecasts if it is very well used for insurance data creation.

II. Related Works

Shi et al. [9] designed HFDA, an ensemble Artificial Intelligence way to deal effectively and distinguish the clinical insurance claims implemented in an online clinical insurance claim framework in China. Tooth et al. [10] proposed another client profitability technique for the insurance business by including risk reserve. The proposed scheme can gauge the genuine insurance client commitment viably by thinking about the recorded buying conduct and foreign income. Wang et al. [11] proposed a novel deep learning model for accident coverage extortion discovery that utilizes Latent Dirichlet Allocation (LDA)- based content examination. In the proposed strategy, LDA is first used to detect the content features covering up in the accident cases' content depictions. Deep neural systems are prepared on the information, incorporating the content features and conventional numeric features for distinguishing fake cases.

Koutsomitropoulos et al. [12] developed OWL metaphysics to recognize insurance forms and delineate information volumes gathered in customary information stores. Under thinking, many semantic questions were shown utilizing the vocabularies in the ontology that can simplify examination and derive understood realities from this information. Lin et al. [13] developed a heuristic bootstrap testing approach joined with the hybrid learning calculation for the insurance business data mining. A parallel analysis that utilizes the equal registering ability and memory cache system improved by Spark and used F-Measure and G-intend to assess the calculation's accuracy. The insurance business information from China Life Insurance Company is used to investigate the proposed model efficiency.

Ren et al. [14] proposed a survival forecast model dependent on graph pattern mining. In the first place, every patient's medical coverage information is developed as a Heterogeneous Information Network (HIN). At that point, visit designs are mined from these HINs, and each successive example is viewed as an element called "design highlight." Finally, the survival time is given by an improved random forest, which can consider the edited information from the graph data mining. The investigation is carried out on a genuine medical coverage informational collection to investigate the utilization of factual strategies to make a standard based heuristic motor that works with self-learning Decision Trees. Rayan et al. [15] present an ensemble system that consolidates expertise in the domain and unsupervised learning procedures to recognize false cases. The examination group is implied with a weighted module of extraordinary circumstances posting the most probable fraud cases with proactive and review investigation comments.

Chae et al. [16] inspected the information disclosure attributes and data mining calculations to investigate how they can produce accurate results and give regular data to hypertension prediction using the Korea Medical Insurance Corporation database. In particular, this examination approved the intensity and core ability of data mining calculations by contrasting the presentation of the decision tree, logistic regression, CHIAD (Chi-squared Automatic Interaction Detection), and C5.0 (a variation of C4.5) utilizing the test set of 4588

recipients and the evaluation set of 13,689 recipients. Viveros et al. [17] addressed the viability of two data mining methods in breaking down and recovering unknown personal conduct standards from gigabytes of information gathered in the medical coverage industry. A scene (claims) database for pathology administrations and a general professionals database were utilized for the analysis. Affiliation rules were applied to the scene database; neural segmentation was used to overlay the two databases. The outcomes acquired from this investigation show the potential estimation of data mining in medical coverage data frameworks by distinguishing designs in the pathology administrations and arranging the general professionals into groups mirroring their practices' nature and style. The methodology produced a higher percentage of results that couldn't have been acquired utilizing traditional procedures.

Jiang et al. [18] presented four significant difficulties existing in artificial intelligence in genuine business models. Standard artificial intelligence calculations can commonly be applied to traditional informational indexes, which are ordinarily homogeneous and adjusted. A proficient cost-sensitive parallel learning framework (CPLF) was used to improve insurance tasks with an in-depth learning approach that doesn't require preprocessing. The methodology contains a novel, unified; start to finish a cost-effective neural system that learns genuine heterogeneous information. An explicitly structured cost-delicate grid that consequently produces a powerful model for understanding minority arrangements and the parameters of both the cost-effective lattice and the half breed neural system is, on the other hand, yet mutually upgraded during processing. CPLF-based design for a certifiable insurance knowledge activity framework showed a misrepresentation discovery and arrangement restoration during this framework's investigation. Wang et al. [19] applied a linkage of the Knowledge Discovery in Databases (KDD) procedure to examine the call community information of the NHIA. The practical techniques of handling, determination, data mining, and assessment for two kinds of data mining investigations: information arranging and information affiliation, Moreover, the investigation results and counsel experts in NHIA for the expert assessment about those outcomes and existing medical coverage arrangements showed the establishment of using intelligence paradigms in the medical insurance field. Senthil Murugan and Usha Devi [20] – [21] have proposed a hybrid classification technique for analyzing many data. Additionally, decipher, and reach determinations from these outcomes using information representation are presented for supportability.

Sato et al. [22] built a novel examination system dependent on past investigations that led with clinical scientists and structured a UI that encourages the development of the preparation rationale in a basic and straightforward way. By indicating the execution consequence of commonplace investigations of insurance information, the created system's viability is examined. Umemoto et al. [23] meant recognizing the changes in the prescription patterns and distinguishing its motivation through an investigation of Medical Insurance Claims (MICs), which involve the details of clinical expenses charged to medical safety insurance providers. The methodology is two-crease. Firstly, proposed an inactive variable model that recreates doctors' drug conduct to precisely imitate month to month medicine time arrangement from the MIC information, where medicine interfaces between the illnesses and meds are absent. Secondly, applied a state-space model with intercession factors to deteriorate the month-to-month remedy time arrangement into various parts, including regularity and auxiliary changes. Testing is fair to extract unnecessary information practically. If it is very well used for insurance data production, the general planning of reported medical clinic visits for prediction tends to include insignificant essential data that can increasingly prompt accurate forecasts.

III. Proposed Methodology

In the recent year, deep learning methods become well known for their high accuracy rate and immense domain applications in various research fields, including image processing [24], [25], speech recognition [26], [27], computer vision [28], [29], authentication system [30]. Convolution neural networks (CNNs) feed-forward artificial neural network (FANN) expectant by standard procedures proposed to classify unique patterns straight away from medical and non-medical image data. Motivated by the great success of CNN in medical research, we employed CNN to target breast cancer tumor segmentation and classification. In convolutional neural networks, convolutional layers are the key building blocks used. The fast application of a variable to an input that results in inactivation is a convolution. Convolution of deep learning infrastructure focused on the neural network that performs medical imaging data analysis at different pipeline levels, including data loading, data increase network architectures, loss functions, and evaluation metrics.

In this paper, we present a new deep learning-based CNN framework architecture to segment and classify breast tumors into two classes (B-Benign and M-Malignant) by the use of CNN fine-tuned models. The proposed system consists of various pipeline stages, including data loading; patch extraction, selection, image segmentation, data augmentation, deep feature extraction, deep feature selection, and classification. The proposed system's detailed flow is shown in Fig. 2., whereas each pipeline stage's detail is described in subsequent sections. Models of deep learning are developed using neural networks. A neural network takes in inputs, and then processed in hidden layers using weights that are changed during preparation. The model spits out a forecast then. For making better predictions, the consequences are adapted to identify trends. A deep learning model is designed to continuously analyze data similar to how a person might conclude with a logic structure. A layered system of algorithms called deep learning applications to use an artificial neural network. The architecture of an artificial neural network is inspired by the human brain's biological neural network, leading to a learning mechanism that is far more capable than that of traditional models of machine learning.

A. Data Loading

In this stage, medical image files are loaded from a medical file format data set. Medical images are stored in different file formats as compared to many other computer vision tasks. These file format stores metadata information like acquisition information (specifies scanner parameter, modality types, etc.), spatial information (including anatomical point of reference and voxel anisotropy), and patients' data. Deep learning is a type of machine learning in which a model learns directly from pictures, text, or sound to perform classification tasks. Typically, deep learning is applied using the design of neural networks. The term deep refers to the number of layers in the network; the more profound the network, the more layers. The processing of medical imaging refers to the handling of images using a computer. This processing requires many methods and practices, such as image collection, storage, presentation, and communication.

B. Medical Image Segmentation

1. Patch Extraction

From the image file, we perform patch extraction over each image Im of size $P \times Q$; this image is then divided into patches Pch_i with a size of 256 × 256 pixels with no overlapping. For each image, the number of patches was different as the size of each image is different.

2. Patch Selection

Each image patches Pch_i are partitioned into sub-region SR_0, SR_1, SR_2, SR_3, SR_4 such that $\bigcup_{i=0}^{3} SR_i = Pch_i$ where ∪ representing the union

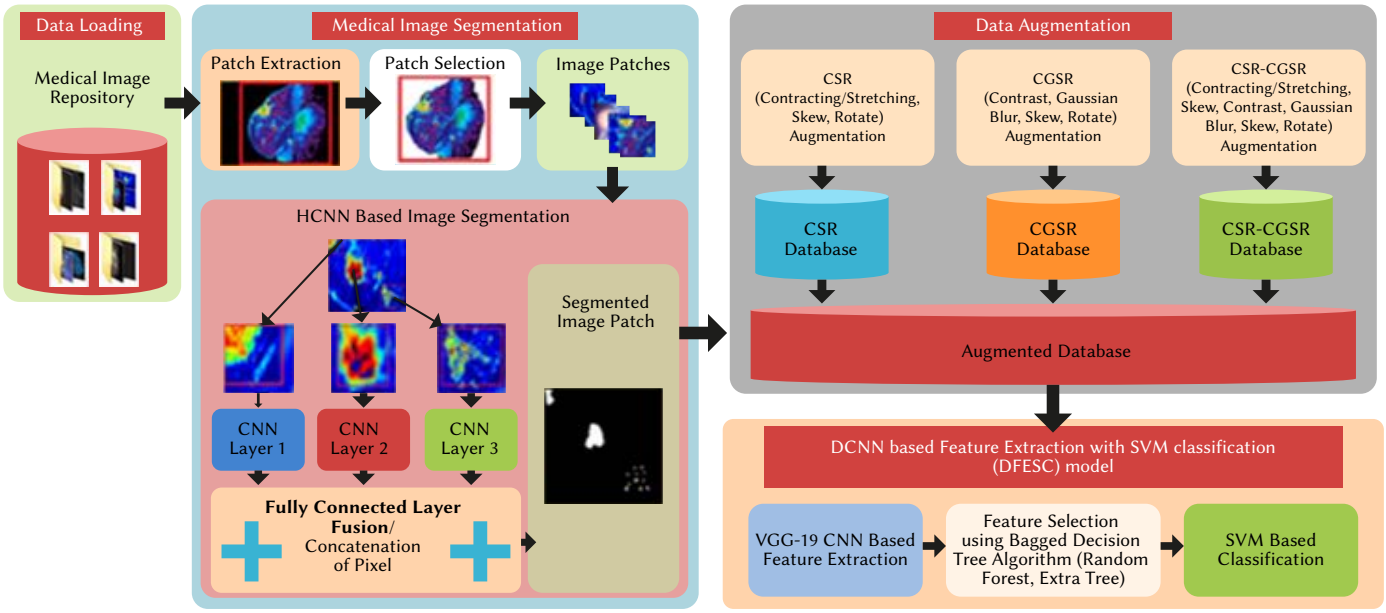


Fig. 2. Proposed Methodology for Tumor Segmentation.

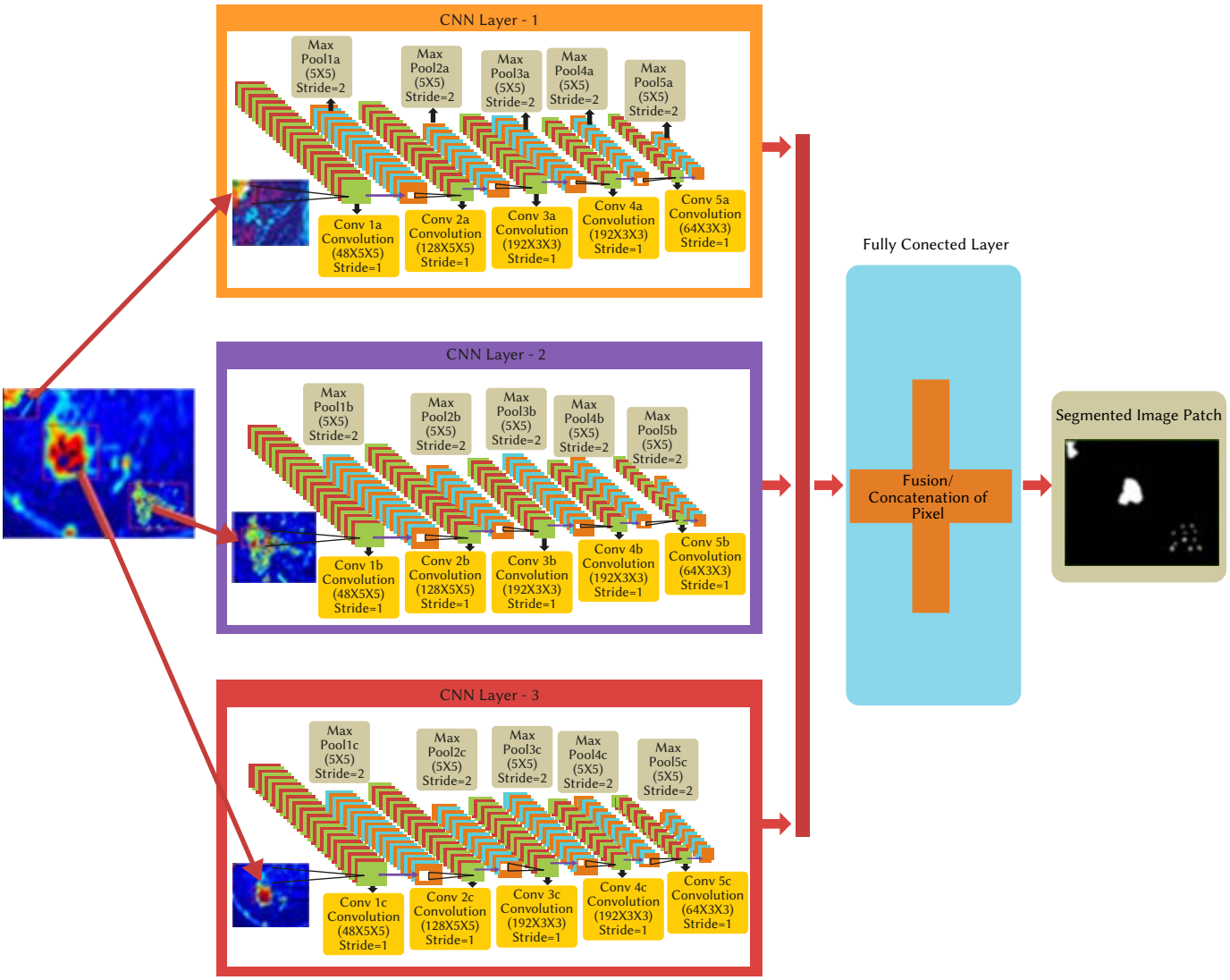


Fig. 3. HCNN architecture for Image Segmentation.

of all four regions. Only those regions are selected that contain tissue sections. Each patch has some noisy background data that is removed using the homogeneity factor, which is defined in (1):

$$H_f = \sum_{j=1}^{I} \quad \sum_{k=1}^{I} \quad \frac{Pr_{jk}}{1+|j-k|} \tag{1}$$

Where, Pr_{jk} is relative position probability of pixel pair (j, k), and I his distinct level of intensity was computed for every patch and optimized up to 60% threshold. Each patch belongs to the tissue region. Patching technology is an effective means of marking the structures of the brain and of other animals. In general, these approaches mark each of the voxels of a key growth by contrasting the image patch centered on the voxel with patches from the Atlas library. A search window is typically used locally based on the target voxel. Various patch-based mark fusion processes are proposed for effective and stable segmentation and illustrated. For works using non-rigid registration, comparable findings have been published. In a variety of computer vision tasks, including texture synthesis, painting and super-resolution, patch-based techniques recently showed high efficiency. Non-local denoising has led to the promotion of the field and to the development of many patch-based segmentation techniques for medical imaging applications.

The algorithm for patch extraction and patch selection is presented below:

Algorithm 1: Patch Extraction and Patch Selection

1. Extract the Patches Pch_i of size 224 × 224 pixel from $Im_{P\times Q}$
2. Partitioned patches into subregion SR_0, SR_1, SR_2, SR_3, SR_4.
3. Select Patches based on tissue section region such that for i=0 to 4:

   Begin:

   $\bigcup_{i=0}^{3} SR_i = Pch_i$

   End
6. Background noise removal using homogeneity factor for pixel pair (j, k).
7. For $j = 1$ *tol*:

   Begin: For $k = 1$ *tol*:

   Begin:

$$H_f = \sum_{j=1}^{I} \quad \sum_{k=1}^{I} \quad \frac{Pr_{jk}}{1+|j-k|}$$

   End

3. HCNN Based Image Segmentation

In the process of segmentation, background tissues are removed from the tumor region in the image. For segmentation, two methods are employed.

- Region-Based Approach in which segmentation is performed based on similarity detection. Few Region-Based approaches include Region growing, merging, and splitting using quad tree decomposition.
- Boundary Based Approach in which detection of discontinuity is performed and then linked to form boundaries of region.

A blend of distinct methods is implemented to maximize the segmentation outcomes. A region-based segmentation and picture analysis with application to medical images has been carried out in this paper. Clustering, object detection, and boundary detection are among the most critical measures in image segmentation. Segmentation of related structures is of utmost significance for several image processing and visualization activities both within and outside the medical image domain. As a result of non-optimal parameter settings, images segmented by area rising techniques often contain either too many regions or too few regions. A blend of distinct methods is implemented to maximize the segmentation outcomes.

The algorithm for boundary detection helps to find the right boundary for noisy pictures. The convergence between the original image and the corresponding mask provides insight into vector data. Finally, the algorithm for boundary detection is implemented to yield accurate input image boundaries.

In this paper, we perform Region of interest extraction using a deep learning-based fully automatic technique called Hierarchical CNN (HCNN). HCNN is different from traditional CNN because of its in-depth image processing. The architecture of HCNN consists of three hierarchical layers, which are fused at fully connected layers. Every pixel of the image is segmented, and the result of segmentation is then merged into the mask after input pixel segmentation. Fig. 3. represents the architecture of HCNN, including convolution layer, pooling layer, Rectified Linear Units Layer (RerLU), and fully connected layer.

Convolution Layer: It performs input image convolution using convolution kernels, which is represented in (2)

$$(R^k)_{x,y} = (Wt^k * Pch_i)_{x,y} + bs^k \tag{2}$$

Where * represent the convolution to the input image patches Pch_i, Wt^k and bs^k represent weight and bias between two neurons, whereas k represent convolution kernel index. $(R_k)_{xy}$ represent the convolution response between k^{th} kernel and pixel with center (x, y). To control the size of output volume of the convolution layer, we use three parameters: depth (number of convolution kernel), stride (control kernel shift amount), and padding (control the spatial size of convolution output volume).

Rectified Linear Units Layer (ReLU): It is used as an activation function that sets all the negative value to zero using the non-linear activation function shown in equation (3)

$$g(R) = max(0, R) \tag{3}$$

Where R represent convolution response output.

Pooling Layer: This layer performs non-linear transformation to reduce spatial dimension and noise elimination activated from the preceding layer. There are different down-sampling strategies used that perform pooling operations, including stochastic pooling, average pooling, and max pooling. Amongst all, Max pooling is most famous for its high-speed performance and convergence optimization.

Fully Connected Layer: It connects every neuron of the preceding layer to all neurons of this layer, called a fully connected layer. The (4) describes it:

$$(r^k)_{x,y} = \sum_f \quad (Wt^{kf} * Pch_i{}^f)_{x,y} + bs^k \tag{4}$$

Where, f represent f^{th} neuron index of input, r^k represent k^{th} neuron output, Wt^{kf} and bs^k represent weight and bias between two neurons Pch_i^f and r^k. In our proposed work, we concatenate the output of all the three convolutions at this layer. Health analytics analyses existing and past awareness of the market for predicting trends, increased scope and much better control the dissemination of diseases. It will include opportunities to enhance health safety, clinical data, diagnosis and organizational management. To overcome many technological challenges and issues that need to be solved to realize this opportunity, the motivational aspect of data analytics and mobile computing is critical for healthcare systems. New capabilities such as artificial learning, data analytics, and computational power have to be upgraded to provide more intelligent and skilled healthcare services for people in advanced healthcare systems.

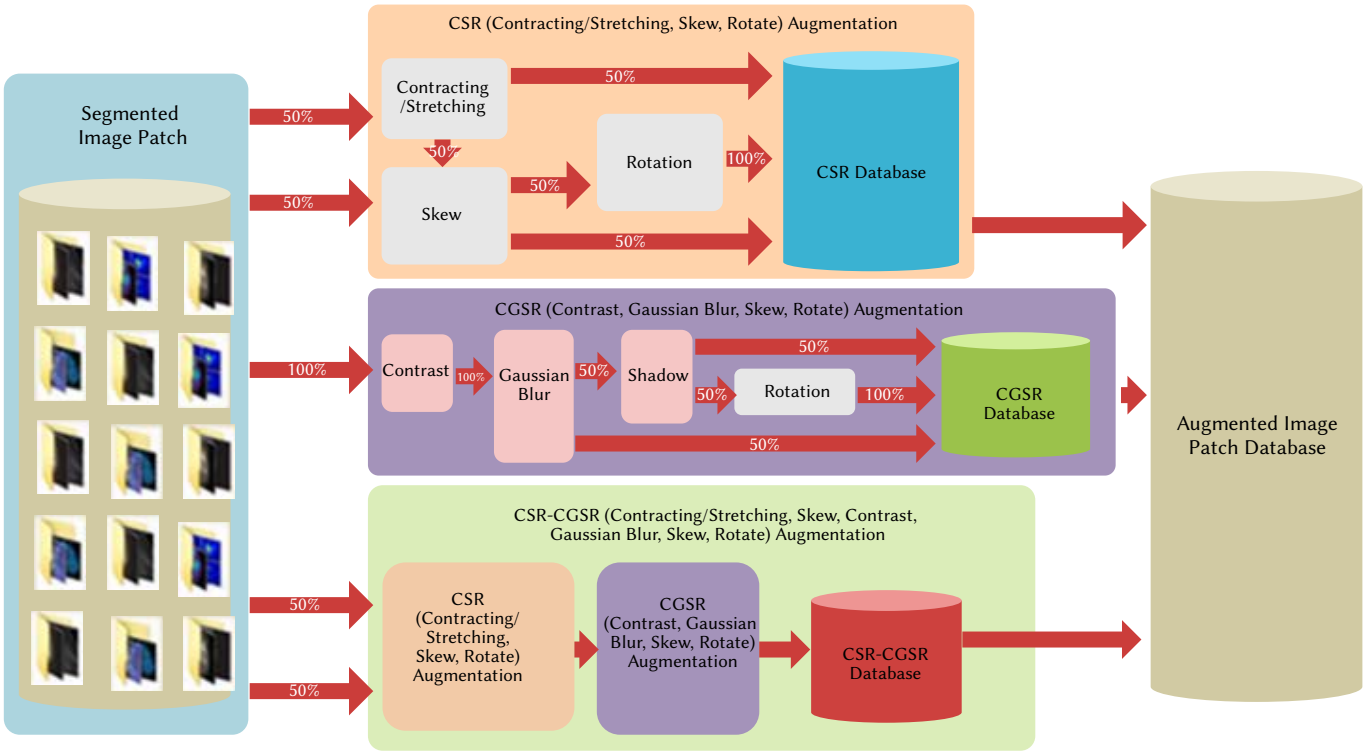


Fig. 4. Data Augmentation using CSR, CGSR, and CSR-CGSR Algorithms.

After performing these steps, a post-processing step is performed to eliminate the noise from the segmented image. The algorithm of the HCNN model for tumor segmentation is mention below:

Algorithm 2: Image Segmentation using HCNN model

1. Input: Image Patches Pch_i, Weight Wt^k and bias bs^k
2. Array Initialization: Number of filter ⟶ $NOF[5] = \{48, 128, 192, 192, 64\}$;
   Number of pixel in Imagepatches $Pch_i = 224 \times 224$;
2. Output: Segmented Image patches Pch_i'.
3. # Three Layer HCNN : for $i = 0$ to 2:
   # HCNNLayer 1: For $j = 0$ to 4
   Begin:
   Number of kernel: $NOK = NOF[j]$;
   if $(j < 2)$ then
   $Conv_{ij}$: $Conv(NOK, 5, 5, ImagePatch_{Size} = (Pch_i, Pch_i, 5))$
   $(R^k)_{x,y} = (Wt^k * Pch_i)_{x,y} + bs^k$
   else
   $Conv_{ij}$: $Conv(NOK, 3, 3, ImagePatch_{Size} = (Pch_i, Pch_i, 3))$
   $(R^k)_{x,y} = (Wt^k * Pch_i)_{x,y} + bs^k$
   $ReLu_{ij}$: $g(R) = max(0, R)$
   $MaxPool_{ij}$:
   $pool_{size} = (3, 3)$;
   $stride = (2, 2)$;
   End:
   End:
4. # *FullyconnectedLayer* 1: *Fusion of out put of all three HCNN layer*

$$(r^k)_{x,y} = \sum_f (Wt^{kf} * Pch_i{}^f)_{x,y} + bs^k$$

5. Output: segmented Image Pathes Pch_i'
6. Exit

C. Data Augmentation

This technique is used to increase the size of the training dataset and reduce overfitting. Data augmentation methods employed for geometric transformation invariance are rotation, shear, skewness, contracting/stretching, and flipping. In contrast, for noise invariance, the techniques used are edge detection, Gaussian blur, sharpen, shadow, and embossing. To make the training model robust and increase the training dataset, we propose three different combinations of augmentation: CSR (Contracting/Stretching, Skew, Rotate), CGSR (Contrast, Gaussian Blur, Skew, Rotate), CSR-CGSR (Contracting/ Stretching, Skew, Contrast, Gaussian Blur, Skew, Rotate). The complex transformation of Image patches is shown in Fig. 4.

D. DCNN Based Feature Extraction With SVM Classification (DFESC) Model

1. Feature Extraction

Feature extraction was carried out by passing the augmented image patches through the pre-trained fine-tuned VGG-19 network. DCNN based VGG-19 network is most popularly used because of its simplicity as it uses only three × three convolutional layers piled up on top of each other to increase depths. For dimensionality reduction, down-sampling of the input image (including image convolution, hidden-layer output matrix, etc.) is employed in this network. It consists of 19 layers, including five stages of 16 convolutional layers, rectified linear units (ReLU) activation, pooling layer of max type, and three fully connected (FC) layers. For better classification accuracy, the last FC layer is connected to the SVM classifier rather than the softmax layer that performs classification.

For feature extraction, first augmented image patches are normalized to zero mean and unit variance before feed into the VGG-19 network. The architecture VGG-19 based DFESC model is shown in Fig. 5 in which the first two layers of convolutional are trailed by max-pooling layer, and the same arrangement is continued for succeeding two layers as shown in Fig. 6. The remaining eight layers are arranged in a group of four convolutional layers followed by max pooling. This

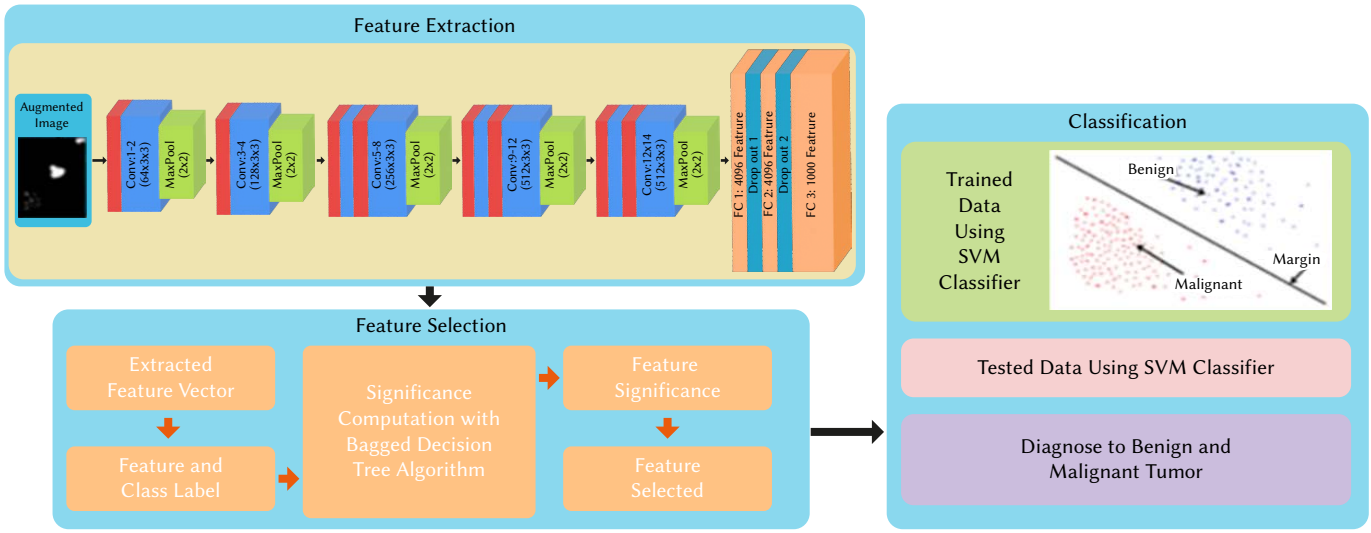


Fig. 5. Architecture of DCNN based Feature Extraction with SVM Classification (DFESC) Model.

Algorithm 3: Data Augmentation using CSR, CGSR, and CSR-CGSR Algorithm

- Input: Segmented Image Patches,
- Transformation parameter: Rotation (-90°, -45°,90°,45°), Gaussian Blur (σ=0.5,1.0,2.0.5), Contrast (lightness value =1.00, 1.5, 2.0, 0.25), Skew (Left, Right, Forward, Backward).
- Output: Augmented Image Patches
- Perform Data Augmentation using CSR, CGSR, and CSR-CGSR Transformation Algorithm
- CSR Augmentation Algorithm:
  i. First, split the segmented image patch database into two equal sets.
  ii. Apply Contracting/Stretching transformation to 50% segmented image patches, and the remaining 50% image patches are undergone through skew transformation.
  iii. Contracting/Stretching (CS) transformed images again split equally into two sets.
  iv. The applied skew transformation over 50% CS transformed image, and the rest of 50% CS transformed image is stored in CSR dataset.
  v. Skew transformed images again, split equally into two sets.
  vi. Applied rotation transformation over 50% skew transformed images, and the rest of 50% skew changed ideas is stored in the CSR dataset.
- CGSR Augmentation Algorithm:
  i. First, all segmented images are modified by the application of contrast transformation.
  ii. Transformed Contrasted images are then passed through Gaussian filters.
  iii. Shadow transformation is then further applied to images passed out from the Gaussian filter.
  iv. Shadow transformed images then split equally into two sets.
  v. Applied rotation transformation over 50% shadow transformed images, and the rest of the 50% shadow transformed image is stored in the CGSR dataset.
  vi. CSR-CGSR Augmentation Algorithm: Two augmentation techniques CSR and CGSR, are combined.
  vii. Initially, image transformation is performed according to the CSR transformation workflow, including Contracting/Stretching (CS), skewing, and rotation.
  viii. Afterward, CSR transformed images then go through CGSR transformation, including contrast reduction, Gaussian blurring, shadowing, and rotation.
  ix. All the transformed images are then stored in a CSR-CGSR dataset.

arrangement is then connected to the last three thoroughly combined (FC) that contains 4096, 4096, and 1000 nodes, respectively. The outcome from these layers resulted in 4096, 4096, and1000 features, respectively.

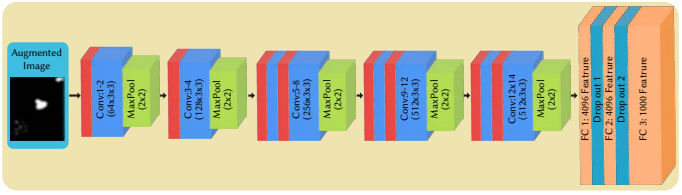


Fig. 6. Architecture of VGG-19 DCNN Based Feature Extraction.

2. Feature Selection

To reduce redundancy, we perform feature selection using bagged decision tree [31] algorithms like Extra Tree and Random Forest. Using these algorithms, we determine the significance of features and then select the quality based on their ranks using (5) in which a part with 95% significance is considered.

$$Significance\ of\ Selected\ Feature = 0.95 * Significance \quad (5)$$

3. Dataset Splitting

Selected feature vectors are then into three parts for training, validation, and testing. In this study, we hierarchically split the dataset. First, the dataset is divided into training and testing datasets in the percentage ratio of 85:15. The training dataset is then further divided into training and validation sets with a percentage ratio of 90:10. A diagrammatic view of the data split is shown in Fig. 7.

Fig. 7. Dataset Splitting.

4. SVM Classification

In this work, augmented feature vectors are combined with SVM classifiers that minimize classification error by determining the best possible separating hyperplanes. With a given labeled feature pair (p_f, q_f) where p_f represent feature vector and $q_f \in (+1, -1)$ represent

whether a given instance corresponds to the class or not. The (6) defines the formation of binary SVM cost function:

$$\frac{1}{2}\, x^T x + \mathrm{C}_r \sum_f \quad \varepsilon_f$$
$$subjecttoq_f \,.\, \left(x^T \emptyset(p_f) + y\right) \geq 1 - \varepsilon_f, \varepsilon_f \geq 0 \tag{6}$$

Where x and y represent separating hyperplane parameter, ε_f represent penalty error for a loose variable (located at the erroneous side of hyperplane margin), C_r represent control parameter for regularization that tradeoff between hyperplane margin and penalty error ε_f, $\emptyset(p_f)$ Represent the input vector non-linear transformation function. The two separatings hyperplane can be defined as $x^T \emptyset\ (p_f) + y = 1$ and $x^T \emptyset\ (p_f) + y = 0$, while margin width is defined as $\frac{2}{\|x\|}$. We use the radial basis function (RBF) kernel is used for SVM classifier, which is defined in (7):

$$K(p_f, p_g) \equiv \emptyset(p_f)^T \emptyset(p_g) = e^{-\gamma \|p_f - p_g\|^2} \tag{7}$$

Algorithm 4: DCNN based Feature Extraction with SVM classification

1. Input: AugmentedImage $Aug_I = 224 \times 224 \times 3$;
2. Array Initialization : Number of filter $\longrightarrow NOF_c[5] = \{64, 128, 256, 512\}$;
3. Feature extraction using VGG-19 net
4. f or $i = 1$ to 8 step 1

Number of kernel: $NOC_c = NOF_c\,[i]$;

Begin:

if $(i \leq 2)$ then

Begin:

for $j = 1$ to 2 step 1

Begin:

$conv_{(i,j)}$: $performconv(NOK_c, 3, 3;$ stride $= 1)$;

$(R^k)_{x,y} = (Wt^k * Aug_I)_{x,y} + bs^k$

$ReLu_{i,j}$:

$g(R) = max\ (0, R)$

End

$MaxPool_i$:

$pool_{size} = (2, 2)$;

$stride = (2, 2)$;

End

else if $(i \geq 3$ && $i \leq 5)$ then

Begin:

for $j = 1$ to 4 step 1

Begin:

$conv_{(i,j)}$: $performconv(NOK_c, 3, 3;$ stride $= 1)$;

$(R^k)_{x,y} = (Wt^k * Aug_I)_{x,y} + bs^k$

$ReLu_{i,j}$:

$g(R) = max\ (0, R)$

End

$MaxPool_i$:

$pool_{size} = (2, 2)$;

$stride = (2, 2)$;

End

else

Begin:

FC_i: Extract Feature

$ReLu_i$: Set all negative value to zero

$drop_i$: Perform 50% dropout

End:

End

5. Perform features election using bagged decision tree algorithm by estimating significance

$Significance\ of\ Selected\ Feature = 0.95 * Significance$

6. Perform SVM classifcation using RBF kernel

IV. Experimental Results

A. Dataset Description

DDSM [32] and CBIS – DDSM are the standard dataset containing a medical image of tumors for breast cancer detection and classification. CBIS – DDSM [33] is the latest version dataset that digitized mammogram images in DICOM standard format, while DDSM contains a lossless-JPEG design. These datasets were downloaded from the website of CBIS-DDSM that consist of 2478 mammography medical image of 1249 female, and these include mediolateral and craniocaudal (CC) oblique view. In this study, each oblique was considered as a separate image. The application of most of the sensors is now focused on evaluating the time domain of acquired sensors, typically by the magnitude and frequency of movement. However, this tentative analytical approach can neglect certain valuable sensing signals, such as identification information

We split the dataset images into three sets: a training set containing 1903 images, a validation set consisting of 199 photos, and a testing set with 376 illustrations. CBIS-DDSM dataset consists of pixel-based annotation for the Region of interest (ROI) with the label: Benign and Malignant. These labels are then further elaborated based on ROI as mass or calcification. Convert the downloaded mammogram images into PNG format and without cropping downsized images to 1152X896 using interpolation.

As we propose patch-based analysis, segmentation, and classification; therefore, we created two patch datasets by sampling patches of the image from the background and ROI region. We extract all image patches of 224×224-pixel size, and this patch size is sufficiently large enough that cover almost all ROI annotation. All patches were classified into one of the five categories: benign calcification, malignant calcification, soft mass, malignant mass, and background.

B. Performance Evaluation

In this section, we discuss the performance of the proposed system for multi-class breast cancer tumor classification. Therefore, performance evaluation of proposed DCNN based Feature Extraction with SVM classification (DFESC) model is discussed for two cases: Before Augmentation and After Segmentation. For the performance assessment of complex structures, there are many methods. However, most of the approaches proposed in the field of medical image processing only face the problem of identifying different metrics that allow precision from a strictly geometric and quantitative point of view to be assessed. Finally, potential future directions for performance assessment research in medical image segmentation are suggested. In essence, data integrity means that the data is correct and has not been wrongfully changed in any way. Inaccurate records can become a significant health concern for patients and an immense liability for providers, leading to fraud, misuse, lack of data, and incorrect or inadequate treatment.

In case 1, classification is performed over the original dataset (without augmentation), while in case 2, classification is performed over an augmented dataset, which is generated by using the proposed data augmentation algorithm (CSR, CGSR, CSR-CGSR). Classification result of proposed DFESC model with and without augmentation is shown in Fig. 8(a) and 8(b), whereas. Their corresponding confusion matrix analysis is shown in Fig. 9(a) and 9(b).

Fig. 8(a) and 8(b) showed that the proposed DFESC model with data augmentation algorithm predicted all five classes correctly with higher probability than without augmentation. It was also observed that background class is easily expected while malignant calcification class is hardest to predict.

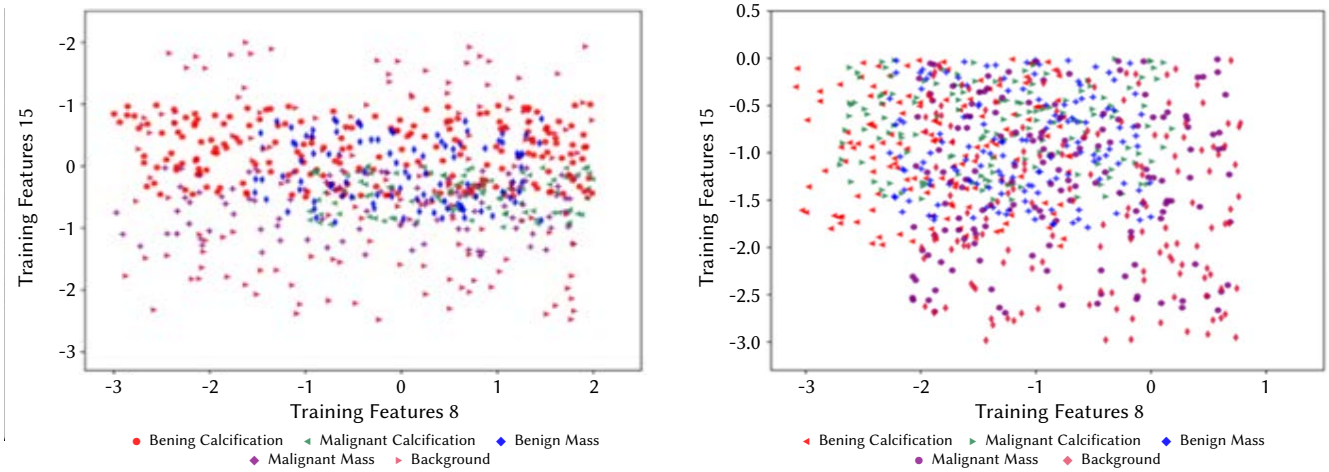


Fig. 8. Classification Result into 5-Class using Proposed DFESC Model (a) Without Augmentation (b) With Augmentation.

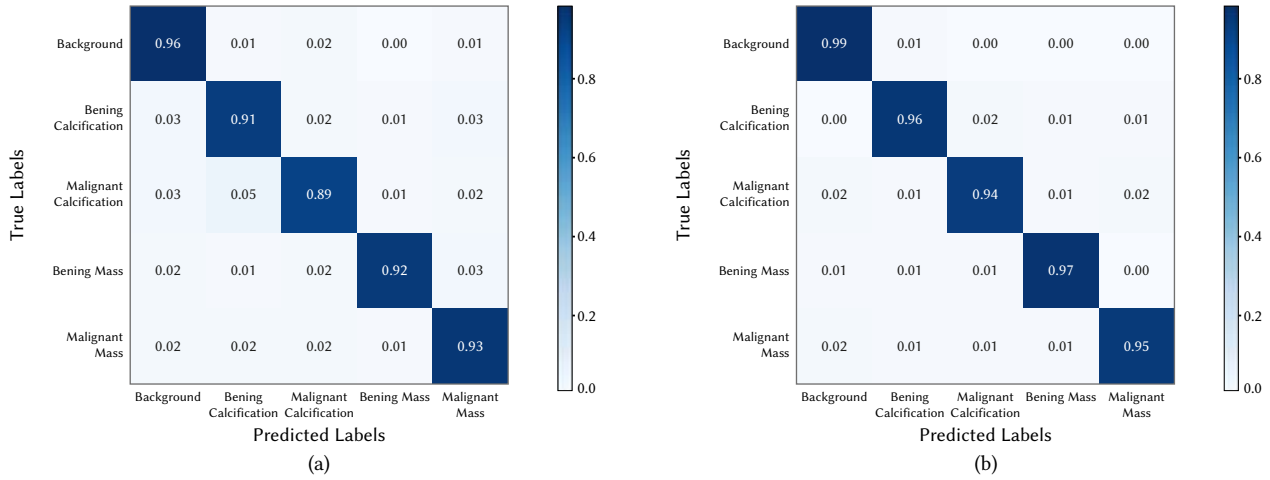


Fig. 9. Tumor Classification into 5-class using Confusion matrix analysis (a) Without Augmentation (b) With Augmentation.

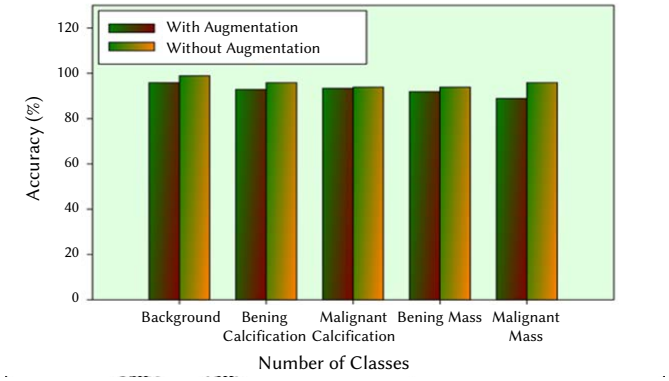


Fig. 10. Accuracy of each Tumor Class with and without Augmentation.

Based on the confusion matrix analysis, we formulate the proposed (DFESC) model's accuracy for both cases. Fig. 9 represents classification accuracy based on confusion matrix analysis for CBIS – DDSM dataset using the proposed classification model for both cases. Fig. 10 observed that classification accuracy after augmentation for background, benign calcification, malignant calcification, benign mass, and malignant mass are 99%, 96%, 94%, 97%, and 95%. In comparison, accuracy before augmentation is 96%, 91%, 93%, 92%, and 89% for background, benign calcification, malignant calcification, benign mass, and malignant mass. Based on this result, the data augmentation algorithm improves the classification accuracy up to a greater extent.

The overall classification accuracy with and without augmentation for the proposed DFESC model using the CBIS-DDSM dataset is shown in Fig. 11. It is marked from Fig. 11 that with the inclusion of the proposed data augmentation algorithm, the accuracy is improved largely. Accuracy starts from 40% for without augmentation while 58 % for the first epoch with augmentation dataset. In addressing these healthcare problems, big data analytics will help. With the help of predictive analytics, healthcare providers will cut healthcare costs and provide quality care. Big data frequently helps to minimize prescription mistakes by improving financial and administrative productivity and reducing hospital admissions.

Further, we assess the performance of the proposed work by computing AUCs per-image over an independent test set for two cases: Without Augmentation and With Augmentation. In case 1, classification is carried out over the original dataset (without augmentation), while classification is carried out over the augmented dataset in case 2, generated using the proposed data augmentation algorithm (CSR, CGSR, CSR-CGSR). The overall classification accuracy for the proposed DFESC model using the CBIS-DDSM dataset with and without augmentation is shown in Fig. 11. The accuracy is significantly enhanced with the inclusion of the submitted data augmentation algorithm.

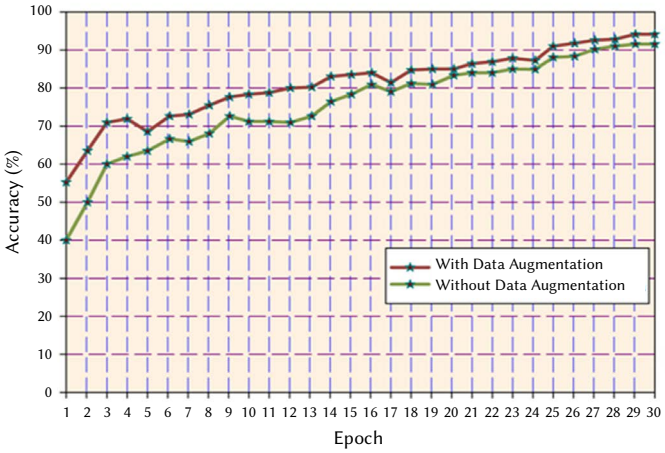


Fig. 11. Overall Accuracy of Proposed Method with and without Augmentation.

Case 1: Without Augmentation

First, we classify the tumor into 5- class over the original dataset (without augmentation) using the proposed DCNN based Feature Extraction with SVM classification (DFESC) model. Next, we generate a ROC curve with AUC computation for the DFESC model without performing augmentation. The ROC curve without augmentation into 5-class is shown in Fig. 12. The corresponding computed AUC values from the ROC curve are 0.91,0.98,0.97, 0.95, and 0.99 for benign calcification, malignant calcification, benign mass, malignant mass, and background. Overall average AUC of the proposed DFESC model before augmentation is 0.96.

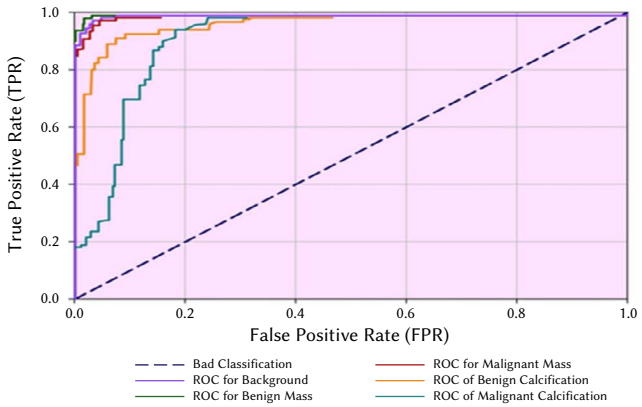


Fig. 12. ROC curve for Proposed DFESC model without augmentation.

Case 2: With Augmentation

In this case, we first augmented the original data using the proposed data augmentation algorithm (CSR, CGSR, CSR-CGSR) and then applied this augmented data to our proposed DCNN based Feature Extraction with SVM (DFESC) model for classification. Fig. 12 represents the ROC curve for benign calcification, malignant calcification, benign mass, malignant mass, and background class with an AUC of 0.990, 0.994, 0.991, 0.990, and 0.998 augmentations. Overall average AUC of the proposed DFESC model after augmentation is 0.994.

The ROC curve represented in Fig. 12 (without augmentation) and Fig. 12 (with augmentation) observed that classification accuracy for augmented data is higher for each class compared to without augmentation. Subsequently, compare the proposed DFESC model's performance using data augmentation algorithm with other CNN architecture, including ResNet, AlexNet, VGG-19, VGG-19 +ResNet. The performance comparison for all classification models using the ROC curve is shown in Fig. 13. It is evident from Fig. 14 that the proposed System represents a high AUC of 0.98 while the AUC of ResNet, AlexNet, VGG-19, VGG-19 +ResNet are 0.92, 0.88, 0.87, and 0.89. It shows that classification accuracy is high if the SVM classifier performs classification using RBF kernel over the feature extracted from the VGG-19 based DCNN.

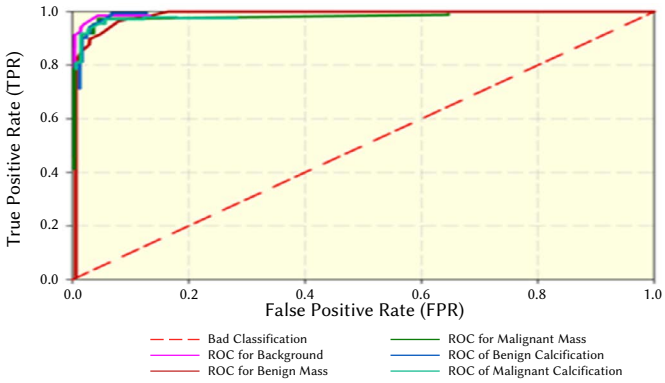


Fig. 13. ROC curve for Proposed DFESC model with Augmentation.

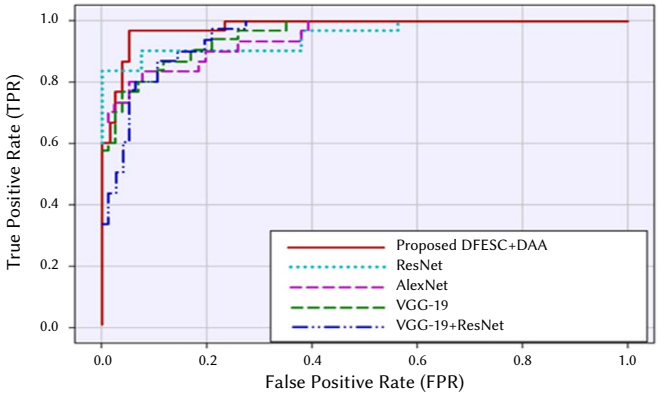


Fig. 14. ROC Curve Representation for Performance Comparison Analysis of Different CNN Classification Architecture with Proposed Method.

V. Conclusion

Deep Learning achieves a high level of accuracy in detecting and classifying multi-class breast cancer tumors. In this paper, we propose a combined approach based on deep learning and machine learning methods that perform classification over augmented data, resulting in better accuracy than other classification techniques. Authors suggest -fold method: Firstly, Authors employed Hierarchical CNN approach for tumor segmentation; Secondly, we propose three data augmentation algorithm-CSR, CGSR, and CSR-CGSR that perform augmentation over the segmented image; thirdly, we proposed VGG-19 CNN based feature extraction followed by feature selection using bagged decision algorithm (Random Forest, Extra Tree); finally, we perform classification using multi-class SVM classifier that classifies the breast cancer tumor image into five different classes: benign calcification, malignant calcification, benign mass, malignant mass, and background with high accuracy. Authors evaluate our proposed system's performance by performing the classification using the proposed DFESC model over original data and augmented data. From the simulation result, we observed that classifying the tumor using the DFESC model over augmented data achieves higher accuracy than without augmentation. Subsequently, the Authors perform a comparative analysis of different deep learning CNN architecture

(like ResNet, AlexNet, VGG-19, VGG-19+ResNet) with our proposed data augmentation based DFESC model. Form the comparison ROC curve, the Authors determine that the proposed method outperforms other CNN architecture. In the future, this work can be extended for fine-grained classification of each class with the examination of light-weight CNN architectures to steadiness the accuracy and efficiency.

References

[1] D. E. O'Leary, "Big Data, the Internet of Things and the Internet of Signs," *Intelligent Systems in Accounting, Finance and Management*, vol. 20, no. 1, pp. 53-65, 2013, doi: 10.1002/isaf.1336.

[2] D. W. Pennington, M. Margni, J. Payet, and O. Jolliet, "Risk and regulatory hazard-based toxicological effect indicators in life-cycle assessment (LCA)," *Human and Ecological Risk Assessment*, vol. 12, no. 3, pp. 450-475, 2006, doi: org/10.1080/10807030600561667.

[3] K. Van Hulle, "Solvency II: state of play and perspectives," *Zeitschriftfür die gesamte Versicherungswissenschaft*, vol. 100, no. 2, pp. 177-192, 2011, doi: org/10.1007/s12297-011-0138-2.

[4] H. K. Cho, K. P. Bowman, and G. R. North, "A comparison of gamma and lognormal distributions for characterizing satellite rain rates from the tropical rainfall measuring mission," *Journal of Applied meteorology*, vol. 43, no. 11, pp. 1586-1597, 2004, doi: 10.1175/JAM2165.1.

[5] H. R. Boveiri, R. Khayami, M. Elhoseny, and M. Gunasekaran, "An efficient Swarm-Intelligence approach for task scheduling in cloud-based internet of things applications," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 9, pp. 3469-3479, 2019, doi: org/10.1007/s12652-018-1071-1.

[6] S. Thamburasa, S. Easwaramoorthy, K. Aravind, S. B. Bhushan, U. Moorthy, "Digital forensic analysis of cloud storage data in IDrive and Mega cloud drive," in *2016 International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India, 2016, pp. 1-6.

[7] V. E. Sathishkumar, and Y. Cho, (2019, December). "Cardiovascular disease analysis and risk assessment using correlation based intelligent system," in *Basic & Clinical Pharmacology & Toxicology,* Hoboken, NJ USA, Wiley, 2019, pp. 61-61.

[8] A. Ahilan, G. Manogaran, C. Raja, S. Kadry, S. N. Kumar, C. A. Kumar, T. Jarin, K. Sujatha, M. K. Priyan, G. C. Babu, N. S. Murugan, and Parthasarathy, "Segmentation by fractional order darwinian particle swarm optimization based multilevel thresholding and improved lossless prediction based compression algorithm for medical images," *IEEE Access*, vol. 7, pp. 89570-89580, 2019, doi: 10.1109/ACCESS.2019.2891632.

[9] Y. Shi, C. Sun, Q. Li, L. Cui, H. Yu, and C. Miao, (2016, March). "A fraud resilient medical insurance claim system," in *Thirtieth AAAI Conference on Artificial Intelligence, (AAAI Press)*, Phoenix, Arizona, 2016, pp. 4393-4394.

[10] K. Fang, Y. Jiang, and M. Song, "Customer profitability forecasting using Big Data analytics: A case study of the insurance industry," *Computers & Industrial Engineering*, vol. 101, pp. 554-564, 2016, doi: org/10.1016/j.cie.2016.09.011.

[11] Y. Wang, and W. Xu, "Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud," *Decision Support Systems*, vol. 105, 87-95, 2018, doi: org/10.1016/j.dss.2017.11.001.

[12] D. A. Koutsomitropoulos, A. K. Kalou, "A standards-based ontology and support for Big Data Analytics in the insurance industry," *ICT Express*, vol. 3, no. 2, pp. 57-61, 2017, doi: org/10.1016/j.icte.2017.05.007.

[13] W. Lin, Z. Wu, L. Lin, A. Wen, and J. Li, "An ensemble random forest algorithm for insurance big data analysis," *IEEE Access*, vol. 5, pp. 16568-16575, 2017, doi: 10.1109/ACCESS.2017.2738069.

[14] Y. Ren, K. Zhang, and Y. Shi, (2019, November). "Survival Prediction from Longitudinal Health Insurance Data using Graph Pattern Mining," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, San Diego, CA, USA, 2019, pp. 1104-1108.

[15] N. Rayan, "Framework for Analysis and Detection of Fraud in Health Insurance," in *2019 IEEE 6th International Conference on Cloud Computing and Intelligence Systems (CCIS)*, Singapore, 2019, pp. 47-56.

[16] Y. M. Chae, S. H. Ho, K. W. Cho, D. H. Lee, and S. H. Ji, "Data mining approach to policy analysis in a health insurance domain," *International journal of medical informatics*, vol. 62, no. (2-3), pp. 103-111, 2001, doi: org/10.1016/S1386-5056(01)00154-X.

[17] M. S. Viveros, J. P. Nearhos, and M. J. Rothman, "Applying data mining techniques to a health insurance information system," in *VLDB'96, Proceedings of 22th International Conference on Very Large Data Bases*, Mumbai (Bombay), India, 1996, pp. 286-294.

[18] X. Jiang, S. Pan, G. Long, F. Xiong, J. Jiang and C. Zhang, "Cost-Sensitive Parallel Learning Framework for Insurance Intelligence Operation," in *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9713-9723, 2019, doi: 10.1109/TIE.2018.2873526.

[19] C. S. Wang, S. L. Lin, T. H. Chou, and B. Y. Li, B. Y. "An integrated data analytics process to optimize data governance of non-profit organization," *Computers in Human Behavior*, vol. 101, pp. 495-505, 2019, doi: 10.1016/j.chb.2018.10.015.

[20] N. S. Murugan, and G. U. Devi, "Feature extraction using LR-PCA hybridization on twitter data and classification accuracy using machine learning algorithms," *Cluster Computing*, vol. 22, no. 6, pp. 13965-13974, 2019, doi:10.1007/s10586-018-2158-3.

[21] N. S. Murugan, and G. U. Devi, "Detecting streaming of Twitter spam using hybrid method," *Wireless Personal Communications*, vol. 103, no. 2, pp. 1353-1374, 2018, doi: org/10.1007/s11277-018-5513-z.

[22] J. Sato, K. Goda, M. Kitsuregawa, N. Nakashima, and N. Mitsutake, "Novel Analytics Framework for Universal Healthcare Insurance Database," *AMIA Summits on Translational Science Proceedings*, vol. 2019, pp. 353-362, 2019.

[23] K. Umemoto, K. Goda, N. Mitsutake, and M. Kitsuregawa, "A Prescription Trend Analysis using Medical Insurance Claim Big Data," in *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, Macao, China, 2019, pp. 1928-1939.

[24] J. Ahmad, K. Muhammad, J. Lloret, and S. W. Baik, "Efficient conversion of deep features to compact binary codes using Fourier decomposition for multimedia big data," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 7, pp. 3205-3215, 2018, doi: 10.1109/TII.2018.2800163.

[25] J. Ahmad, K. Muhammad, and S. W. Baik, "Medical image retrieval with compact binary codes generated in frequency domain using highly reactive convolutional features," *Journal of medical systems*, vol. 42, no. 2, pp. 24, 2018, doi:10.1007/s10916-017-0875-4.

[26] J. Ahmad, K. Muhammad, S. I. Kwon, S. W. Baik, and S. Rho, "Dempster-Shafer fusion-based gender recognition for speech analysis applications," in *2016 International Conference on Platform Technology and Service (PlatCon)*, Jeju, Korea (South), 2016, pp. 1-4.

[27] K. Muhammad, J. Ahmad, Z. Lv, P. Bellavista, P. Yang, and S. W. Baik, "Efficient deep CNN-based fire detection and localization in video surveillance applications," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 7, pp. 1419-1434, 2018, doi: 10.1109/TSMC.2018.2830099.

[28] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad and S. W. Baik, "Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features," in *IEEE Access*, vol. 6, pp. 1155-1166, 2018, doi: 10.1109/ACCESS.2017.2778011.

[29] K. Muhammad, J. Ahmad, and S. W. Baik, "Early fire detection using convolutional neural networks during surveillance for effective disaster management," *Neurocomputing*, vol. 288, pp. 30-42, 2018, doi: org/10.1016/j.neucom.2017.04.083.

[30] M. Sajjad, S. Khan, T. Hussain, K. Muhammad, A. K. Sangaiah, A. Castiglione, C. Esposito, and S. W. Baik, "CNN-based anti-spoofing two-tier multi-factor authentication system," *Pattern Recognition Letters*, vol. 126, pp. 123-131, 2019, doi: 10.1016/j.patrec.2018.02.015.

[31] D. Guan, W. Yuan, Y. K. Lee, K. Najeebullah, and M. K. Rasel, "A review of ensemble learning based feature selection," *IETE Technical Review*, vol. 31, no. 3, pp. 190-198, 2014, doi: org/10.1080/02564602.2014.906859.

[32] M. Heath, K. Bowyer, D. Kopans, P. Kegelmeyer, R. Moore, K. Chang, and S. Munishkumaran, "Current status of the digital database for screening mammography," in *Digital mammography*, Springer, Dordrecht, 1998, pp. 457-460.

[33] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Scientific data*, vol. 4, pp. 170177, 2017.

Cheng Zhang

Cheng Zhang is an associate professor of the Department of Sociology in Jiangnan University, China. Her research interests include Social Security, Social Services, and Project Management, more than 9 papers published and 2 books published.

Dr. B. Vinodhini

Vinodhini. B. is presently an Assistant Professor in the Department of Computer Science & Engineering, SNS College of Technology, affiliated to Anna University-Coimbatore, Tamilnadu, India. Her research interests include Data Analytics, Computer Networks, and Wireless Communications. She filed 4 Patents, published 20 Research Articles and Published papers in International and National Conferences.

Dr. BalaAnand Muthu

Dr. BalaAnand Muthu is working as Associate Professor in the Department of Computer Science & Engineering at Adhiyamaan College of Engineering, India. His area of interest includes Big Data Analytics, Social Networks, Internet of Things in Healthcare. He is a member of IEEE and ACM. Has published many research articles in SCI, SCIE, Scopus indexed peer review journals. Also, handled Guest lectures, Intensive Workshop, Hands on programming in Hadoop, Spark, Grid & Cloud Computing at various technical institutions around Tamil Nadu. He is serving as reviewer in Computer Communication, IEEE Access, Multimedia Tools & Applications, International Journal of Parallel Programming, Enterprise Information System, Computer Networks, Measurement, Computer & Electrical Engineering, Wireless Personal Communication, Cluster Computing, Computational Intelligence, IET Transport Systems and so on.

A Comparative Evaluation of Bayesian Networks Structure Learning Using Falcon Optimization Algorithm

Hoshang Qasim Awla[1*], Shahab Wahhab Kareem[2,3], Amin Salih Mohammed[3,4]

[1] Department of Computer Science, Faculty of Science, Soran University, Erbil (Iraq)
[2] Department of technical information system Engineering, Erbil Technical Engineering College, Erbil Polytechnic University, Erbil (Iraq)
[3] Department of Information technology, College of Engineering and computer science, Lebanese French University, Erbil (Iraq)
[4] Department of Software and Informatics, College of Engineering, Salahaddin University-Erbil (Iraq)

Received 16 January 2020 | Accepted 16 September 2022 | Published 19 January 2023

Abstract

Bayesian networks are analytical models that may represent probabilistic dependent connections among variables and are useful in machine learning for generating knowledge structure. Due to the vastness of the solution space, learning Bayesian network (BN) structures from data is an NP-hard problem. The score and search technique is one Bayesian Network structure learning strategy. In Bayesian network structure learning the Falcon Optimization Algorithm (FOA) is presented and evaluated by the authors. Inserting, Reversing, Moving, and Deleting, are used in the method to create the FOA for finding the best structural solution. The FOA algorithm is based on the falcon's searching technique during drought conditions. The suggested technique is compared to the score metric function of Pigeon Inspired search algorithm, Greedy Search, and Antlion optimization search algorithm. The performance of these techniques in terms of confusion matrices was further evaluated by the authors using a variety of benchmark data sets. The Falcon optimization algorithm outperforms the previous algorithms and generates higher scores and accuracy values, as evidenced by the results of our experiments.

Keywords

Bayesian Network, Falcon Optimization Search Algorithm, Global Search, Local Search, Score And Search, Structure Learning.

DOI: 10.9781/ijimai.2023.01.004

I. Introduction

In machine learning, Bayesian networks (BN) are one of the main analytical models for developing the probabilistic structure of knowledge [1]. They may be used in a variety of contexts, including knowledge design, argumentation, and inference [2]. There are two stages to learning a Bayesian network: parameter learning and structure learning. The focus of this paper is on Bayesian network structure learning. In structure learning, three procedure is needed such as strategies on the conditional independence, calculating score for optimization technique, and combining different approaches [3]. In Bayesian network, directed acyclic graph (DAG) is the main structure, and this structure contain two components key: parameters and network structure. The structure displays interrelationships between variables, whereas the parameters represent conditional probabilities. Without a great search technique, it's hard to solve the Bayesian network's learning structure. Meanwhile, although learning the Bayesian network structure from a dataset to produce the best result is NP-hard [4] a lot of work has gone into developing estimate methods for learning the network structure. Generally, constraint based approach and score-and-search strategy are two different mechanism in structure learning of Bayesian network [5]. The main mechanism for searching on the Bayesian network space is score and search mechanism, and continuously evaluate each potential network structure until the correct metric value is found.

Score-based methods utilize a metric to quantify the network and data available before looking for a structure that maximizes the score [6]. The scoring function method was implemented using two key criteria: one of them is Bayesian score, and the second is Information-theoretic score. Information-theoretic score has been used by the Normalized Minimum Likelihood (NML), BDeu (Bayesian Dirichlet equivalent uniform ("u" for uniform joint distribution), Bayesian Information Criterion (BIC) and log-likelihood (LL), Minimum Description Length (MDL), and Akaike Information Criterion (AIC) [7]. The Bayesian score is used in K2, BDe (Bayesian Dirichlet ("e" for likelihood-equivalence), BD (Bayesian Dirichlet), and Mutual Information Tests, (MIT) [8]. There are several sorts of search strategies for discovering the optimal solution to the structure learning issue. Simulated Annealing Algorithm [7], Particle Swarm optimization [9], Ant Colony Algorithm [10], Antlion optimization [3], Hybrid Algorithms

* Corresponding author.
E-mail address: hoshang.awla@soran.edu.iq

([11], [12], [13], [14], [15] [16]), Bacterial Foraging Optimization [17]., Breeding Swarm Algorithm (20), Genetic Algorithms (GOMEA) [18][19], Falcon optimization is another reducing met heuristic for Bayesian network structure learning. These findings suggest and examines this method for addressing the Bayesian network structure learning difficulty. BNs are increasingly useful mechanisms for risk assessment, risk evaluation, resource planning for data science and environmental management.

BNs is simple and straightforward graphical presentation that is used to manage causal inference and risk monitoring, so they have a lot of benefits over regression-based methods. The Bayesian network is frequently used this to clearly visualize the connection between the emergence of several major illnesses and patient related variables during the time [22]. The rest of this paper will be arranged as follows. The notion of structure learning in Bayesian networks is introduced in Section II. The Falcon Optimization Search Algorithm in Section III, is briefly introduced. In section IV, we go through the approach in depth and show the results of the experiment. Section V contains the conclusions.

II. Structure Learning of Bayesian Networks

The Bayesian network is essentially made up of two parts: (G, P).

The DAG G(V; E) denotes a collection of conditional probable distributions (CPD) that includes all variables Xi. P = P (Xi | Pa (Xi) denotes a set of conditional probability distributions (CPD) that includes all parameters Xi (vertices from a graph). Pa(Xi) commonly denotes the parents of the node Xi in G [29]. Probabilistic as a simple pairing for a (G; P) network may be outlined to apply this equation (1):

$$P(X_i, \dots X_n) = \prod_{i=1}^{n} P(X_i | Pa(X_i)) \tag{1}$$

A scoring function, is based on a number of factors, including Bayesian techniques, information and entropy, and the length of the minimal description [30]. Bayesian network posterior probability may be stated as follows using Bayesian inference rules:

$$P(G \mid D) = P(D \mid G) \cdot P(G) / \sum_{G'} P(D \mid G')P(G') \tag{2}$$

P(G') is the posterior probability and reflects the parameters of the model in equation (2). As a result, as long as the minimum probability of all potential structures is known, It is possible to establish the prior probability of the network structure. [31]. P(D|G) stands for marginal likelihood and is defined: P(D) is used as a normalizing constant:

$$P(D \mid G) = \int P(D \mid G, \theta)P(\theta \mid G)d\theta \tag{3}$$

P(D) in Byesian network structure is supposed to be independent of network G. Structure learning methods compare the present and prior scores of the structure using score-based methods. [32] is the final representation of the score:

$$\text{BDe (G, D)} = \prod_{i=1}^{p} BDe(\text{Xi},$$

$$\Pi\text{Xi}) = \prod_{i=1}^{p} \prod_{j=1}^{qi} \left\{ \frac{\Gamma(\alpha\text{ij})}{\Gamma(\alpha\text{ij} + \text{nij})} \prod_{k=1}^{ri} \frac{\Gamma(\alpha\text{ij} + \text{nijk})}{\Gamma(\alpha\text{ijk})} \right\} \tag{4}$$

where:

- p is the number of nodes in G;
- ri is the number of classes regarding node Xi;
- qi is the number of preparations from the groups of Xi's parents;
- nijk denotes the amount of participants who might have node Xi's jth class and its parents' kth arrangement.

III. Falcon Optimization Algorithm

Metaheuristics are algorithms that are inspired by nature and are used to find approximate solutions to computationally difficult optimization problems. Metaheuristics have been used to exploit swarming characteristics of animals such as the Firefly-BAT [33], Cuckoo [34],GWO [35], Deep multi-model fusion [36]antlion, pigeon, fish, bee, and others. Homogeneity, adaptability, illation-free tools, and the capacity to avoid local optima are all characteristics of metaheuristics [37]. The suggested metaheuristic algorithm in [38] was inspired by the falcon's hunting activity. The For probabilistic inhabitants' tasks, the Falcon optimization search algorithm is a dependable and stable process that encourages parameter values for its three item resolution.

The proposed strategy was inspired by the chasing style of falcons when on the hunt for prey while in flight. Falcons strategy foe hunting is determined by their needs. However, specific strategies emerge.

Based on several studies [39], [40]. Falcons are high-performance fliers among birds. The suitable targets are examined for the boundaries of flying achievement in distinct stages of heightened hunting [41]. Determining the physical power of flight, calculating average flight velocities, and responding to wind are some of the flight implementation strategies in the framework. [41]. Falcons are one of the fastest animals on the planet, with stoops reaching speeds of above 300 km/h. Falcons can breathe freely due to little thin tubercles in their noses that direct air via high-speed stoops. The majority of the hunting will be done in the morning and at night. The predominant source of food is small-medium-sized birds, with insects such as cicadas, moths, and locusts arriving only occasionally [42].

Falcons approach their prey in a number of ways while flying. The route is divided into two sections: the first section is logarithmic curve in which the falcon keeps it's own head straight whereas peering slantingly the prey in the outcomes acuity, the second one is a straight segment in which the falcon wants to fly to a prey if in the vision and dives when it becomes close to it. As a result, the falcons mainly obtain a movement which can be separated into three phases: the First Phase, which involves prey exploration; the Second Phase, which involves improving the look into the logarithmic curve; and Third Stage, which involves the dive itself, which can lead to the success outcome, like as picking a prey. Instead, depending on its prior experiences, the falcon immediately changes its behavior. The five steps of a quick method to adopting FOA are shown below [38].

Step 1: Determine the parameters of the optimization task, such as falcons number (NP), limit of speed (Vmax), rate of (cc) cognitive rate, the social constant (sc), the following constant (fc), probability of dive (DP), and the alertness probability (AP) (AP).

Step 2: Based on the boundary conditions where each falcon's position is established, assign the falcons' velocity and location in a D-dimensional space at random, while keeping the number of NP candidates in all D dimensions in mind. Between the Vmax and Vmin limitations, which are established as follows, the velocities are generated at random:

$$\text{Vmax} = 0.1^{*}\text{ub} \tag{5}$$

$$\text{Vmin} = -\text{Vmax} \tag{6}$$

Where ub denotes the upper border of each dimension's boundary region. Create the pairings of values (pAP, pDP) for each falcon at random to compare with the awareness and dive probability.

Step 3: Find the best (xbest) and global (gbest) locations by calculating the fitness value. This fitness value, of, is calculated for each bird. The chosen positions will be utilized to create new locations based on the logic that governs the dive's movement and the probabilities of awareness.

Algorithm: Structure Learning of Bayesian Network based on falcon optimization algorithm

INPUT: - datasets

Population size, NP;

Maximum speed, Vmax;

Values of cognitive Cc, social, Sc and following Fc, constant.

Value of awareness Probability (AP) and Dive probability (DP);

t_{max}*: maximum number of iteration number:* X_{max} *:upper boundary, and* X_{min} *: - lower b*oundary

OUTPUT: - learning Bayesian Network

1. *The initialized empty structure and initialize parameters of FOA algorithm (dimension space D_"s " size of population NP, the constant value ofCc, Sc and Fc, Awareness AP and Dive DP probability, the number of iteration number, upper boundary and lower boundary,* $(G^t_{\text{best } i\, t})$.
2. *Set the velocity and position for all Falcon randomly. Comparing each falcon by BDe score function, and find the best in the current position* $(P^t_{\text{best } i\, d})$.
3. *For loop to maximum iteration number*
4. *For loop to size of population*
5. *Generate the random value pAP,pDP. Select a new best position by comparing the BDe score function of each falcon.*
6. *if pAP < AP, update falcon velocity* $(V_t\, d)$ *using equation 7.*
   else if pDP > DP update faicon velocity using Equation 8 .
   else compare the score function of the current and previous one if its better update falcon velocity using Equation 9 otherwise use equation 10.
7. *Update the position X.*
8. *Evaluate BDeu score function of new position* $(X^t_{i,d})$
   a. *If current position* $(X^t_{i,d})$ *is better than the best position* $(P^t_{\text{best } i\, d})$ *then update the best position by* $((P^t_{\text{best } i\, d}) = (X^t_{i,d}))$
   b. *If* $(G^t_{\text{best },i,d})$ *< current position then update the best solution for global by* $(G^t_{\text{best },id} = (X_1, \beta))$
   c. *The best score value and solution are saved.*
   d. *If* $X_{min} \geq X_{max}$*, stop the iteration process, and the results are present. If not, move into Step 5.*
9. *Return the maximum BDe score.*

Fig. 1. FOA for Bayesian Network Structure learning.

Step 4: Make new locations, repositioning the falcon as required. Based on its own and other falcons' experiences, the falcon examines the pAP to the aware probability AP, and if the attention probability AP is greater than the pAP, the falcon avoids pursuing preys:

$$\text{Xiter} + 1 = \text{Xiter} + \text{Viter} + \text{cc}\,(\text{Xbest}, \text{Xiter}) + \text{sc}\,(\text{gbest}, \text{Xiter}) \quad (7)$$

where Viter represents existing velocity and Xiter represents the falcon's current location.

Compare the probability of dive DP with pDP if AP is smaller than pAP. If DP is smaller than pDP, the falcon (Xchosen) selects one of the targets as prey and completes the first phase in the hunting process. The logarithmic spiral is calculated as follows:

$$\text{X}_{\text{iter}+1} = \text{X}_{\text{iter}} + |X_{\text{chosen}} - X_{\text{iter}}| \exp(bt) \cos(2\pi t) \quad (8)$$

where a fixed number is b and takes the position of the logarithmic spiral that matches 1, and number as a random will be t in the range (-1,1) that defines the falcon's next exact location [38].

While AD is more than pAP, compare the preferred prey's score function to the falcon's scoring function, and the falcon will follow the preferred prey wherever it is most suitable, like in a dive step:

$$Xiter + 1 = Xiter + Viter\,1 + fc * rand\,(Xchosen - xiter) \quad (9)$$

The falcon, on the other hand, continues to fly around the optimal position:

$$Xiter + 1 = Xiter + Viter\,1 + cc\ \hat{}\ * rand\,(Xbest, Xiter) \quad (10)$$

Later, new location is assessed in terms of velocities and location bounds. The new scoring function is then computed, as well as the updated values of gbest and Xbest.

Step 5: After that, the assessments from step 4 are repeated until the iterations of maximum number (itermax) is obtained. Fig. 1 shows the falcon optimization method for structure learning Bayesian networks as a proposed technique. Falcon G0, which illustrates a DAG using arcs in Fig. 2, tries addition, move, reversal, and deletion, going to new solutions G1, G2, G3, and G4.G3 will be chosen since it has the greatest score; the falcon will then continue to investigate using a similar strategy to arrive at G+3 as the next alternative. If the G+3 BDeu score is higher than the G+1 BDeu score, the G+3 BDeu score is used, the falcon will do a similar operation. The methods iterate while the score of BDeu is fix or repetition loop reaches its maximum length. During the whole operation, the falcon must pick between deletion, movement, reversal, and addition.

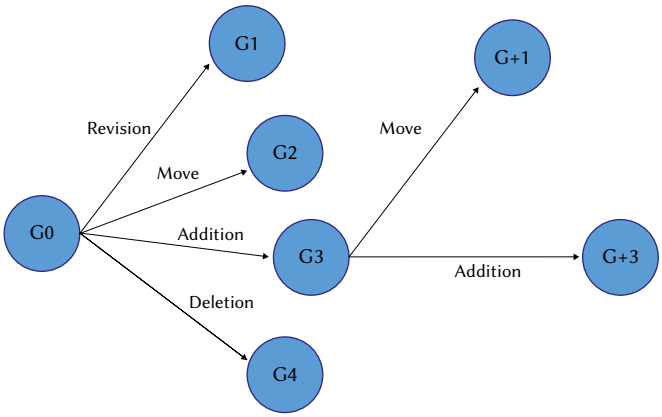


Fig. 2. Searching steps for one Falcon [12].

TABLE I. Score Function Calculation for PIO, ALO, FOA, and Greedy Within Execution Times of 2, 5, and 60 Minutes

	FOA	PIO	ALO	Greedy	FOA	PIO	ALO	Greedy	FOA	PIO	ALO	Greedy
Dataset	2-Min	2-Min	2-Min	2-Min	5-Min	5-Min	5-Min	5-Min	60-Min	60-Min	60-Minu	60-Minu
Hepatitis	-1016.3403	-1327.73	-1326.58	-1350.16	-1011.875	-1327.73	-1327.73	-1350.16	-1010.015	-1327.73	-1327.7	-1350.16
Parkinsons	-1598.9078	-1486-86	-1488.52	-1732.76	-1598.9078	-1439.09	-1441.27	-1721.16	-1598.9078	-1439.09	-1442.87	-1700.36
Imports	-1773.194	-1811.99	-1811.99	-1994.15	-1768.8996	-1811.99	-1811.99	-2012.21	-1755.3093	-1811.99	-1811.25	-1995.76
Heart	-2348.9413	-2426.8	-2424.49	-2576.93	-2335.4483	-2423.8	-2424.81	-2560.43	-2325.483	-2423.8	-2422.57	-2527.44
Mashroom	-3345.924	-3160.87	-3162.28	-3734.22	-3345.924	-3160.87	-3162.45	-3706.66	-3000.9687	-3160-87	-3019.91	-3588.69
WDBC	-6668.5114	-6660.43	-6658.43	-8089.41	-6603.9566	-6660.43	-6662.24	-7954.65	-6574.2007	-6660.43	-6662.25	-7841.35
win95pts	-45978.549	-46779.5	-46772.8	-83749.3	-43850.275	-46779.5	-46779.5	-83150.7	-39814.781	-46779.5	-46780	-81779.5
Sensors	-60343.344	-60710.3	-60341.9	-69200.3	-59895.45	-60710.3	-60343.3	-69150	-58291.874	-60710.3	-60343.3	-68364
Hepar	-160095	-160095	-160095	-169497	-160082	-160095	-160095	-169881	160055	-160095	-160095	-168871
Letter	-173090.07	-175200	-175185	-184307	-173090.07	-175200	-175200	-184916	-173090.07	-175200	-175200	-184118
Epigenetics	-177511.65	-176636	-176641	-225346	-176451.98	-176636	176637	-224172	-176235.58	-176636	-176642	-217246
Adult	-20598.489	-207809	-207805	-211844	-20551.489	-207809	-48572	-211781	-20535.927	-207809	-207457	-211762

IV. Experimental Evaluation

A common assessment approach is used to evaluate the performance of FOA, which employs probabilistic datasets collected from prominent Bayesian networks benchmarks. A PC with the following characteristics serves as the experimental platform: The method is implemented in Java and runs on a 4GB RAM, 2.1GHz CPU, Core i3, operating system (Ubuntu 14.04). We looked examined the suggested algorithm's characteristics in a number of static datasets, including: Asia (8 variables, 8 arcs, and 3000 instance), Static Banjo (33 variables and 320 instance), Letter (17 variables and 20000 instance), Heart(22 variables and 267 instance), Epigenetics (30 variable and 72228 instance), Alarm (37 variables, 46 arcs, and 10000 instance), Hailfinder (56 variables, 66 arcs, and 2656 instance), WDBC (9 variables and 1000 instance), Hepar (70 variables, 123 arcs, and 350 instance), Water (32 variables, 66 arcs, and 10083 instance), Child (20 variables, 25 arcs, and 230 instance), Imports(22 variables and 205 instance), Sensors(25 variables and 5456 instance), Insurance (27 variables, 52 arcs, and 3000 instance), win95pts (76 variables, 112 arcs, and 574 instance), Andes (223 variables, 338 arcs, and 500 instance), Hepatitis(35 variables and 137 instance), Soybean (35 variables and 307 instance), Lucas01(10 variables and 10000 instance), Adult (16 variables and 30162 instance), Parkinsons (23 variables and 195 instance), Mushroom (23 variables, 1000 instance), and Lucap02 (143 variables and 10000 instance) [43].

The learning datasets we looked at stationary sets, and this study is built on the assumption of stationary data. Extending the FOA technique to Andes and sensor benchmarks or other types of stream data sets in online is a hard task that may attempted after a thorough evaluation of its effectiveness on stationary data sets.

The authors used relevant metrics for the datasets to compare the outcomes with Pigeon optimization algorithm (PIO), Greedy Search (GS), and Antlion optimization algorithm (ALO). We assessed all techniques under the identical settings after determining the parameters of the FOA algorithm. For the experiments in the FOA, the following values were used: N is the population size, AP= 0.3, and tmax is 1,000. Sc = 3, Cc = 2, Fc = 4, (t) is a random value within the range of [-1,1], Vmax = 0.1 ub (ub is 100, and Vmax is 10), and DP is 0.85 are fixed value of the FOA's optimization. Pigeon Inspired algorithms have parameters such as dimension space D, population size Np, factor R for map and compass, number of iterations Nc1 max and Nc2 max for two operators, and Nc2 max > Nc1 max.The Antlion optimization algorithm parameters are: dimension space D, population size NE, number of iterations, upper and lower boundaries (Xmax and Xmin), and Xmax > Xmin..The algorithms were implemented in three distinct time frames: two minutes, five minutes, and 60 minutes.

Table I displays the scores for those algorithms which is known in this paper in the specified datasets, as well as time values. In all circumstances, the recommended approach outperforms the default Greedy Search, Antlion Algorithms, as indicated in the table. This illustrate that FOA obtain the best score in the quickest time possible.

Confusion matrix implemented for all data sets and network structure to assess the efficacy of structure identification. The metrics FN, TP, TN, and FP, the criteria Sensitivity (SE), F1 Score, Average Hamming Distance (AHD), and Accuracy (Acc), have been computed for each network per method.

$$\text{sensitivity} = \frac{TP}{TP + FN} \tag{11}$$

$$\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{12}$$

$$\text{F1 Score} = \frac{2 * TP}{2TP + FP + FN} \tag{13}$$

$$\text{AHD} = \frac{FN + FP}{TP + TN + FP + FN} \tag{14}$$

Defining these metrics illustrated as: A TP is a learning network arc (vertex or edge) that is located in the correct location. The arc that travels through neither the learning nor the regular networks is known as TN. The arc of the learning network, not the arc of a regular network, is FP. The FN is the arc in a conventional network, but not in a learning network. PIO, FOA, ALO, and Greedy Sensitivity Results, are illustrate in Fig. 3. The FOA produces best values than the PIO, Antlion, and Greedy search in diffrent datasets.

As demonstrated in Fig. 4, the suggested technique has higher accuracy values in the most dataset than the PIO, ALO, and Greedy methods. The suggested FOA Learning Algorithm is effective in determining the correct structure. As a consequence, in most datasets, the Iterative FOA method outperforms other algorithms in terms of prediction accuracy, and the FOA also outperforms other algorithms findings, we utilized F1 as a metric of the model's accuracy for performance metrics.

The Falcon optimization algorithm's performance is evaluated using the Precision, Recall, F1-score. In these cases, Precision is the The number of total network edges in anticipated BN splited with the number of successfully identified directed edges. Recall is achived by dividing the directed edges number identified by total number for edges in the BN. It recognizes that the harmonic average of accuracy and recall is F1. The scenario is depicted in Fig. 5. FOA, ALO, PIO, and Greedy searches are compared. As demonstrated in Fig. 5, the suggested approaches outperform the ALO, Greedy search, and PIO methods. Furthermore, accuracy is an important criterion for measuring model performance since the model's ultimate purpose is to provide a

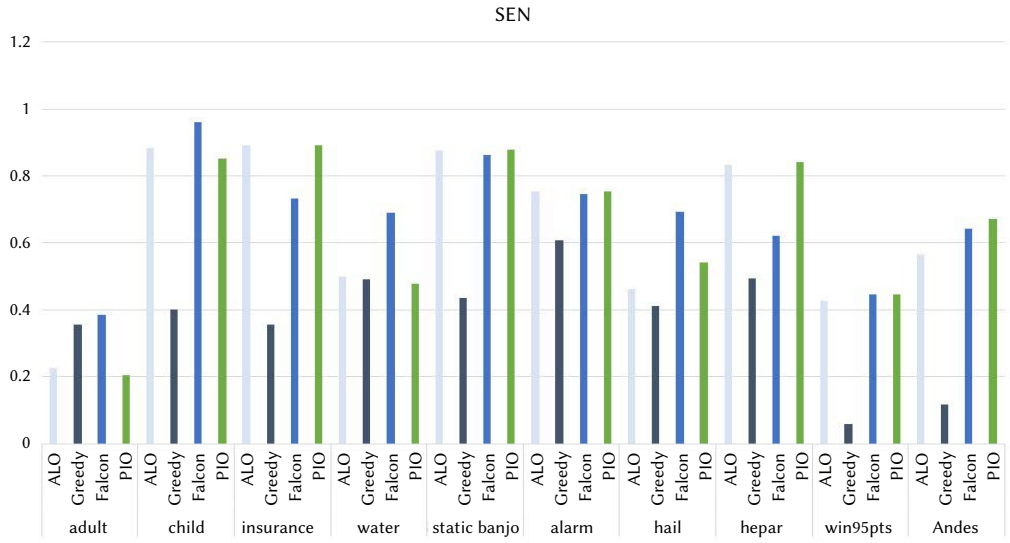


Fig. 3. Sensitivity for SA, GS, FOA, and PIO.

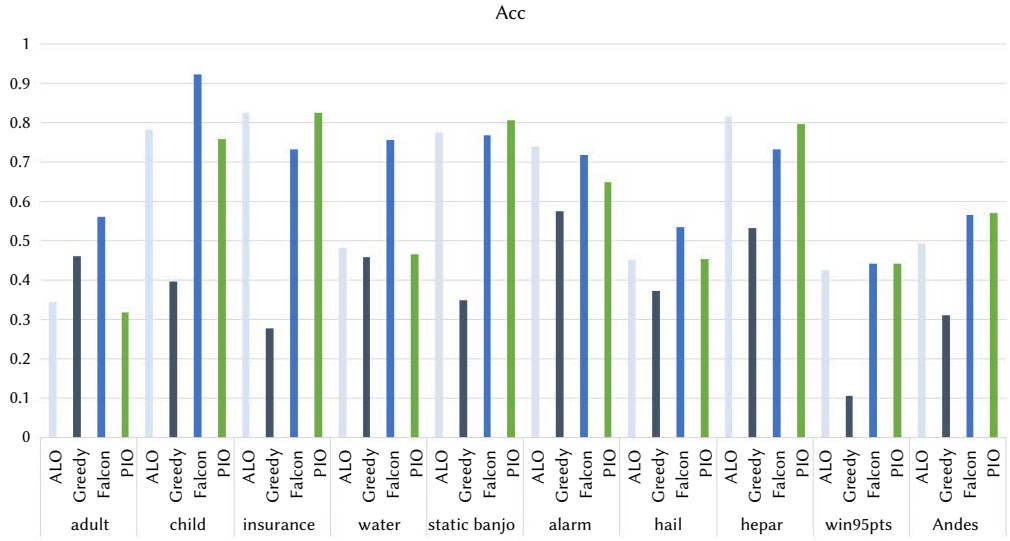


Fig. 4. Accuracy for SA, GS, FOA, and PIO.

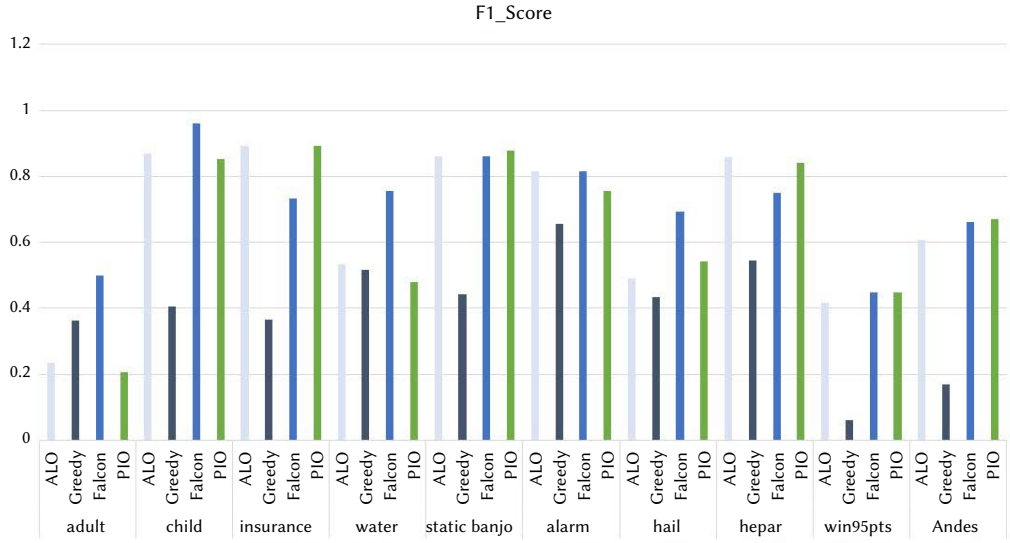


Fig. 5. F1_Score for SA, GS, FOA, and PIO.

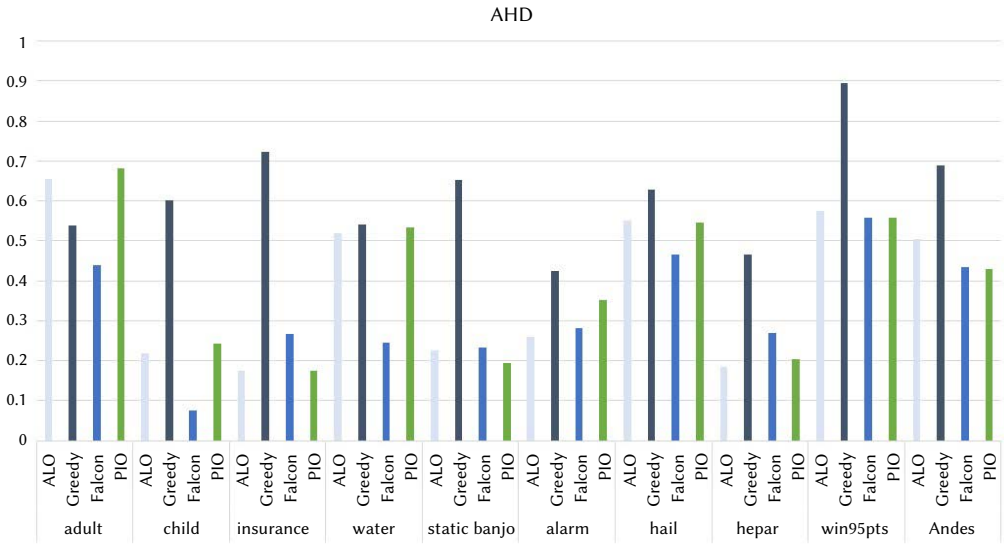


Fig. 6 AHD for SA, GS, FOA, and PIO.

useable illustration of the real world. In terms of Hamming distances, the proposed approach beats the DAG space algorithm, which is always much smaller. Because local networks are entirely focused on exploration rather than inference, main assessment measures for BN structure learning is hamming distances because it directly suits the structure of learners. The Average Hamming Distances for the methods presented are shown in Figure 6. The findings show that the proposed strategy delivers higher performance values than the other strategies we looked at.

V. Conclusion

The authors have focused on Bayesian network structure learning and used Falcon Inspired Optimization method to tackle the problem. We employed the search and score strategy using the FOA algorithm as search function and BDeu as the scoring function. FOA is a stochastic optimization technique based on falcon navigational behavior.

FOA is a method for locating a discrete solution search space that may be applied to any task. The falcon can employ FOA to lead a logarithmic spiral to the lowest usable solution space, which allows for quicker concentration to the global extremum. The proposed technique has a greater search capability, which implies it can find better structure solutions, calculate THE VALUE score function, and properly measure network structure. The strategies help to speed up global convergence and improve global search efficiency. We want to investigate other important aspects of the FOA, such as efficiency, resource use, and run time analytics, BY using THE BEST data sets and experimental configurations.

References

[1] Ji, J., Wei, H. & Liu, C., "An Artificial Bee Colony Algorithm For Learning Bayesian Networks," Springer-Verlag Berlin Heidelberg, 2012.

[2] N. Fortier, J. Sheppard and K. G. Pillai, "Abductive Inference Using Overlapping Swarm Intelligence," In IEEE Symposium On Swarm Intelligence, 2013.

[3] H. Q. Awla, S. W. Kareem, A. S. Mohammed, "Bayesian Network Strucutre Discovery Using Antlion Optimization Algorithm," International Journal Of Systmatic Inovation, 2022

[4] Sun X., Chen C., Wang L., Kang H., Shen Y., Chen Q., "A Hybrid Optimization Algorithm For Bayesian Network Structure Learning Based On Database," Journal Of Computers, VOL. 9, 2014.

[5] D. Margaritis, "Learning Bayesian Network Model Structure From Data," Carnegie-Mellon University, Pittsburgh, Pa. Available As Technical Report Cmu-, 2003.

[6] A. S. Fast, "Learning The Structure Of Bayesian Networks With Constraint Satisfaction, Massachusetts," Phd Thesis , Department Of Computer Science, University Of Massachusetts, February 2010, 2010.

[7] A. S. Hesar, "Structure Learning Of Bayesian Belief Networks Using Simulated Annealing Algorithm," Middle-East Journal Of Scientific Research 18, Pp. 1343-1348, 2013.

[8] G. F. Cooper And E. Herskovits, "A Bayesian Method For The Induction Of Probabilistic Networks From Data," In Machine Learning, 9, 1992.

[9] J. Cowie, L. Oteniya, R. Coles , "Particle Swarm Optimisation For Learning Bayesian Networks," Engineering And Physical Sciences Research Council., 2007.

[10] Salama, K.M., Freitas, A.A. (2012), "ABC-Miner: An Ant-Based Bayesian Classification Algorithm," Lecture Notes in Computer Science, vol 7461. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-32650-9_2

[11] C. -C. He and X. -G. Gao, "Structure Learning Of Bayesian Networks Based On The LARS-MMPC Ordering Search Method," *2018 37th Chinese Control Conference (CCC)*, 2018, pp. 9000-9006, doi: 10.23919/ChiCC.2018.8483049

[12] S. Li and B. Wang, "A Method For Hybrid Bayesian Network Structure Learning From Massive Data Using Mapreduce," *2017 ieee 3rd international conference on big data security on cloud (bigdatasecurity), ieee international conference on high performance and smart computing (hpsc), and ieee international conference on intelligent data and security (ids)*, 2017, pp. 272-276, doi: 10.1109/BigDataSecurity.2017.42.

[13] Dai, J., Ren, J., Du, W. Et Al., "An Improved Evolutionary Approach-Based Hybrid Algorithm For Bayesian Network Structure Learning In Dynamic Constrained Search Space," Neural Computing & Applications, Pp. 1413-1434, 32 2020.

[14] Sun, X.; Chen, C.; Wang, L.; Kang, H.; Shen, Y.; Chen, Q., "Hybrid Optimization Algorithm For Bayesian Network Structure Learning," Information 10, No. 10: 294. P. Https://Doi.Org/10.3390/Info10100294, 294 10 2019.

[15] K. Shahab, And C. Mehmet, "Bayesian Network Structure Learning Using Hybrid Bee Optimization And Greedy Search," Adana / TURKEY, 2018.

[16] A. C. Constantinou, "Learning Bayesian Networks With The Saiyan Algorithm," Acm Trans. Knowl. Discov. Data, P. Doi:Https://Doi.Org/10.1145/3385655, 4 14 2020.

[17] Cuicui Yang , Junzhong Ji , Jiming Liu , Jinduo Liu And Baocai Yin, "Structural Learning Of Bayesian Networks By Bacterial Foraging Optimization,", International Journal Of Approximate Reasoning, P. 69, 2016.

[18] P. Larraiiaga , M. Poza, "Structure Learning Of Bayesian Networks By Genetic Algorithms," Springer-Verlag Berlin Heidelberg Gmbh, 1996.

[19] Kalia Orphanou, Dirk Thierens, And Peter A. N. Bosman, "Learning Bayesian Network Structures With GOMEA," Kyoto, Japan, 2018.
[20] Kareem, S. W., & Okur, M. C. (2021)., "A Novel Method Forbayesian Networks Structure Learning Based On Breeding Swarm Algorithm," Soft Computing, Nr. 9, 9 2018.
[21] Jingyunwang, And Sanyangliu, "Novel Binary Encoding Water Cycle Algorithm For Solving Bayesian Network Structures Learning Problem," Knowledge-Based Systems, Bd. 150, 2018.
[22] Kareem, S.W., & Okur, M.C. (2019), "Bayesian Network Structure Learning Based On Pigeon Inspired Optimization," International Journal Of Advanced Trends In Computer Science And Engineering, Bd. 8, Nr. 1.2, Pp. 131-137, 2019.
[23] Fan, X., Yuan, C., & Malone, B.M. (2014), "Tightening Bounds For Bayesian Network Structure Learning," In Association For The Advancement Of Artificial Intelligence, 2014.
[24] Changheyuan, B. And Xiaojian W., "Learning Optimal Bayesian Networks Using A* Search," In NSF Grants IIS-0953723 And EPS-0903787, 21 IJCAI. , Barcelona , 2011.
[25] Patrick O. Djan-Sampson And Ferat S., "Structural Learning; Of Bayesian Networks From Complete Data Using The Scatter Search Documents," In IEEE International Conference On Systems, Man And Cybernetics, 2004.
[26] M. B. A. Askari and M. G. Ahsaee, "Bayesian Network Structure Learning Based On Cuckoo Search Algorithm," In 2018 6th Iranian Joint Congress On Fuzzy And Intelligent Systems (CFIS), Kerman, Iran, 2018.
[27] Thibaud R., Sylvain M., Stephane G., Florence F., "Fast Bayesian Network Structure Learning Using Quasi-Determinism Screening," HAL, Bd. 2, Pp. 14-24, 2019.
[28] Safiye S., Ercan O., Harun T. And Orhan T., "Bayesian Structural Learning With Minimum Spanning Tree Algorithm," In The World Congress In Computer Science, Computer Engineering And Applied Computing, 2013.
[29] J. Cowie, L. Oteniya, R. Coles, "Particle Swarm Optimisation For Learning Bayesian Networks," In Engineering And Physical Sciences Research Council, 2007.
[30] L. M. D. Campos, "A Scoring Function For Learning Bayesian Networks Based On Mutual Information And Conditional Independence Tests," Journal Of Machine Learning Research, No. 7, Pp. 2149-2187, 2006.
[31] Zhang S., Liu L., "Mcmc Samples Selecting For Online Bayesian Network Structure Learning." Kunming,, 2008.
[32] D. Heckerman, D. Geiger, And D. M. Chickering,, "Learning Bayesian Networks: The Combination Of Knowledge And Statistical Data," Machine Learning,, Pp. 197-243, 1995.
[33] Thippa Reddy, G., & Khare, N., "FFBAT-Optimized Rule Based Fuzzy Logic Classifier For Diabetes," International Journal Of Engineering Research In Africa Trans Tech Publications, Pp. 137-152, 2016.
[34] Gadekallu, T. R., & Khare, N., "Cuckoo Search Optimized Reduction And Fuzzy Logic Classifier For Heart Disease And Diabetes Prediction," International Journal Of Fuzzy System Applications (IJFSA), Pp. 25-42, 2017.
[35] Ding, Q., & Xu, X., "Improved GWO Algorithm for UAV Path Planning on Crop Pest Monitoring," International Journal Of Interactive Multimedia And Artificial Intelligence, Vol. 7(Special Issue on Multimedia Streaming and Processing in Internet of Things with Edge Intelligence), 30-39, 2022.
[36] Seyedali M., Seyed M. M., Andrew L., "A Grey Wolf Optimizer," Advances In Engineering Software, Bd. 69, Pp. 46-61, 2014.
[37] Verma, K.K.; Singh, B.M., "Deep Multi-Model Fusion for Human Activity Recognition Using Evolutionary Algorithms," International Journal Interactive Multimedia Artificial Intelligence 2021, 7, 44–58.
[38] Segundo H. De V., Cocco V., Dos Santos Coelho L., "Design Of Heat Exchangers Using Falcon Optimization Algorithm," Applied Thermal Engineering, Pp. 119-144, 2019.
[39] V. Tucker, "Gliding Flight: Drag And Torque Of A Hawk And A Falcon With Straight And Turned Heads, And A Lower Value For The Parasite Drag Coefficient," Journal Experimental Biology 203, P. 3733–3744., 2000.
[40] V. Tucker, "Gliding Flight: Speed And Acceleration Of Ideal Falcons During Diving And Pull Out," Journal Experimental Biology 201, P. 403–414., 1998.
[41] Anders H. Mikael R., Susanne Å., And Fernando S., "Flight Performance During Hunting Excursions In Eleonora's Falcon Falco Eleonorae," The Journal Of Experimental Biology, P. 2029–2039, 1999.
[42] D. Dekker, "Hunting Tactics Of Peregrines And Other Falcons, Wageningen," The Netherlands Wageningen University, 2009.
[43] R. Nagarajan, M. Scutari And S. Lèbre, "Bayesian Networks In R With Applications In Systems Biology," New York: Springer, 2013.

Shahab Wahhab Kareem

Shahab Wahhab Kareem his BSc in Control and Computer Engineering from the University of Technology Baghdad in 2001, MSc in Software Engineering from Salahadeen University in 2009, and Ph.D. in Yasar University Izmir, Turkey in 2020. My research interests include Machine learning and BIG DATA. I'm a lecturer at the Information System Eng. (ISE) Department (2011-till now). He can be contacted at email: shahab.kareem@epu.edu.iq or shahabwk@yahoo.com.

Amin Salih Mohammed

Dr. Amin Salih Mohammed is the President of the Lebanese French University in Erbil, Iraq's Kurdistan region. He was the university's Vice President for Scientific Affairs before being promoted to President. He has more than 15 years of expertise in both teaching and research. Amin is a dedicated scholar who has served as a resource for a variety of seminars and faculty development programs hosted by various schools. He earned his undergraduate, postgraduate, and doctoral degrees in computer engineering from the Kharkiv National University of Radio Electronics in Kharkiv, Ukraine. Computer networks, wireless networks, and cloud computing are among his research interests.

Hoshang Qasim Awla

Hoshang Qasim received his BSc in Computer Science/ Soran University /Iraq in 2012 and his MSc in Computer Engineering from Hassan Kalyouncu University/ Turkey in 2016. and a Ph.D student in data science from Soran University. Assistant lecturer at the computer science department/ faculty of science at Soran University, email address: hoshang.awla@soran.edu.iq

Multi-Agent and Fuzzy Inference-Based Framework for Traffic Light Optimization

Abdelouafi Ikidid*, Abdelaziz El Fazziki, Mohammed Sadgal

Cadi Ayyad University, Marrakesh (Morocco)

Received 30 March 2021 | Accepted 30 September 2021 | Published 16 December 2021

Abstract

Despite the fact that agent technologies have widely gained popularity in distributed systems, their potential for advanced management of vehicle traffic has not been sufficiently explored. This paper presents a traffic simulation framework based on agent technology and fuzzy logic. The objective of this framework is to act on the phase layouts represented by its sequences and length to maximize throughput and fluidize traffic at an isolated intersection and for the whole multi-intersection network, through both inter- and intra-intersection collaboration and coordination. The optimizing of signal layouts is done in real time, and it is not only based on local stream factors but also on traffic stream conditions in surrounding intersections. The system profits from agent communication and collaboration as well as coordination features, along with decentralized organization, to decompose the traffic control optimization into subproblems and enable the distributed resolution. Thus, the separate parts can be resolved rapidly by parallel tasking. It also uses fuzzy technology to handle the uncertainty of traffic conditions. An instance of the proposed framework was validated and designed in the ANYLOGIC simulator. Instantiation results and analysis denote that the designed system can significantly develop the efficiency at an individual intersection as well as in the multi-intersection network. It reduces the average travel delay and the time spent in the network compared to multi-agent-based adaptative signal control systems.

Keywords

Agent Technology, Coordination, Communication, Fuzzy Logic, Traffic Signal Regulation.

DOI: 10.9781/ijimai.2021.12.002

I. Introduction

THE optimization of signal light control in urban areas is at the forefront of research in the field of Artificial Transportation Systems (ATS). ATS can be implemented by different approaches and technologies. The widely used artificial intelligence techniques for optimizing traffic signals are the Genetic Algorithm, Artificial Neural Network, Fuzzy logic, Multi-Agent System (MAS), Case-Based Reasoning, and Reinforcement Learning (RL). In this paper, we combine agent technology and fuzzy logic to design a cooperative real-time traffic signal optimization system, where the signal control plan is frequently updated to meet the non-stationary traffic state. Agent technologies have been widely accepted as one of the most responsive tools to deal with a wide-reaching distributed system. That's why agent-based systems are well suited for the traffic and transportation domain, since these systems are geographically distributed in a non-stationary environment [1]. Agents can use perceptive data and received information from other agents to achieve their goals. Each agent can cooperate with neighboring agents and adjust his reactions in real time to his surroundings as they change. Therefore, multi-agent technology treats a complicated system in a distributed manner; it splits the complex control system into simple subtasks, therefore allowing parallel and fast decision-making [2].

* Corresponding author.
E-mail address: a.ikidid@gmail.com

With this being considered, the Multi-agent Cooperative Traffic Signal Optimization (MCTSO) is proposed to maximize the signalized intersection throughput and reduce congestion in urban arteries with three contributions: (1) the real-time optimization is introduced to adapt the system in a timely way to the continuously changing conditions and disturbances, supported by online monitoring of the optimum indicators to detect congestion and maintain the system not far off from the suitable operating point as much as possible. (2) Two-stage coordination, including intra-junction coordination, which is enabled to prioritize the higher congested stream, and inter-junction coordination, which is used to generate a fluidized scenario downstream of the congested stream, is used. (3) Distributed collaboration control, splitting the network into sub-areas whose control is easier, is used to allow parallel-tasking. Therefore, the functionality of an MAS will not reside in the agents themselves, but will be ubiquitously distributed to allow the system to perform tasks in parallel, avoiding an additional computational cost [3].

In this article, we propose a distributed and adaptative, as well as online, optimized traffic signal control scheme enabled by a decentralized multi-agent system, where each group of agents represents a signalized intersection control unit, each group coordinates and collaborates with adjacent surrounding groups, and each group achieves local optimization, taking into consideration global network optimization. We use an artificial fuzzy logic algorithm to tackle the fuzzy condition of the road environment. Our proposed MCTSO differs from existing approaches due to agent specialization. The group contains specialist agents for each role, and it is designed and adapted to a specific task, which allows us to improve the agent

efficiency and make its role more accurate. Additionally, in MCTSO the combination of two-stages of coordination and collaboration aims to develop a clearer view of the environment, make decisions of common benefit, and avoid local optimization.

The rest of the paper is organized as follows: The second section analyzes and discusses the related works about intelligent traffic signal control and artificial intelligence techniques. The third section gives a global overview of the traffic control problem. The fourth section details the proposed multi-agent system. The fifth section provides the detailed results of the simulation tests carried out on the AnyLogic platform. Finally, we conclude in the sixth section.

II. Related Works

The MAS is rapidly growing as one of the most powerful popular technologies proposed to solve complicated problems in different fields, such as electrical engineering, computer science, electronic commerce, civil engineering, and transportation systems.

In a transportation system and with the diversity of actors involved, agent technology can be used in the various components of the system, such as drivers and vehicles [4] [5], traffic light [6], phases [7], and to handle diverse aspects, e.g., congestion [8], the green transportation system [9], and route guidance[10]. In urban traffic networks, signalized intersections are one of the most important and influential ingredients, and the traffic signal is the most utilized instrument for scheduling and managing traffic flow. In what follows we analyze and discuss succinctly several studies that use a multi-agent system and artificial intelligence techniques to perform intelligent traffic signal control.

Regarding the architecture of multi-agent-based signal control, most approaches usually divided the road network into regions or sub-parts that cover one or more intersections. These sub-parts are controlled by a cell of one or more agents. The organizational structure of agents can be modeled in various ways. The organizational structure determines the interactions, roles, and structures of the agent's community. It can be designed in many forms, such as flat, hierarchical, holonic, teams, and federation [11].

Many studies have reported using a hierarchical scheme to manage the traffic signal. Jin and Ma [12] use reinforcement learning to introduce a hierarchical multi-agent-based control scheme. The agents are categorized as the region agent (RA), intersection agent (IA), and turning movement agent (TA), listed in the order of the hierarchy. Communication and cooperation between agents at equal levels are elevated through the decentralized representation of the framework. Nevertheless, agents at the lower level have to reach an accord between their own goals and those given by the agents on the next level up. Like Jin and Ma, Xu et al. [13] introduced a three-layer optimizing control system that includes intersection controller agents (ICAs), sub-zone controller agents (SZCAs), and network controller agents (NCAs), which represent the lowest, middle, and highest layers, respectively. The interaction takes place across all levels to optimize the signal timing strategy, while coordination is granted by the SZA. Nonetheless, besides the overcharge data at higher levels, the focal decision process might produce a bottleneck in these levels, lengthen the response time, and limit the system's scalability. Flat [14] and holonic [15] structures are also proposed for multi-agent-based traffic signal control. Otherwise, it is widely recognized that there is no specific operating multi-agent architecture that is absolute for all traffic signal control systems; additionally, various operating models can be combined.

Pre-timed signal control cannot adapt to the non-stationary traffic state. It has been a while since interactive system control became a trend in traffic management. The first appearance of adaptative traffic control was in the last decade of the second millennium, with the release of the cycle and offset optimization technique (SCOOT) in the 1980s, the Sydney cooperative adaptive traffic system (SCATS), and the green link determining (GLIDE) system. Thereafter, these adaptative control systems were implemented in many countries to manage traffic control in metropolitan areas, and others have been developed (for a review of the self-adaptive traffic signal control, see [16]).

Recently, more focus has been placed on multi-agent-based systems for urban traffic management [17]. It has been proposed that several transport system problems be solved in a distributed manner. However, disturbance management requires particular abilities that a MAS cannot guarantee alone. Consequently, to create intelligent traffic signal controllers, a MAS integrates various intelligent techniques. For example, many models combined the multi-agent approach with the RL approach to optimize a signal timing plan [18]. The agents employ their ability to communicate with the environment to learn and optimize their decision-making behavior. Foremost among the model-free RL methods, Q-learning (QL) is the model most used by researchers using multi-agent reinforcement learning (MARL) in intelligent traffic light control. A work [19] uses fuzzy Q-learning and agent technologies to develop a traffic lights control framework. Each agent interacts with neighbor agents by getting a reward from each decision. . The control decision is made by using the number of vehicles input to schedule green phase duration. The aim is to maximize the reward and decrease average delay time. El-Tantawy et al. [20] improve the travel time and overall delay using QL and a decentralized junction-based model. The model-free RL can be implemented when dealing with a non-deterministic model of the environment, as it does not require pre-assignment of the environment.

Concurrently, some researchers investigated the potential of fuzzy-logic-based control, which has a rule-based inference system and is based on human reasoning. FL is suitable for handling the control of a single intersection [21] characterized by uncertainty, fuzzy circumstances, inexact data, and typically controlled by rules. Because the MAS has a restricted capability to deal with fuzzy circumstances, the incorporation of an MAS and fuzzy inference can show considerable effectiveness in enhancing signal settings in traffic light control [22] [23].

In these studies, the cooperation mechanism is limited at the inter-junction level, which reduces the local control efficiency in favor of global control. Also, the concentration of fuzzy logic in one level creates an overload at fuzzy components. Our proposed multi-agent control system is a model based on the two levels of coordination and collaboration, local at the intersection and within the surrounding neighbors. Each intersection is represented by a controller group in which the decision is made via two levels of fuzzy logic and coordination with adjacent group controllers.

III. Traffic Control Problem Description

According to the US Census Bureau, metropolitan areas will contain 6.7 billion people [24]. This growing urbanization increases the traffic road demand because of a high number of vehicles seeking to use the road infrastructure. Road traffic in urban areas is a nested phenomenon, on the one hand because of the many contributors that act autonomously and on the other hand because of the uncertainty of the road network. When the number of vehicles on an infrastructure exceeds its capacity, traffic congestion occurs, resulting in slow movements and queues that stretch over time. Therefore, the congestion is a parallel evolutionary anomaly, in both space and time. Consequently, to inexpensively mitigate this anomaly, we can optimize

traffic space occupation with an acceptable delay. Signal control is the basic method and an effective one to alleviate congestion as well as to fluidize traffic at the intersection [25]. Optimized signal control can significatively increase infrastructure capacity and reduce travel time [26]. Additionally, it helps to reduce fuel consumption and the emission of air pollutants and improves the health of citizens, too [27].

A. Signalized Intersection Features

A road intersection is a crossing of several roads that contains three functional zones (Fig. 1) managed by a tricolor traffic light; the red queues the vehicles in a storage area, the green gives access to the exit zone through the conflict zone, and the yellow is a transition period from green to red to allow the vehicles to evacuate the conflict zone.

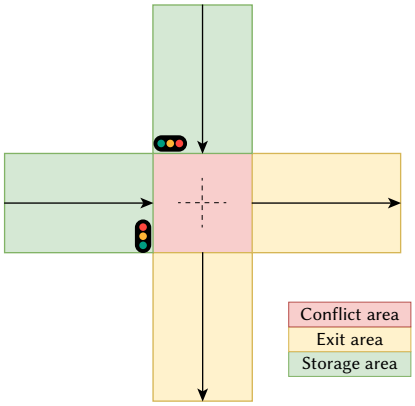


Fig. 1. Functional areas of an intersection of two one-way roads.

B. Intersection Network Modeling

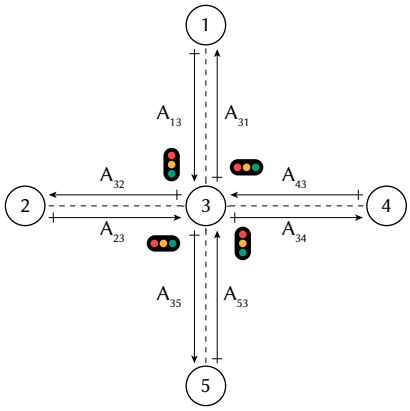
The intersection network is viewed as a disturbed system that is modeled by a strongly connected oriented graph $G = (C, A)$, where (C) is a set of nodes that represent the intersections and (A) is a set of arcs that represent the links that connect these intersections. Each intersection, as a component of the disturbed system, has its own requirements; therefore, it coordinates with its adjacent intersection. Two intersections connected by an arc are considered adjacent. Adjacent intersections cooperate and share their data to achieve a common goal of the system, which is to optimize traffic flow management.

Each arc A_{ij} is bounded by two intersections: i, the initial junction and the arc flow origin, and j, the terminal junction and the arc flow destination. Downstream of an arc is the group of succeeding arcs $succ(A_{ij}) = \{A_{jk}, (i, j, k) \in C\}$, where the outflows of the arc can be routed. Upstream of an arc is the set of predecessor arcs $pred(A_{ij}) = \{A_{ki}, (i, j, k) \in C\}$ where the inflows of the arc arrive.

An arc is characterized by static information, such as the storage area length and max capacity, and dynamic information, namely the state of the traffic signal at the stop line of the arc (green or red). The concentration T (1) at a given segment is the number of vehicles N (in private vehicle units, PVUs) using this segment at a time t, relative to the segment length.

$$T_{\Delta x}(t) = T(p \rightarrow p + \Delta p, t) = \frac{N}{\Delta p} \quad (1)$$

An intersection has a set of incoming arcs $E(i) = \{A_{ji}, (i, j) \in C\}$ and a set of outgoing arcs $S(i) = \{A_{ji}, (i, j) \in C\}$. All intersections are controlled by a signal light, concerning the colors to be used and also their succession or order of appearance. A green phase is a lap of time during which a group of compatible arcs is activated, i.e., the arc flows are allowed to cross the intersection. The cycle is the appearance order of all of the phases, and a traffic control strategy is the scheduling method that defines how the phases participate in the cycle and their layout (length and sequence). Fig. 2(a) shows an intersection with 4 incoming and outgoing arcs and 4 adjacent intersections. Fig. 2(b) gives a representation of a cycle, phase, and the arcs that are activated during each phase.



(a)

Phase		Arc							
		A_{13}	A_{31}	A_{23}	A_{32}	A_{43}	A_{34}	A_{53}	A_{35}
Cycle	$\varphi 1$	1	1	0	1	0	1	0	1
	$\varphi 2$	0	1	1	1	0	1	0	1
	$\varphi 3$	0	1	0	1	1	1	0	1
	$\varphi 4$	0	1	0	1	0	1	1	1

(b)

Fig. 2. Representation of activated and not-activated arcs in a 4-phase traffic light intersection. (a) Intersection with 4 incoming arcs and 4 outgoing arcs; (b) truth table of 4-phases and the arc cycle for each phase.

An intersection is considered congested if it does not manage to evacuate all of the storage areas of the activated arcs after a green phase time; in other words, it is considered congested if the stop time of an incoming arc exceeds the cycle time duration.

IV. Agent Modeling

The organizational design of the urban traffic-responsive control system (UTCS) is spatially and functionally distributed. Each intersection is viewed as a network sub-section and controlled by a community of autonomous, cooperative, and intelligent agents. Commonly, agents are perceived as analyzing at a level with an abstraction upper than components and objects, which makes a MAS suitable for complex and distributed problems.

The proposed MAS has a decentralized architecture with two levels of collaboration. Each signalized intersection is controlled by an intersection control group (ICG), which defines the signal control strategy. This strategy optimizes phase layouts while it is executed to meet the continuously changing surrounding environment, whereas the control of the whole intersection network is fully distributed and is accomplished through the collective and coordination capability of ICGs. In sum, the system goal is achieved with two levels of coordination, which are the following:

- Inter-junction collaboration: This allows coordination between connecting ICGs.
- Intra-junction collaboration: This allows interactions between the agents belonging to the same ICG.

We build our MAS by applying the concept of the model-driven architecture (MDA) [28] to construct our system. We propose to create

an increasingly detailed system from the abstract to a concrete concept following a process in five stages as follows:

1. Select the organizational structure of the MAS.
2. Analyze the system requirements.
3. Structure the UTCS into groups of agents.
4. Identify agents and roles.
5. Implement the generic structure of an agent-oriented system in the AnyLogic simulator.

A. The Organizational Structure of the MAS

The selection of the organizational structure is a very essential stage in MAS development and has an impact on the succeeding stages. Various specifications drive the definition of the organizational structure, including the environment characteristics, the architecture of the real-world organization, the ability of the MAS to support the computation and coordination complexity of the scenario, and the necessity of respecting the organizational rules and minimizing the complexity of the design.

Our proposed MAS has a decentralized architecture with two levels of collaboration based on the metamodel AALAADIN [29], which is built on three main notions: agent, group, and role. Fig. 3 shows a diagram of this model.

Agent: The agent is defined as an active entity that communicates and plays a specific role inside its group. The metamodel does not pose any constraint on the internal architecture of agents.

Group: The group is an atomic aggregate of agents sharing services with other groups. Each agent belongs to a group; in our case, the concept of belonging to a group is limited to one group.

Role: The role is an abstract representation of an agent's tasks, function, or activities. Each agent can have multiple roles, and each role is accomplished by an agent group.

We define the organizational structure as a decentralized set of group sharing services.

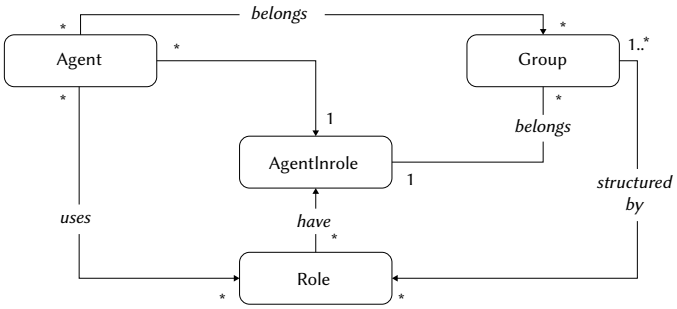


Fig. 3. Organizational structure of the group of agents.

B. Analyzing the System Requirements

The functional architecture of the UTCS includes a set of components. Fig. 4 represents the main components of a regulation system. These components generate an optimized traffic control strategy based on the following scenario:

- The optimization process is initialized after each recurring interval to update the traffic signal control plan.
- The captures are placed at the stop line of the incoming arc to develop a local view of the traffic conditions by observing the storage areas.
- The incoming arcs are monitored to define and update the arc traffic state indicators. These indicators are calculated by observing the local state collected by the captures and considering that of succeeding arcs.
- During the optimizing process, if the degree of saturation in the downstream is intolerable, the upstream indicators are adjusted to slow evacuation and relieve saturation.
- The coordinator provides the traffic condition stat of connecting intersections and shares the local intersection stat.
- The phase managers use the traffic state provided by arc monitors to define the phases' states and request a traffic signal control update.
- The cycle time is prefixed and divided between all intersection phases. Unused lap time will be reallocated to other phases or subtracted from the cycle time.
- The intersection controller updates the control strategy during the progress of the cycle.
- Each arc has a right to green time one and one time only in the cycle, and all the links with at least one queued vehicle at the stop line should have green time.
- The pedestrian phase is outside of the scope of our approach.

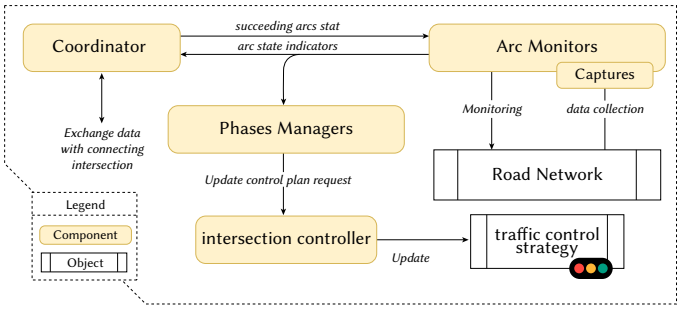


Fig. 4. Components of intersection control.

C. Structure the UTCS Into Groups of Agents

We assume that the structural aspect of a MAS consists of two parts: a classification structure and a role structure (see section D). The classification structure indicates how agent groups are determined and how they interact with each other. This classification is based on the roles of agents and their social interactions. To structure the agent community into groups, we rely on the classic software engineering rule: high cohesion and low coupling. Therefore, agents sharing more roles and goals will be in the same group, and agents that do not share roles, or have few common roles, are placed in two distinct groups.

The representation of a UTCS by a MAS is based on a mapping between the UTCS and MAS that we propose (Table I). Commonly, each intersection is managed by a group control that consists of five main components: an arc monitor for each incoming arc; two phase managers, one for the green phase and the other for the red phase; a coordinator; and an intersection controller.

TABLE I. Alignment of UTCS / MAS

UTCS component	Agent Group
Arc monitor	Intersection Control Group (ICG)
Phase managers	
Coordinator	
intersection controller	

D. Identify Roles and Agents

The MCTSO applied to the signalized intersection network contains a set of ICGs. Each ICG was assigned to an intersection and charged with full control over the incoming streams. Each ICG includes a number of agents classified into 5 types: an ARC agent, which is associated with each incoming arc, two Phase agents (the Active Phase

agent (APA) manages the current green phase and the Inactive Phase agent (IPA) manages other phases), a Coordinator agent (CA), and a Decision agent (DA).

1. ARC Agent

Each incoming arc is managed by an agent. The goal of this kind of agent is to monitor the arc storage area in a timely and continuous manner. Arc agents have only a local view of the environment. To minimize the complexity degree of the system, no agent can have a full overview of the network. They use sensors placed at the stop line that cover the whole storage zone to define the arc state, taking into account the outflow streams. Depending on the signal state at the arc stop line, the arc stat is defined by the urgency indicators when the signal is red, which are the stop ratio (SR) and congestion ratio (CR). They are calculated using data collected from sensors and the congestion ratio from downstream (CRd) received from the Coordinator agent. When the signal is green, the arc stat is defined by extend indicators, which are the CR and CRd.

SR (2) represents the waiting time ratio of vehicles in the storage area and is defined as the duration of elapsed red time since the last switch (t_s) divided by the cycle length (c) minus the total yellow signal length (t_y). *CR* (3) is the number of enqueued vehicles in the arc storage zone over the capacity of the arc.

$$SR = \frac{t_s}{c - ty} \tag{2}$$

$$CR = \frac{T_t}{Tmax} \tag{3}$$

where *Tmax* is the maximum concentration of vehicles in the arc; T_t is the concentration at an instant t; t_s is the vehicle stop time on the red signal; c is the cycle length; and ty is the yellow signal length.

To reduce the phase transitions when there is no traffic, arcs waive their green turn by setting the urgency indicators or extend indicators equal to 0 when

- There are no enqueued vehicles in arcs, since the empty arc does not need green time.
- The CRd is equal to or greater than 1, which means that the concentration downstream surpasses or reaches its maximum capacity, since the congested outgoing arc is not able to get more inflow.

This type of agent will be conscious of all of the other intersection agents. It cooperates with the Coordinator agent to define the arc traffic condition state, with the phase agents to propose the suitable phase layout update and with the Decision agent to implement the optimized control strategy.

2. Phase Agents

A phase is seen as a state machine. This automaton has two states: Active and Not Active (Fig. 5). Depending on the states of a phase, we have adopted two agents to manage all phases in an intersection: the APA, which manages the activated phase, and the IPA, which manages the not-activated phases.

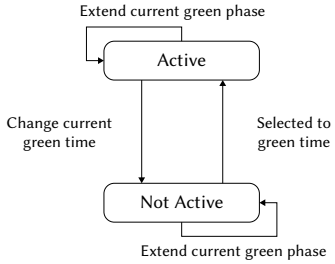


Fig. 5. Phase state.

The goal of the APA is to maximize the green time allocated to the arcs that make up the current green phase, while the IPA's goal is to reduce the stop time of the other phases. The phase agents compete; each agent seeks to extend its active time and otherwise limit the time of the other phase agents.

a) IPA

The IPA controls the phase sequences. It selects a phase from all phases, except the current phase and the already activated phase in the cycle, to be a candidate for the next green time. This agent examined the urgency degree of approved phases by evaluating arc urgency indicators provided by arc agents; a phase is represented by the most urgent arc. The phase urgency degree is obtained by the fuzzy process after the verification of the max/min constraint.

Selection of candidate phase:

The candidate phase is calculated through the collaboration and coordination with the **Arc** and **Coordinator** agents. The **IPA** is the controller of phase scheduling and sequence. It proposes an advisable phase order for the current traffic state. The flowchart of the candidate phase selection process is presented in Fig. 6 and consists of six steps:

Step 1. The **IPA** starts the phase selection process by creating a *collaboration-group* and initializes it to the list of all intersection arc agents classified in the phase set (one arc agent may belong to many phases).

Step 2. The **IPA** sends a request to the *collaboration-group members* to inform them that the phase selection process has been started and orders them to begin calculating the arcs' urgency indicators.

Step 3. Each arc agent of the *collaboration-group* calculates its urgency indicators.

Step 4. The **IPA** receives all responses and calculates the urgency degree of each phase using a *fuzzy selection mechanism*. The highest urgency phase will be selected and suggested to receive green time. The selected phase and its urgency degree value will be sent to the **Decision** agent.

Step 5. If the suggested phase gets the green time, the **IPA** removes it from the *collaboration-group* list; it also removes their arc agent if they did not belong to other phases of the current *collaboration-group members*.

Step 6. The phase selector waits a predetermined time (min. red time) and returns to **step 2** while the *collaboration-group* is not empty.

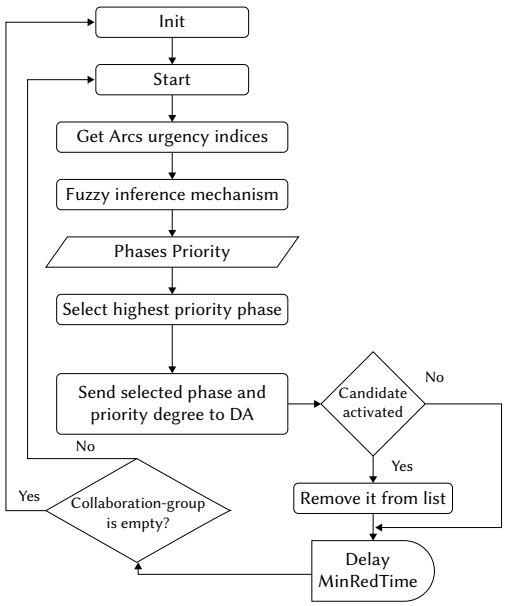


Fig. 6. Flowchart of the candidate phase selection process.

Fuzzy selection mechanism:

The IPA combines the inputs CR, CRd, and SR to create the urgency degree output of the candidate phases. The outputs of the phase agents are used as input in the DA process to make the final decision.

The membership functions SR, CR, and CRd are standardized. According to this, there are four membership functions, including Small (S), Medium (M), Large (L), and Very Large (VL), for these inputs. The linguistic variables as well as the membership functions are shown in Fig. 7. The technique used in all of the defuzzification process is the Center of Gravity (COG) method.

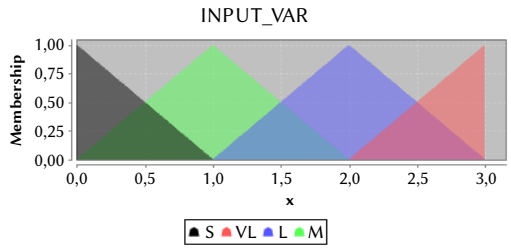


Fig. 7. Membership functions of inputs.

To define the most appropriate phase for green time, a set of rules is defined for the phase selection process. Fig. 8 shows several rules that were used to perform the system simulation. The "AND" operation is performed by using the max t-norm (Łukasiewicz).

```
RULE 11 : IF CR IS M AND CRd IS S AND SR IS S THEN UrgencyDegree IS S;
RULE 12 : IF CR IS M AND CRd IS S AND SR IS M THEN UrgencyDegree IS M;
RULE 13 : IF CR IS M AND CRd IS S AND SR IS L THEN UrgencyDegree IS L;
RULE 14 : IF CR IS M AND CRd IS M AND SR IS S THEN UrgencyDegree IS S;
RULE 15 : IF CR IS M AND CRd IS M AND SR IS M THEN UrgencyDegree IS M;
RULE 16 : IF CR IS M AND CRd IS M AND SR IS L THEN UrgencyDegree IS L;
RULE 17 : IF CR IS M AND CRd IS L AND SR IS S THEN UrgencyDegree IS S;
```

Fig. 8. Fuzzy rules of phase urgency degree determination process.

b) APA

This type of agent is charged with managing the activated phase. Its goal is to maintain if possible the green time for the current phase until it has evacuated all of its enqueued vehicles. Using the extended indicator provided by active arc agents, it calculates the phase extended degree. The extended degree indicates the extended green time need level. To define the extension need degree, the **APA** collaborates with the set of arc agents involved in the current green phase.

The phase extender process is executed after each 1/3 of allocated green time, which the same as the min. red time. This will synchronize the two-parallel process of phase agents. The APA starts the extend process by sending a request message to the agents managing the active arcs demanding the extend indicators.

The extended indicators of all active arcs will be passed into the fuzzy mechanism to determine the phase extend degree and send it to the DA to request an extension. As mentioned previously, the input CR and CRd are standardized. An example of the rules used to define the extended degree is shown in Fig. 9.

```
RULE 3 : IF CR IS S AND CRd IS M THEN ExtendDegree IS PNo;
RULE 4 : IF CR IS S AND CRd IS L THEN ExtendDegree IS PNo;
RULE 5 : IF CR IS M AND CRd IS S THEN ExtendDegree IS PYes;
RULE 6 : IF CR IS M AND CRd IS M THEN ExtendDegree IS Maybe;
RULE 7 : IF CR IS M AND CRd IS L THEN ExtendDegree IS PNo;
RULE 8 : IF CR IS L AND CRd IS S THEN ExtendDegree IS Yes;
```

Fig. 9. Fuzzy rules of phase extend degree determination process.

3. Coordinator Agent

The objective of the CA is to coordinate with the connecting control group. It represents the communication interface agent of the ICG and plays a mediator role in all external communications. It exchanges the state of incoming arcs with the adjacent CA group member. It takes the succeeding arcs stat request from the local arc agents and contacts the CAs of the appropriate groups to get the requested data and response to the request. For its part, it provides the local arc stat to other groups. The CA controls all of the interaction flow with the ICG outside the environment, and it assures coordination and collaboration with others.

4. Decision Agent

The DA is the agent axis of our architecture; it decides to extend the active phase or switch to the selected phase. The decision is made in a collaborative way to avoid local optimization. The DA receives simultaneous requests from phase agents and then decides via fuzzy inference to either extend the current phase or to switch to the candidate phase. This agent sends the final decision to the phases and arc agents in real time.

The DA starts the decision process by checking the parameters of the phases to evaluate if the max. elapsed time of red and green time is reached. Then, it uses a fuzzy mechanism to make the decision and informs the phase and arc agents. Fig. 10 shows the decision-making process.

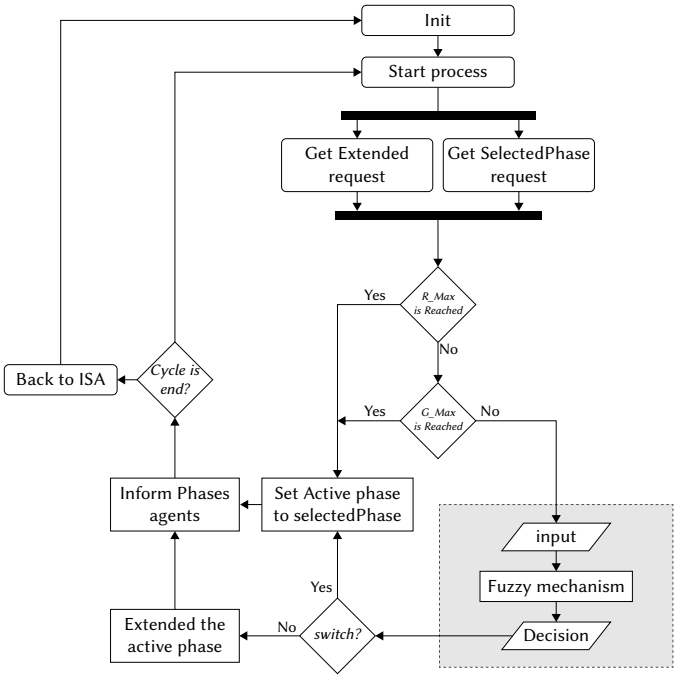


Fig. 10. Fuzzy decision mechanism.

The DA uses the phase urgency degree and extend degree provided by the phase agents to make the final decision. As in Fig. 11, there are five membership functions, including Zero (Z) Low (L), Medium (M), High (H), and Very High (VH), for the phase urgency degree. For the extended degree, there are five membership functions, including No (N), Perhaps No (PNo), Maybe Yes (MYes), Perhaps Yes (PYes), and Yes (Y). Finally, there are only two membership functions for the decision to switch to a candidate phase: No and Yes.

The decision-making process is based on a set of fuzzy rules. Fig. 12 shows an example of these rules.

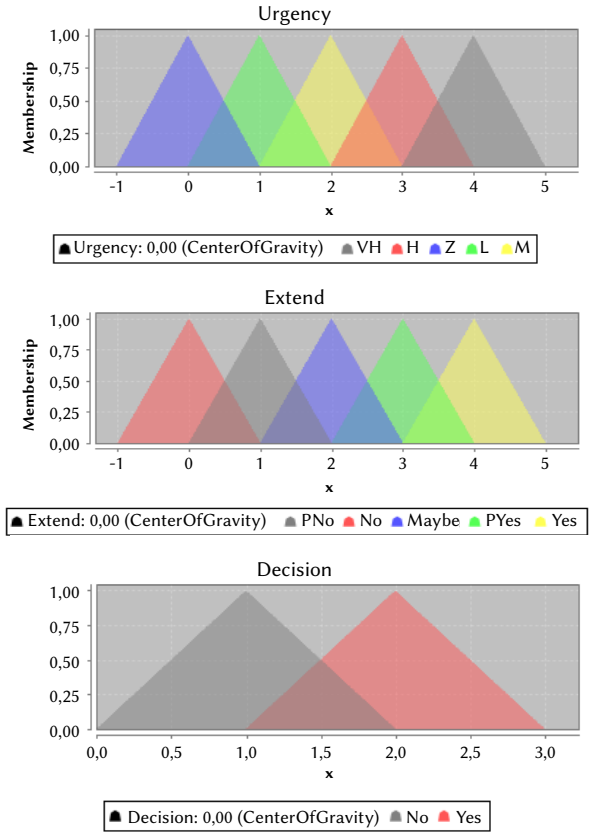


Fig. 11. Membership functions.

```
RULE 13 : IF ExtendDegree IS Mybe AND UrgencyDegree IS Mybe THEN Decision IS No ;
RULE 14 : IF ExtendDegree IS Mybe AND UrgencyDegree IS PYes THEN Decision IS Yes ;
RULE 15 : IF ExtendDegree IS Mybe AND UrgencyDegree IS Yes THEN Decision IS Yes ;
RULE 16 : IF ExtendDegree IS PYes AND UrgencyDegree IS No THEN Decision IS No ;
RULE 17 : IF ExtendDegree IS PYes AND UrgencyDegree IS PNo THEN Decision IS No ;
RULE 18 : IF ExtendDegree IS PYes AND UrgencyDegree IS Mybe THEN Decision IS No ;
```

Fig. 12. Fuzzy rules of the decision process.

The interaction of different agents in the designed MCTSO is illustrated in Fig. 13. All agents have a communication model to perceive their environment and handle the exchanged data flow within the agent society. A common Agent Communication Language (ACL) has been used to fulfill the communication model goals.

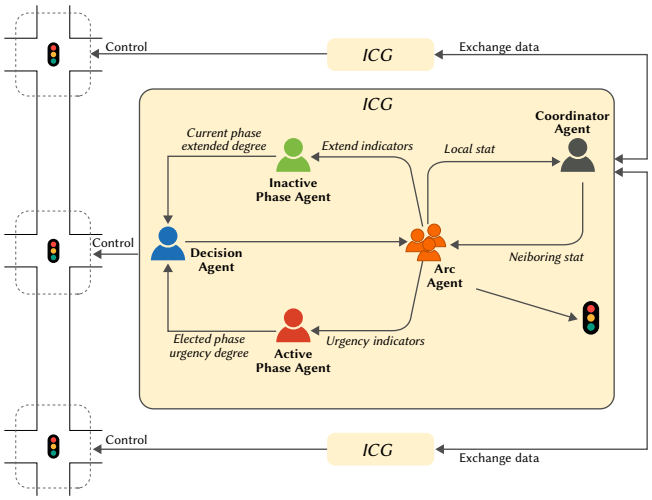


Fig. 13. The architecture of the MCTSO.

V. Experimental Results and Performance Analysis

The performance of the proposed system is validated and designed in the ANYLOGIC simulator, which is used to handle both the agent modeling and traffic simulation using a virtual road network that is shown in Fig. 14. It consists of 9 intersections controlled by 9 ICGs.

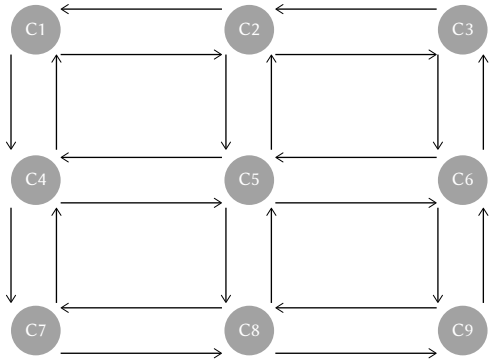


Fig. 14. Road network.

The fuzzy inference mechanism has been programmed in the JAVA language, using jFuzzyLogic, an open-source Fuzzy Logic library, and an FCL language implementation, which offers a fully functional and complete implementation of a fuzzy inference [30]. To link jFuzzyLogic to AnyLogic, we add the jFuzzyLogic library to Java external libraries in AnyLogic [31].

Fig. 15 describes the procedure of setting up a simulation for the MCTSC system. First, the road network is extracted in image format. AnyLogic is then used to convert the image into a simulation network. After obtaining the simulation network, vehicle mobility is generated using an origin-destination matrix. The arrival rate is adjusted to simulate the different scenarios of traffic demand. Then, the agent-based-modeling approach of AnyLogic is used to implement agents, and jFuzzyLogic library to handle the fuzzy decisions.

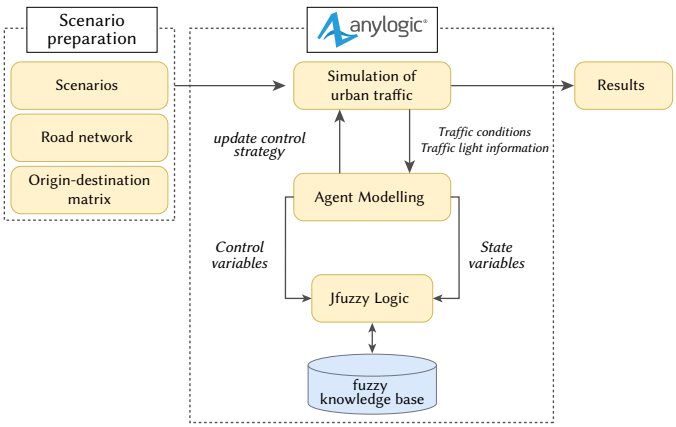


Fig. 15. The procedure of simulation.

A. Experimental Setup

In the study area, the frequency of entry at the source point is adjusted to simulate the varying traffic demand from the peak hour to the slack period. During the simulation, new vehicles are generated with an origin-destination matrix (OD). The OD simulates requests for network uses and represents possible situations of urban traffic conditions. The evaluation is carried out at both local intersections, to evaluate the local optimization, and for the whole network, to evaluate a global optimization.

Arcs are monitored up to the storage area length. We assume that the arc storage area can be varied proportionally to the lane number and the length of the given link. For links longer than 400 m, we monitor the point 150 m from the stop line, or 30 PVU, with 5 m as the typical car length, and all of the links in the other cases.

B. Results and Analysis

We assess the performance of MCTSO by referring to two control methods, namely the adaptative traffic signal optimization and adaptative multi-agent traffic signals control proposed by [32]:

- Adaptative traffic signal optimization (ATSO): standard version of the proposed MCTSO without agents.
- Adaptative multi-agent traffic signals control (AMTSC): represents a control traffic signals method based on multi-agent systems to control the traffic signals. The agents are organized in holonic architecture and a holonic Q-learning method is adopted to learn signals timing in two holarchical levels.

Since a feasible approach should smoothly deal with different traffic conditions, all control systems are tested on similar conditions and under 3 different scenarios: the first scenario allows the assessment of the performance of methods under low traffic demand, with 18000 PVU/hour as arrival rate. The second scenario describes medium traffic demand and represents a moderate congestion situation, with an arrival rate of 27000 PVU/hour. The third scenario provides results for high traffic demand with an arrival rate of 36000 PVU/hour. Each method is run for 180 minutes in each scenario case. Each case is repeated for 50 iterations to increase the reliability of the collected results.

In this study, the vehicle travel time and travel speed are chosen to build up an overview of the general performance of the control methods. The travel time represents the time between the departure of the vehicle from the origin point and the arrival at the destination point. Such criteria will provide us with the optimization level of our approach, and it includes the average stop time and network throughput indices.

Table II depicts the average vehicle's travel time and speed for each assessment scenario. It particularly shows that MCTSO allows the fastest travel time under all scenarios, and by consequence improves the number of vehicles that can use the network and reach their destinations compared to the other control methods. The results show that all control methods have acceptable performance under the first scenario, due to the low level of traffic flow. Increasing the arrival rate causes the congestion to get worse; the ATSO method presents an unacceptable travel time and speed. Thus, ATSO is not proficient with medium and higher congestion levels. At the higher level of congestion, our proposed method optimizes both travel time and speed, and it reduces these criteria compared to other controllers. We notice that in low traffic demand the average travel time attained by our proposed framework is 16,71% lower than that in ATSO, 6,82% than the AMTSC. While in medium traffic demand the corresponding improvement of the proposed framework is 20,36% and 11,08, respectively. The improvements become more important in high traffic demand and attain 37,65% compared to ATSO and 22,05% compared to AMTSC. As to the average travel speed, we observed that the MCTSO provided 8,26% according to the ATSO model and 4,82% improvement according to AMTSC in low traffic conditions, while this improvement rises 20,91% and 13,21% according to ATSO and AMTSC respectively in normal traffic density. In a heavy density scenario, the proposed model has better speed performance about 26,57% compared to the ATSO model and 16,76% compared to AMTSC. Furthermore, the standard deviation of the vehicle's travel time and travel speed of the proposed approach is lower than that in the other control methods. A high standard deviation means that there is a large amount of variability among the data, while a low standard deviation means that the data is less spread, thus very reliable. Consequently, the proposed approach is more reliable.

In addition, the ANOVA two factor with replication test yields a p-value of ≈ 0 (3,39E-40 for travel time and 2,18E-97 for travel speed) that is much smaller than the 0,05 level of alpha significance, meaning that the changes in used control methods had statistically a significant impact on the travel time in different traffic demand.

The reduction in travel time is due to the reduction in the set of key performances and by consequence, in a set of intersection indices. Fig. 16 summarizes the intersection metrics' key performances. The measurements are first locally aggregated in each intersection and at each time over evaluation scenarios; then, the performance average and other indices are calculated. The measurements show that MCTSO outperforms other controllers' methods in almost all metrics. Unfortunately, other methods failed to optimize green time management to mitigate traffic conditions.

Regarding the indices in Fig. 16, our proposed controller gives more throughput cars with less green time in all scenarios. This is due to optimizing the splitting of green time over all phases, reutilizing unused green time in phases with no cars in the storage zone. Also, the reduction in the average red time minimizes the travel time. The results also show a reduction in the number of cycles per intersection, which means that the system is suspended due to an empty storage zone in all intersection arcs. This augments the performance of the control system and makes evacuating the surrounding intersection more likely.

The results show that the proposed system is a practical approach and works smoothly with different traffic conditions.

VI. Conclusions

In this paper, a Multi-agent Cooperative Traffic Signal Optimization (MCTSO) is proposed to reduce congestion on urban roads by optimizing traffic light control with three contributions. First, the MCTSC interactive system involves real-time optimization. Second, there are two levels of coordination, the inter-junction and intra-junction, to avoid local optimization and build a control strategy

TABLE II. Percentage Improvement in Travel Time and Speed Over the Other Control Methods

	low traffic demand			medium traffic demand			high traffic demand		
Parameters	**ATSO**	**AMTSC**	**MCTSO**	**ATSO**	**AMTSC**	**MCTSO**	**ATSO**	**AMTSC**	**MCTSO**
Avg. travel time (s)	76,00	69,56	65,12	96,00	88,60	79,76	130,96	116,12	95,14
Standard deviation of delay	7,84	6,78	4,07	7,62	6,77	4,55	7,33	6,57	4,55
improvement travel time (%)	16,71%	6,82%	N/A	20,36%	11,08%	N/A	37,65%	22,05%	N/A
Avg. travel speed (km/h)	38,76	40,03	41,96	29,13	31,11	35,22	22,51	24,4	28,49
Standard deviation of delay	8,34	7,76	5,08	7,86	7,49	4,15	8,61	6,26	4,22
improvement travel time (%)	8,26%	4,82%	N/A	20,91%	13,21%	N/A	26,57%	16,76%	N/A

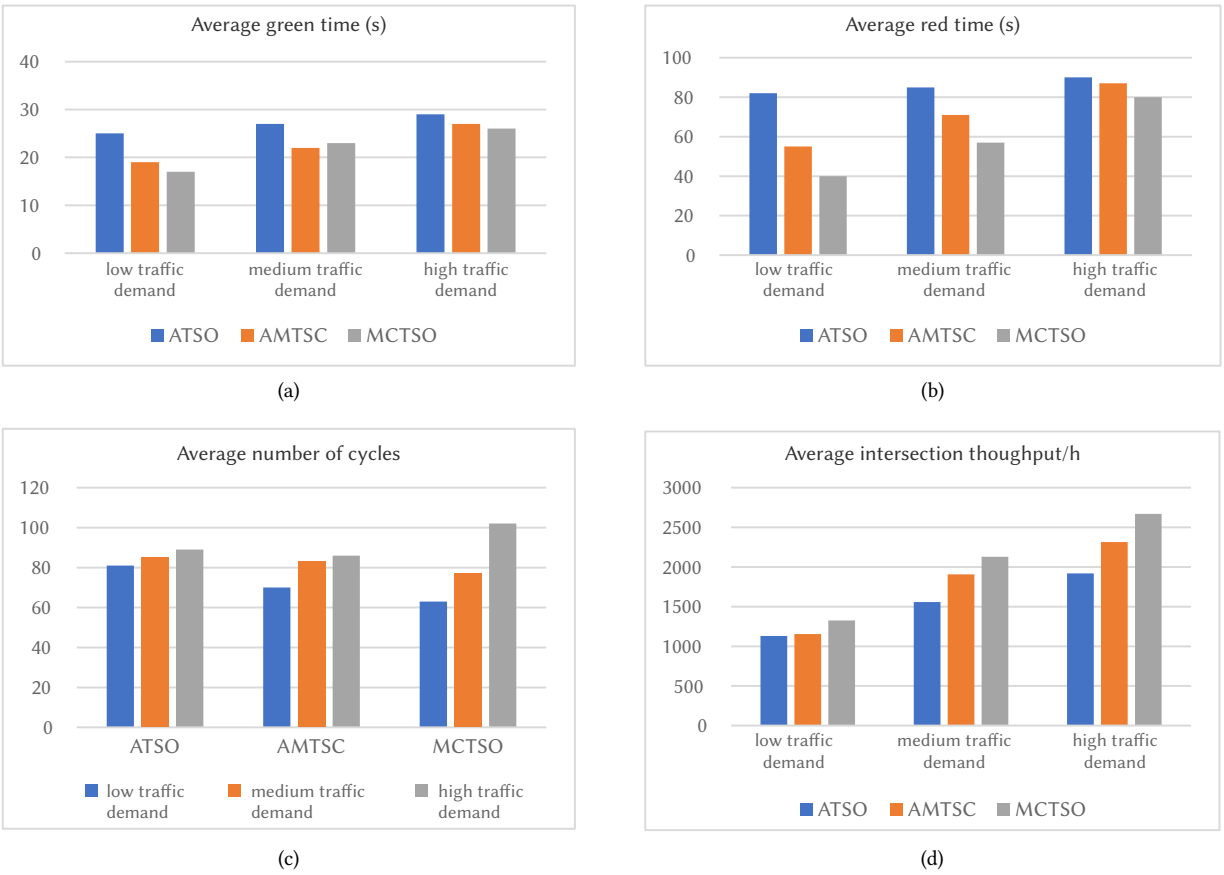


Fig. 16. The intersection metrics' key performances. The first chart shows the average green time assigned to active phases, the second chart shows the average red time assigned to inactive phases, the third chart shows the average number of cycles, and the fourth chart shows the average intersection throughput. All these measurements had been taken for each method and under different scenarios.

that takes into consideration all connecting intersections. Third, distributed control restricts the cooperative scope in neighbors and allows the parallel control.

The proposed system can handle a large multi-intersection network with many alterations in the road infrastructure, and hence facilitates extensibility. The system also increases the robustness and throughput of the network, as shown in the simulation executed in the AnyLogic simulator. The performance of the proposed approach has been compared to the same approach without agents and another adaptative multi-agent optimization. In these performance comparisons, the average travel time and speed were selected as signaling performance criteria. In low traffic demand scenarios, the proposed Multi-agent Cooperative Traffic Signal Optimization model provided 16,71%–6,82% improvement in average travel time and 8,26%–4,42% improvement in average travel speed compared to both adaptative traffic signal optimization and adaptative multi-agent traffic signals control respectively. These improvement values become more important when traffic demand increases and the traffic congestion goes worst, and they are respectively up to 37,65%–22,05% for average travel time and to 26,57%–16,76% for average travel speed compared to adaptative traffic signal optimization and adaptative multi-agent traffic signals control models in high traffic demand. Both local and network performance keys are investigated based on the computational experiments in different traffic condition scenarios. The proposed agent-based optimization shows a better result and can adapt smoothly with different traffic demands. It can significantly optimize performance keys such as travel time, stop time, intersection throughput, and so on.

In the future, the framework shall be further extended to other traffic control fields. For traffic signal control, one extension of this approach is to include priority vehicles and add the special management of priority links. Meanwhile, it is necessary to develop the intelligent optimization approach for operations concerning large uncertainties in the road network, such as disturbances and emergencies.

REFERENCES

[1] B. Chen and H. H. Cheng, "A review of the applications of agent technology in traffic and transportation systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 2, pp. 485–497, 2010, doi: 10.1109/TITS.2010.2048313.

[2] M. R. Evans and D. S. Elston, "Agent-Based Modeling and Simulation for Transportation, VASTO: Evolutionary Agent System for Transportation Outlook," pp. 1–88, 2013.

[3] de la P. F. Román J.A., Rodríguez S., *Highlights of Practical Applications of Scalable Multi-Agent Systems. The PAAMS Collection*, vol. 616. Cham: Springer International Publishing, 2016.

[4] K. Małecki, "A computer simulation of traffic flow with on-street parking and drivers' behaviour based on cellular automata and a multi-agent system," *Journal of Computational Science*, vol. 28, pp. 32–42, 2018, doi: 10.1016/j.jocs.2018.07.005.

[5] H. Hamidi and A. Kamankesh, "An Approach to Intelligent Traffic Management System Using a Multi-agent System," *International Journal of Intelligent Transportation Systems Research*, vol. 16, no. 2, pp. 112–124, 2018, doi: 10.1007/s13177-017-0142-6.

[6] T. Anagnostopoulos, C. Luo, J. Ramson, K. Ntalianis, V. Kostakos, and C. Skourlas, "A multi-agent system for distributed smartphone sensing cycling in smart cities," *Journal of Systems and Information Technology*,

vol. 22, no. 4, pp. 119–134, 2020, doi: 10.1108/JSIT-12-2018-0158.

[7] Z. Zhang, J. Yang, and H. Zha, "Integrating independent and centralized multi-agent reinforcement learning for traffic signal network optimization," *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, vol. 2020-May, pp. 2083–2085, 2020.

[8] M. C. Ho, J. M. Y. Lim, K. L. Soon, and C. Y. Chong, "An improved pheromone-based vehicle rerouting system to reduce traffic congestion," *Applied Soft Computing Journal*, vol. 84, p. 105702, 2019, doi: 10.1016/j.asoc.2019.105702.

[9] K. L. Soon, J. M. Y. Lim, and R. Parthiban, "Coordinated Traffic Light Control in Cooperative Green Vehicle Routing for Pheromone-based Multi-Agent Systems," *Applied Soft Computing Journal*, vol. 81, p. 105486, 2019, doi: 10.1016/j.asoc.2019.105486.

[10] A. Eydi, S. Panahi, and I. N. Kamalabadi, "User-based Vehicle Route Guidance in Urban Networks Based on Intelligent Multi Agents Systems and the ANT-Q Algorithm," *International Journal of Transportation Engineering*, vol. 4, no. 3, pp. 147–161, 2016.

[11] A. Dorri, S. S. Kanhere, and R. Jurdak, "Multi-Agent Systems: A Survey," *IEEE Access*, vol. 6, pp. 28573–28593, 2018, doi: 10.1109/ACCESS.2018.2831228.

[12] J. Jin and X. Ma, "Hierarchical multi-agent control of traffic lights based on collective learning," *Engineering Applications of Artificial Intelligence*, vol. 68, no. January, pp. 236–248, 2018, doi: 10.1016/j.engappai.2017.10.013.

[13] M. Xu, K. An, L. H. Vu, Z. Ye, J. Feng, and E. Chen, "Optimizing multi-agent based urban traffic signal control system," *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, vol. 23, no. 4, pp. 357–369, 2018, doi: 10.1080/15472450.2018.1501273.

[14] S. Darmoul, S. Elkosantini, A. Louati, and L. Ben Said, "Multi-agent immune networks to control interrupted flow at signalized intersections," *Transportation Research Part C: Emerging Technologies*, vol. 82, pp. 290–313, 2017, doi: 10.1016/j.trc.2017.07.003.

[15] I. H. Tchappi *et al.*, "A critical review of the use of holonic paradigm in traffic and transportation systems," *Engineering Applications of Artificial Intelligence*, vol. 90, no. March 2019, p. 103503, 2020, doi: 10.1016/j.engappai.2020.103503.

[16] Y. Wang, X. Yang, H. Liang, and Y. Liu, "A review of the self-adaptive traffic signal control system based on future traffic environment," *Journal of Advanced Transportation*, vol. 2018, 2018, doi: 10.1155/2018/1096123.

[17] M. Vasirani, F. Klügl, E. Camponogara, and H. Hattori, "Special issue on intelligent agents in traffic and transportation," *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, vol. 19, no. 1, pp. 1–2, 2015, doi: 10.1080/15472450.2013.856719.

[18] P. Mannion, J. Duggan, and E. Howley, "An Experimental Review of Reinforcement Learning Algorithms for Adaptive Traffic Signal Control," *Autonomic Road Transport Support Systems*, pp. 47–66, 2016, doi: 10.1007/978-3-319-25808-9.

[19] A. Daeichian and A. Haghani, "Fuzzy Q-Learning-Based Multi-agent System for Intelligent Traffic Control by a Game Theory Approach," *Arabian Journal for Science and Engineering*, vol. 43, no. 6, pp. 3241–3247, 2018, doi: 10.1007/s13369-017-3018-9.

[20] S. El-Tantawy, B. Abdulhai, and H. Abdelgawad, "Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (marlin-atsc): Methodology and large-scale application on downtown toronto," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 3, pp. 1140–1150, 2013, doi: 10.1109/TITS.2013.2255286.

[21] M. Collotta, L. Lo Bello, and G. Pau, "A novel approach for dynamic traffic lights management based on Wireless Sensor Networks and multiple fuzzy logic controllers," *Expert Systems with Applications*, vol. 42, no. 13, pp. 5403–5415, 2015, doi: 10.1016/j.eswa.2015.02.011.

[22] A. Latif and P. Megantoro, "Traffic Light Regulatory System Based on Fuzzy Algorithm Using Microcontroller," *Journal of Physics: Conference Series*, vol. 1464, no. 1, 2020, doi: 10.1088/1742-6596/1464/1/012034.

[23] Y. Bi, D. Srinivasan, X. Lu, Z. Sun, and W. Zeng, "Type-2 fuzzy multi-intersection traffic signal control with differential evolution optimization," *Expert Systems with Applications*, vol. 41, no. 16, pp. 7338–7349, 2014, doi: 10.1016/j.eswa.2014.06.022.

[24] "World Population Projections - Worldometer." [Online]. Available: https://www.worldometers.info/world-population/world-population-projections/. [Accessed: 05-Feb-2021].

[25] M. Xu, Z. Ye, H. Sun, and W. Wang, "Optimization model for transit signal priority under conflicting priority requests," *Transportation Research Record*, vol. 2539, pp. 140–148, 2016, doi: 10.3141/2539-16.

[26] X. Li and J. Q. Sun, "Signal Multiobjective Optimization for Urban Traffic Network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 11, pp. 3529–3537, 2018, doi: 10.1109/TITS.2017.2787103.

[27] A. C. Olivera, J. M. García-Nieto, and E. Alba, "Reducing vehicle emissions and fuel consumption in the city by using particle swarm optimization," *Applied Intelligence*, vol. 42, no. 3, pp. 389–405, 2015, doi: 10.1007/s10489-014-0604-3.

[28] M. Kempa and Z. A. Mann, "Model Driven Architecture," *Informatik-Spektrum*, vol. 28, no. 4, pp. 298–302, Aug. 2005, doi: 10.1007/s00287-005-0505-2.

[29] J. Ferber and O. Gutknecht, "A meta-model for the analysis and design of organizations in multi-agent systems," *Proceedings - International Conference on Multi Agent Systems, ICMAS 1998*, pp. 128–135, 1998, doi: 10.1109/ICMAS.1998.699041.

[30] P. Cingolani and J. Alcalá-Fdez, "jFuzzyLogic: A Java Library to Design Fuzzy Logic Controllers According to the Standard for Fuzzy Control Programming," *International Journal of Computational Intelligence Systems*, vol. 6, no. SUPPL1, pp. 61–75, 2013, doi: 10.1080/18756891.2013.818190.

[31] "AnyLogic: Simulation Modeling Software Tools & Solutions for Business." [Online]. Available: https://www.anylogic.com/. [Accessed: 28-Mar-2021].

[32] M. Abdoos, N. Mozayani, and A. L. C. Bazzan, "Holonic multi-agent system for traffic signals control," *Engineering Applications of Artificial Intelligence*, vol. 26, no. 5–6, pp. 1575–1587, 2013, doi: 10.1016/j.engappai.2013.01.007.

Abdelouafi Ikidid

He is a PhD candidate at the Computer Science Department of Cadi Ayyad University in Marrakesh, Morocco. He received his Master's degree in Information Systems Engineering from the same university in 2016. His research interests are in software engineering, focusing on multi-agent systems and artificial Intelligence.

Abdelaziz El Fazziki

He received the M.S. degree from the University of Nancy, France, in 1985, and the Ph.D. degree in computer science from the Cadi Ayyad University in 2002. He is a Professor of computer science at Cadi Ayyad University, where he has been since 1985. He is the author of over 50 papers on software engineering. His research interests are in software engineering and focusing on information system development

Mohammed Sadgal

He received the Ph.D. degree in computer science from the University of Lyon in 1989, and the Ph.D. degree in computer science from Cadi Ayyad University in 2005. From 1985 to 1987, he was an Associate Researcher with Lyon I, France. He is currently a Professor with Cadi Ayyad University, Marrakesh, Morocco. His research interests include computer vision, artificial intelligence, and multi-agent systems.

Digit Recognition Using Composite Features With Decision Tree Strategy

Chung-Hsing Chen[1,2], Ko-Wei Huang[1] *

[1] Department of Electrical Engineering, National Kaohsiung University of Science and Technology, Kaohsiung City (Taiwan)
[2] Plustek Inc., Taipei City (Taiwan)

Received 11 February 2022 | Accepted 21 September 2022 | Published 14 December 2022

Abstract

At present, check transactions are one of the most common forms of money transfer in the market. The information for check exchange is printed using magnetic ink character recognition (MICR), widely used in the banking industry, primarily for processing check transactions. However, the magnetic ink card reader is specialized and expensive, resulting in general accounting departments or bookkeepers using manual data registration instead. An organization that deals with parts or corporate services might have to process 300 to 400 checks each day, which would require a considerable amount of labor to perform the registration process. The cost of a single-sided scanner is only 1/10 of the MICR; hence, using image recognition technology is an economical solution. In this study, we aim to use multiple features for character recognition of E13B, comprising ten numbers and four symbols. For the numeric part, we used statistical features such as image density features, geometric features, and simple decision trees for classification. The symbols of E13B are composed of three distinct rectangles, classified according to their size and relative position. Using the same sample set, MLP, LetNet-5, Alexnet, and hybrid CNN-SVM were used to train the numerical part of the artificial intelligence network as the experimental control group to verify the accuracy and speed of the proposed method. The results of this study were used to verify the performance and usability of the proposed method. Our proposed method obtained all test samples correctly, with a recognition rate close to 100%. A prediction time of less than one millisecond per character, with an average value of 0.03 ms, was achieved, over 50 times faster than state-of-the-art methods. The accuracy rate is also better than all comparative state-of-the-art methods. The proposed method was also applied to an embedded device to ensure the CPU would be used for verification instead of a high-end GPU.

Keywords

Decision Tree, E13B Fonts, Feature Extraction, Image Classification, Multilayer Perceptron.

DOI: 10.9781/ijimai.2022.12.001

I. Introduction

It is important to determine whether a scanned image for recognition can be used without employing a magnetic ink reader. Several studies on using images for character recognition and handwritten characters have been conducted. However, recently researched architecture, such as deep learning, is not suitable for embedded devices because the high computational resources and elapsed time cannot meet the requirements of a compact embedded system.

A. Motivation

The scanner market has begun to shrink, and scanner manufacturers have turned to special-purpose scanners or readers. Considering the common barcode reader on the market as an example, barcode recognition may be achieved with mobile phones or software; however, in commercial applications, customers prefer a device that does not need to rely on the computing power of the cash register to directly read the barcode. The identification is implemented on the reader. For the cash register, the reader is just a HID (Human Input Device), just like someone helping you type. This application scenario is used on magnetic ink character recognition (MICR) as well. We want to design a device like a MICR reader, which means that we need to complete image recognition and HID output on a single-board controller. This requirement necessitates a simple algorithm that does not require additional GPU or NPU devices—the primary goal of the current research. Assuming that the recognition speed of each character is 50 ms, the total recognition time of one check can be computed to be 1,400 ms. This value is more than one second and hence, does not meet the recognition time requirement. The desired computation time is not achievable using the technologies proposed in recent years. Hence, smaller data features and simpler recognition decisions are needed to accomplish the desired computation time.

We cannot use the image of a character to reduce the amount of information required for recognition. Instead, a method must be developed to reduce the amount of required information for quick calculations. Before using an image as input data for machine learning or deep learning, 1050 pixels need to be processed. This value is

* Corresponding author.
E-mail addresses: I110154101@nkust.edu.tw (C. H. Chen), elone.huang@nkust.edu.tw (K. W. Huang).

greater than the 784 pixels of MNIST's handwritten digital database [1]. Additionally, more explicit features are needed to classify the input data, which can reduce the required amount of input data during classification and facilitate faster decisions and classification. The specifications for checks were defined by the International Standards Organization in 1977, and the latest version is ISO 1004-1:2013, comprising 10 numbers and 4 symbols. E13B is one of the two primary MICR fonts used for printing checks and other payment documents. This font is popular in North America and most of Asia, while the other major MICR font, CMC7, is the standard font used in most of Europe and South America. The E13B font uses characters with unique characteristics designed to produce a distinctive pattern when scanned by a magnetic reader. Both the CMC7 and E13B are read using the unique magnetic characteristics of the font. The characters of E13B are read by detecting the strength of the magnetic signal in a continuous "waveform" pattern from left to right, while CMC7 is read like a barcode, performing a series of "on/off" tests from left to right.

The first and second rows consist of 1, 2, 3, 4, 5, and 6, 7, 8, 9, 0, respectively, as shown in Fig. 1.

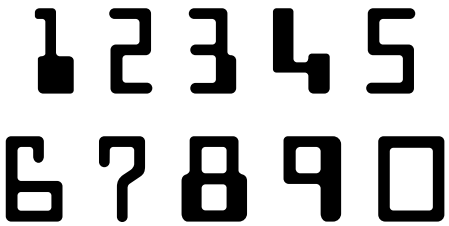


Fig. 1. Illustration of E13B font digits.

The E13B [2] font is specifically defined for MICR, comprising ten digits (0–9) and four symbols. The symbols comprise three separate squares or rectangular symbols with specific arrangements, as shown in Fig. 2.

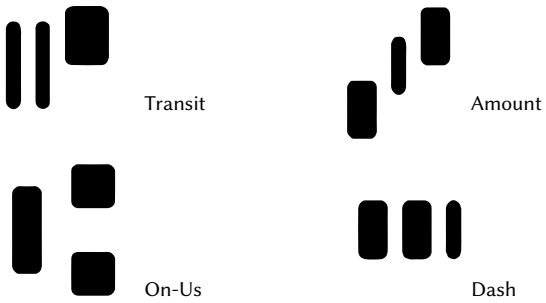


Fig. 2. Illustrations of the E13B font symbols.

In this study, we propose a method to identify all the characters of E13B using feature recognition and relevant correlations to prove the validity and accuracy of the proposed method. Five sets of control groups are applied using the same sample neural network model for training: a four-layer and five-layer MLP [3]–[5] network structure, LeNet-5, Alexnet and hybrid convolutional neural network–support vector machine (CNN-SVM) control groups [6]. Artificial neural networks, such as the relatively new LeNet-5 [7]–[11] have made considerable progress in character recognition.

The neural networks of the five control groups used herein have all ten recognition categories consistent with E13B digital fonts. In this study, we use statistical features [12]–[20] such as image density, pattern, and target relative position features, using a decision tree diagram [21] to achieve classification.

B. Contribution

A low-energy-consuming algorithm is the main contribution of this research. For the sake of accuracy, advanced research continuously increases the depth of the neural network and the input parameters, but this research direction cannot meet the requirements of special-purpose scanners. Restricted by the hardware of special-purpose scanners, computational requirements and I/O requirements of image data limit the feasibility of neural networks for use in image recognition. Although this study uses a more traditional approach to image recognition, the text features are captured more creatively, and the decision tree is not trained by machine learning but by general logic because of the need to work around the design in cases where there is a limit on computational resource consumption.

We used a laptop with an i5-2520m 2.5Ghz CPU and 4GB RAM for verification. The process in this experiment used the CPU instead of GPU to demonstrate the cost-effectiveness of the proposed method. Our proposed method obtained all test samples correctly, with a recognition rate close to 100%. A prediction time of less than one millisecond per character, with an average value of 0.03 ms, was achieved, over 50 times faster than state-of-the-art methods. The accuracy rate is also better than all comparative state-of-the-art methods.

The remainder of this paper is organized as follows. The literature review is presented in Section II. The research methodology is described in Section III. The performance evaluation is outlined in Section IV. Finally, conclusions and suggestions for future research are provided in Section V.

II. Literature Review

Text recognition can be divided into several stages. In addition to the pre-processing of the image and the layout segmentation of the text [22], past research papers can be roughly divided into two aspects, one is feature extraction and the other is classifier. Fumitaka and Shridhar used the static Zoning topo to represent the contour of a character [23]. According to the direction of the contour, it is divided into four groups, namely horizontal, vertical, and diagonal in both directions (45° and 135°). The number of contours in each group is its features.

Singh and Budhiraja proposed several features for recognizing handwritten Gurmukhi text, such as projection histogram features, zoning, distance profile features, and background directional distribution features [24].

Verma and Aki published a study on feature extraction methods and classification techniques used in OCR systems [25]. The features used in the study were statistical features and structural features. In their research, statistical feature techniques include zoning, moments, projection histograms, n-tuples, crossing and distances etc. Structural features included convexities, concavities, number of end points, number of holes, etc.

Dimpy et al. proposed a feature extraction of pneumonia data using DensenNet-169 architecture [26]. The original image database was a 3-channel image that was resized from 1024 x 1204 to 224 x 244 pixels. The resizing reduces the need for heavy computation and speed up the processing. After feature extraction through DenseNet-169, a one-dimensional vector of 50,176 x 1 was obtained and input to a different classifier. The findings of this study showed that the best classifier model was SVM (RBF kernel) with an AUC value of 0.80022.

Aimin Yang et al. proposed a feature extraction approach for recognizing tumors by using a local binary model algorithm for image preprocessing [27]. The local binary patten and convolutional neural network algorithm are used to image and extract features from tumor CT images in the medical field, and the recognition rate of this method for medical images was 99.7%.

Lehal proposed a powerful font-independent Gurmukhi OCR system [28]that used four classifiers. The first two classifiers are a binary tree classifier and k-NN classifier. They operate sequentially and use structural features for feature extraction. The third classifier is an SVM using a Gabor filter with a vector size of 189. The fourth classifier is also an SVM, which operates on certain structural and statistical features and achieved a 98.18% accuracy.

Kobayashi et al. proposed the use of histogram of gradient (HOG) to extract candidate features from an image located in a grid. Moreover, they applied principal component analysis to obtain vectors [29]. This method used linear SVM to detect the pedestrian/non-pedestrian and achieved an accuracy of 99.3%.

Singh et al. proposed the use of the Gabor filter for handwritten Gurmukhi character recognition [30]. They performed 5-fold cross-validation on the entire database using the RBF-SVM classifier and achieved 94.29% accuracy.

Shawon et al. proposed a Bengali handwritten letter recognition using 76 features and MLP as a classifier [31]. The feature set developed for recognizing handwritten characters of the Bengali alphabet consists of 24 shadow features, 16 centroid features and 36 longest-run features. The recognition accuracy of the MLP designed to process this feature set was 86.46% and 75.05%, respectively, on the training set and test set samples. This method is useful for the development of a complete OCR system for handwritten Bengali text.

Rajinikanth et al. proposed a meta-heuristic algorithm to solve the multi-thresholding of RGB scale images by using entropy value [32].

Acharya et al. published a new image dataset for the Devanagari script called the Devanagari Handwritten Character Dataset (DHCD) [33]. It consisted of 92,000 images across 46 unique classes of characters of Devanagari script segmented from handwritten files. In addition, they proposed a deep learning architecture (CNN) for the recognition of these characters. The accuracy of the proposed system was 98.47% using the DHCD dataset.

Ramadhan et al presented a comparative analysis of the accuracy and process length of each algorithm [34]. The use of K-Nearest Neighbor (KNN) and Decision Tree (DT) algorithms to detect DDoS attacks was analyzed. In addition, they used the CICIDS2017 dataset, which consists of the world PCAP data format. The results of their study showed that the accuracy of the DT algorithm in detecting DDoS attacks was higher than the KNN value algorithm. The accuracy of DT was 99.91%, while the accuracy of k-NN was was 98.94%.

Assegie et al. proposed a method for recognition of handwritten numbers using a decision tree classification model [35]. Decision tree classification is a machine learning method that uses predefined labels from past known sets to determine or predict classes for future datasets for which the class labels are unknown. They used a standard Kaggle digits dataset to train and recognize handwritten digits using a decision tree classification model. This experiment was trained using a Kaggle dataset containing 42,000 rows and 720 columns. The method had an accuracy of 83.4%. Based on its accuracy, we can see that a decision tree classification model for handwritten number recognition is quite efficient.

Ahlawat et al. proposed a hybrid model of CNN-SVM, a hybrid model of a powerful convolutional neural network (CNN) and supporting machine (SVM) [36], for handwritten digit recognition in the MNIST dataset. A hybrid model of CNN-SVM was proposed for handwritten digit recognition that utilized automatic feature generation of CNN and output prediction using SVM. Experimental results show that their proposed method achieved a classification accuracy of 99.28% on the MNIST dataset.

Barbhuiya et al. proposed a sign language recognition system using Alexnet and VGG16 for feature extraction with SVM as a classifier [37]. The system uses Alexnet and VGG16 to pre-train the American Sign Language dataset, and then uses SVM to pre-train the feature classification, enabling the classification of gestures. The proposed American Sign Language recognition system, has been compared with some state-of-the-art recognition approaches, including both a random 70–30 and leave-one-subject-out validations. The proposed method using the modified AlexNet and SVM classifier has a recognition accuracy of 99.82%, which is the highest among the compared methods using a random 70–30 cross-validation.

Varun et al. proposed a multi-classifier for classifying imbalanced financial news datasets [38]. The proposed architecture uses the SMOTE method to generate similar synthetic samples, which can balance the original imbalanced dataset. Past research has proven that imbalanced datasets have adverse effects on machine learning, thus making the author ensure that the dataset was balanced in his architecture. The proposed architecture was compared with other compared architectures, with the worst recognition accuracy of 34% and the best recognition accuracy of 99%. The proposed architecture achieved a recognition accuracy of 100% with a dataset balanced by random forest with SMOTE.

Manju et al. proposed an RGB and RGB-D static gesture recognition method [39] using the fine-tuned VGG19 model. The fine-tuned VGG19 model uses feature concatenation layers from RGB and RGB-D images to improve the accuracy of the neural network. The proposed model was compared with different CNN models, such as VGG16, CaffeNet, VGG19, and Inception V3, which were not fine-tuned. It achieved 94.8% accuracy using the test results of the ASL dataset. The maximum recognition accuracy of these four models was 88.15%, much lower than that of the proposed model.

III. Methodology

In this study, 24,000 E13B digits were extracted from 1,079 check samples. The treatment and control groups were used to verify the recognition rates. The treatment group utilized the multi-feature method and directly classified the 24,000 samples to obtain the recognition results. The control group worked with a two-layer and three-layer MLP, LeNet-5, AlexNet and CNN-SVM using the k-Fold cross-validation method, where k was 2, 5, and 10 [40].

A. Collecting Data

For this study, 1,079 checks were collected from related companies, and 24,000 characters were extracted from them to train the AI model. The study methodology did not require a physical copy of the checks owing to its feature characteristic recognition.

The check is an important accounting document of expenditure or income in the company; it is confidential information within each company. We have legally collected nearly 2,000 Taiwan check samples from our customers through Plustek Inc. After discarding the bad samples, the remaining 1,079 checks were used in this study. These check samples are based on company confidentiality principles, even though Plustek Inc. has eliminated the customer's astute data before providing the check images for this study. In principle, these check images cannot be made public, but images and labels that have been cut into single characters can be provided free of charge for academic purposes or licensed to other for-profit organizations for commercial use.

Using the control group of the neural network of the MLP model, we applied the k-fold cross-validation method by dividing the data into k groups. Three groups with k = 2, 5, and 10 were used to validate the accuracy and speed of the proposed method. Fig. 3 represents the checks used in Taiwan; the information on the MICR is shown as the

Fig. 2 that identified as “Transit” check number ”Transit” ”OnUs” bank’s routing number ”OnUs” interchange code ”Transit” account number ”Transit” ”Amount” amount ”Amount”.

Fig. 3. Checks used in Taiwan. (The amount and the respective symbol are not displayed in this image because the check has not been cashed and processed in the clearing house).

Fig. 4 shows a check used in the United States. The information on the MICR is shown as the Fig. 2 that identified as “Transit” check number “Transit” “OnUs” bank’s routing number ”OnUs” account “Transit” “Amount” amount “Amount”. Although the format of Taiwan’s check is different from that of the United States, it conforms to the ISO 1004-1:2013 standard. In Taiwan, the check format has an additional clearing house code after the bank code because Taiwan uses a computerized exchange system to automatically exchange each clearing house’s notes.

Fig. 4. Checks used in the United States.

B. Pre-Processing

The following algorithm is used for pre-processing, where the background is separated from the foreground, the foreground is the text, and the background is the non-text area. HTGS(k) is a statistical histogram, whose value can be evaluated using the formula given below, where k is the grayscale ranging from 0 to 255, p represents each pixel, and gray is the grayscale value that utilizes the average of the R, G, and B channels. The HTGS(k) can be calculated as indicated in Eq. (1):

$$HTGS(k) = |\{p \mid k = \lfloor (Gray(p) + LastGray(p))/2 \rfloor, \; 50 < Gray(p) < Mean, \; |Gray(p) - LastGray(p)| > 40\}| \quad (1)$$

where LastGray represents the grayscale value of the previous image and Mean represents the average grayscale value of the image. The obtained value of HTGS(k) is equal to k_{max}, which is the binarization threshold of the entire area.

C. Characters in Digit Recognition

1. Statistical Features

The statistical feature of this method utilizes the surface density component to divide an image into four equal blocks: upper left (UL), lower left (LL), upper right (UR), and lower right (LR), as shown in Fig. 5. The density refers to the ratio of the black dots to the total area. The number of black dots in the upper-left block is denoted by nUL. The density in the upper left block (dUL) can be obtained by dividing the quarter of the total area (Width*Height) with nUL. The formulae are as follows:

1. Area = $|(Width/2)| * |(Height/2)|$
2. nUL = black pixel count in Upper Left Block.
3. dUL = nUL/Area; Density in Upper Left Block.
4. nLL = black pixel count in Lower Left Block.
5. dLL = nLL/Area; Density in Lower Left.
6. nUR = black pixel count in Upper Right Block.
7. dUR = nUR/Area; Density in Upper Right.
8. nLR = black pixel count in Lower Right Block.
9. dLR = nLR/Area; Density in Lower Right.

A description of the examined objects and tools used during the experiment, and the factors that could affect the experimental results are discussed in subsequent sections.

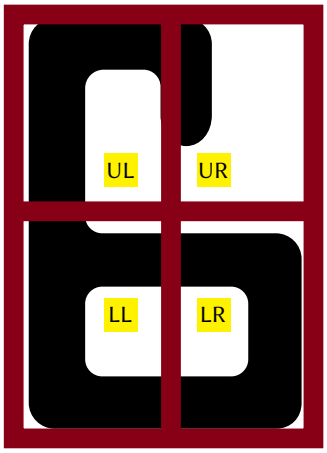


Fig. 5. Schematic of statistical features.

2. Pattern Features

The E13B font used in the check is composed of ten numbers ranging from 0 to 9 and four symbols. The numbers and symbols are identified separately in this study. The pattern features were analyzed by calculating the total number of black pixels in an area. Six probes were designed for the digit characters 3, 5, 6, 8, and 0. These feature labels distinguish the five-digit characters, as shown in Fig. 6. $Probe_6$ detects the blank area of the number 6 in the upper-right corner. There is a visible empty block with an E13B print on number 6, which indicates that the block should not contain any black pixels.

Fig. 6. Schematic of geometric features.

In Table I, the coordinates of the rectangles are denoted by (X’, Y’, W’, H’), where X’ and Y’ are the coordinates of the leftmost and highest points of the rectangles, respectively, and W’ and H’ are the widths and heights of the rectangles, respectively. The width and height of image are represented by W and H, respectively. $Probe_0$, $Probe_8$, $Probe_3$-Upper, $Probe_3$-Lower, $Probe_5$, and $Probe_6$ denote the number of black pixels in the six rectangles, respectively.

TABLE I. Reference Labels for Each Coordinate

Label	Coordinates (X', Y', W', H')
$Probe_0$	W / 3, H / 4,
	W - ((W / 3) * 2), W - (H / 4) * 2
$Probe_8$	(W / 5) * 2, H / 3, W / 5, H / 12
$Probe_3$-*Upper*	0, (W / 4), W / 2, H / 7
$Probe_3$-*Lower*	0, (H / 6) * 4, W / 2, H / 7
$Probe_5$	W / 2, (H / 6), W / 2, H / 6
$Probe_6$	W - (W / 4), 0, W / 4, H / 4

3. Decision Tree

The recognition of numbers uses hybrid features from density features and geometric features, based on those decision conditions to create a decision tree model. As shown in Fig. 7, the three numbers 0, 2, 5 have unique features that can be easily classified, while the remaining numbers are divided into two groups using density features. The first group consists of 3, 7, 9, and the second group consists of 1,4,6,8. The detection area of $Probe_0$ is binarized without any black pixels, i.e., the threshold value of $Probe_0$ binarization is calculated using the HTGS method. If the result is not zero, the second set of feature conditions is compared. The number 2 is the only one of the ten numbers whose dLL is larger than its dUL and whose dUR is larger than its dLR. Hence, if the candidate being tested has the above-mentioned features, then the number is classified as 2.

If the result is not 0 or 2, a third set of feature conditions is compared. The next feature condition is the geometric feature $Probe_3$-Lower and $Probe_5$, and both are zero. If the candidate has this feature, then the candidate can be predicted to be number 5.

The fourth set of feature conditions divides the numbers into two groups, that is, (3, 7, 9) and (1, 4, 6, 8). The fourth decision condition is that the dUR must be larger than the dLL. If this condition is met, the number is in the first group. If the condition is not met, the number is in the second group.

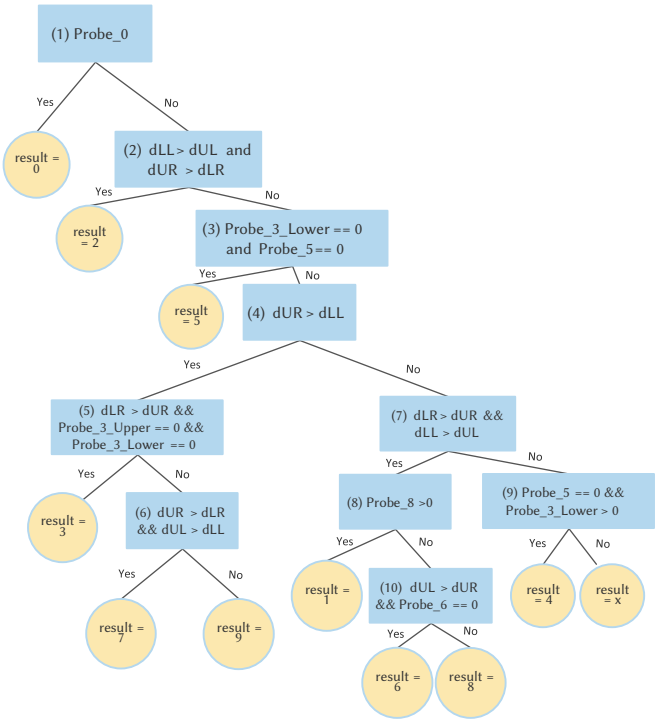


Fig. 7. Decision tree for digit recognition.

The fifth set of feature conditions divide the remaining numbers into two groups. The first group consists of 3 and the second consists of 7 and 9. The condition is that the dLR must be larger than the dUR. The sixth feature conditions are that dUR must be greater than dLR and dUL must be greater than dLL. If the condition is met, the candidate is predicted to be 7 and if not, it is predicted to be 9.

For the group of numbers 1, 4, 6, 8, a seventh set of feature conditions is used. The condition is that the dLR must be larger than dUR and dLL must be larger than dLR. The numbers 1, 4, 6, 8 are divided into two groups, 1, 6, 8 and 4. The next group is 1, 6, 8, and the decision condition is the eighth group of the feature conditions , which is the geometric feature detection area $Probe_8$. If there is a black pixel in $Probe_8$, the candidate can be predicted as number 1. If there is no black pixel, then the number can be either 6 or 8. A tenth set of feature conditions to be passed is that the dUL must be larger than the dUR and the geometric feature detection area $Probe_6$ is zero. If both these conditions are met, the candidate is predicted to be 6 and if not, it is predicted to be 8.

For the last number 4, the ninth set of feature conditions is compared, wherein if the values of the geometric feature $Probe_5$ and $Probe_3$-Lower are zero, the candidate can be predicted to be 4. If this set is not valid, the candidate is not an E13B number and will predicted to be not a digit number that will be classified into the x.

The following illustrates the decision path for each number:

1. Number 0 : (1)
2. Number 1 : (1), (2), (3), (4), (7), (8)
3. Number 2 : (1), (2)
4. Number 3 : (1), (2), (3), (4), (5)
5. Number 4 : (1), (2), (3), (4), (7), (9)
6. Number 5 : (1), (2), (3)
7. Number 6 : (1), (2), (3), (4), (7), (8), (10)
8. Number 7 : (1), (2), (3), (4), (5), (6)
9. Number 8 : (1), (2), (3), (4), (7), (8), (10)
10. Number 9 : (1), (2), (3), (4), (5), (6)

According the pseudo code as showing in Algorithm 1. From the above decision-making, it can be inferred that number 0 is compared once, number 2 is compared twice, number 5 is compared three times, number 3 is compared five times, numbers 1, 4, 7, and 9 need to be compared six times, and numbers 6 and 8 require seven comparisons. Therefore, it can be concluded that the maximum number of comparisons for this decision tree is seven, the minimum number of comparisons is one, and the average number of comparisons is 4.6.

D. Symbol Recognition

The difference between symbols and digits is that symbols are divided into three targets, whereas digits are not. Therefore, the comparison methods differ. The comparison method applied in this study comprises the following steps. First, the selected symbolic targets are placed in a queue of length 3 from left to right, and the relative positions of targets 1 and 2 are compared. If they are valid, the next target is compared. If not, target 1 is discarded, target 2 moves to the position of target 1, and target 3 moves to the position of target 2. This comparison model test is repeated.

The conditions for symbol comparison are detailed in Table II, and the comparison flow is illustrated in Fig. 8.

Algorithm 1: Pseudo code for proposed decision tree

```
Input: statistics features: dUL dLL dUR dLR and
geometric features : Probe0 Probe8 Probe3-Upper Probe3-Lower Probe5 Probe6
Output: classification result of input data
if number of Probe0 == 0 then
    result = "0" ;
else if density in Lower Left > density in Upper Left and density in Upper Right > density in Lower Right then
    result = "2" ;
else if number of Probe3-Lower ==0 and number of Probe5==0 then
    result = "5" ;
else if density in Upper Right > density in Lower Left then
    if density in Lower Right > density in Upper Right and number of Probe3-Upper == 0 and number of Probe3-Lower ==0 then
        result = "3" ;
    else if density in Upper Right > density of Lower Right and density Upper Left > density in Lower Left then
        result = "7" ;
    else
        result ="9" ;
else if density in Lower Right > density in Upper Right and density in Lower Left > density in Upper Left then
    if number of Probe8>0 then
        result = "1" ;
    else if density in Upper Left > density in Upper Right and number of Probe6 == 0 then
        result = "6" ;
    else
        result ="8" ;
else if number of Probe3-Lower >0 and number of Probe5==0 then
    result = "4" ;
else
    something else;
return result
```

TABLE II. Relative Feature Chart

Item	Conditions
Condition 1	The distance between Target 2 and 1 is greater than 2 and less than 10.
Condition 2	Height of Target 1 is greater than that of Target 2, and height of Target 2 is greater than that of Target 3. Additionally, the width of Target 3 is greater than that of both of Targets 1 and 2.
Condition 3	The correlation height of Target 1 is greater than that of Targets 2 and 3, and the y-axis position of Target 1 is greater than that of Targets 2 or 3.
Condition 4	The relative y-axis distance of Target 3 is less than that of Target 2, and relative y-axis distance of Target 2 is less than that of Target 1.
Condition 5	The correlation height of Target 1 is greater than that of Target 3, and the correlation width of Target 2 is greater than that of Target 3.

IV. Experimental Results

The accuracy rate of character recognition was calculated using the above-mentioned characteristics along with the decision-making process. The complete data set consisted of 24,000 E13B characters. The control experiment was conducted with the same data set in Two Layer MLP, Three Layer MPL, LetNet-5 [7], AlexNet [11] and CNN-SVM [36]. The data of this study and the control group were compared to verify the validity and performance of the proposed method.

A. Comparison Algorithms

1. Multilayer Perceptron

The control group is divided into two groups of multi-layer perceptron hidden layers; in two sets and three sets of hidden layers, respectively. The input is a 28 × 28 pixel image, and uses one applicable channel at a time. There are 784 neurons (1 × 28 × 28 input neurons); each input layer is between 0 and 1, and a pixel value between 0-255 is converted to a floating-point number between 0 and 1. Next, the 28 × 28 matrices are converted into a single layer before input, and the 28 columns are converted into a 784 one-dimensional array, referred to as the first input layer. In the first set of the two hidden layers, as shown in Fig. 9, 256 neurons are arranged in the second hidden layer, and ten neurons in the third hidden layer. The fourth layer acts as the output layer. As shown in Fig. 10, three hidden layers are arranged in the second control group, with 512 neurons in the second hidden layer, 256 neurons in the third layer, and ten neurons in the fourth layer. The fifth layer acts as the output layer. The output consists of digits ranging from 0–9, and thus, according to the softmax function, the final output contains ten neurons.

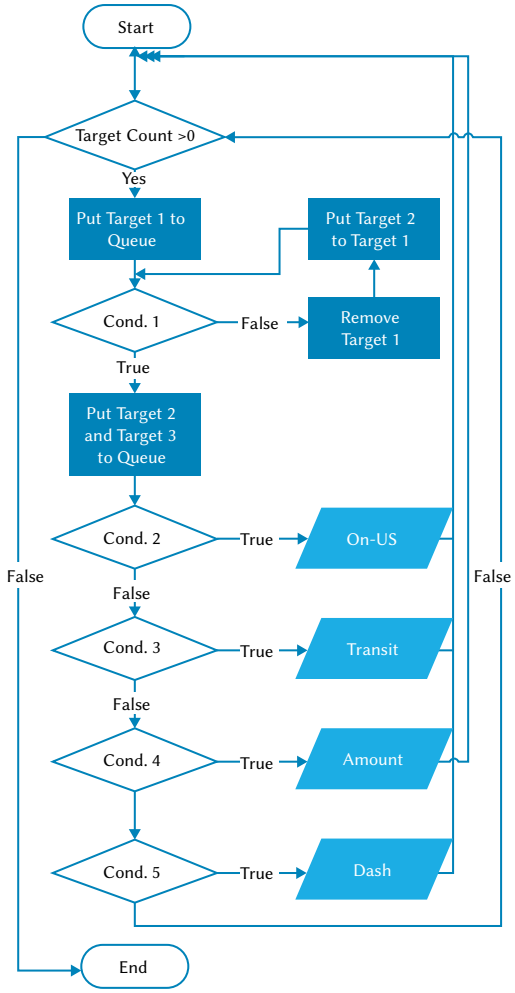


Fig. 8. Flow chart of proposed algorithm.

Neural networks mimic the chain of responses generated by the stimulation of the brain's nerve cell. Transferring neurons between each layer eventually yields an output result.

2. LeNet-5

LeNet-5 is a convolutional network algorithm proposed by LeCun in 1998 [7]. This network is also the basis of today's deep learning models. We will use this algorithm as another control group for our experiments. The first layer of LeNet-5 is a 32*32 grayscale image, implying that it is a two-dimensional array with only one channel, which is different from the three channels of an RGB image. The second layer is a 2 x 2 Max pooling layer with a stride of 2. The output of this layer is 14 x 14 x 6. The third layer consists of 16 convolutions of size 5 x 5, with a stride of 1. The fourth layer is a pooling layer that is the same as the second layer with a stride of 2. The output layer is the fifth layer with 120 convolution kernels. The sixth layer is a fully connected layer, and the hidden layer has 84 neural nodes. The last layer consists of 84 hidden nodes corresponding to 10 outputs.

3. Alexnet

Alexnet is a CNN proposed by Alex Krizhevsky in 2012 [11] and it won the ImageNet LSVRC competition in the same year. We used this algorithm as another control group for our experiments. The AlexNet architecture has eight layers and uses a total of five convolutional layers and three fully connected layers, which is deeper than the LeNet-5 model. The first to fifth layers are convolutional layers, where a Maxpooling layer of size 3 x 3 and stride of 2 is used after the first, second and fifth convolutional layers . The input layer is larger than

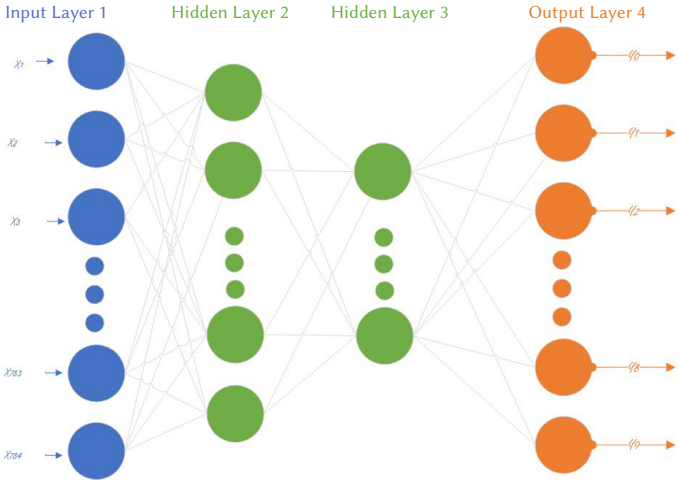


Fig. 9. Two Hidden Layers Multilayer Perceptron Structure Chart.

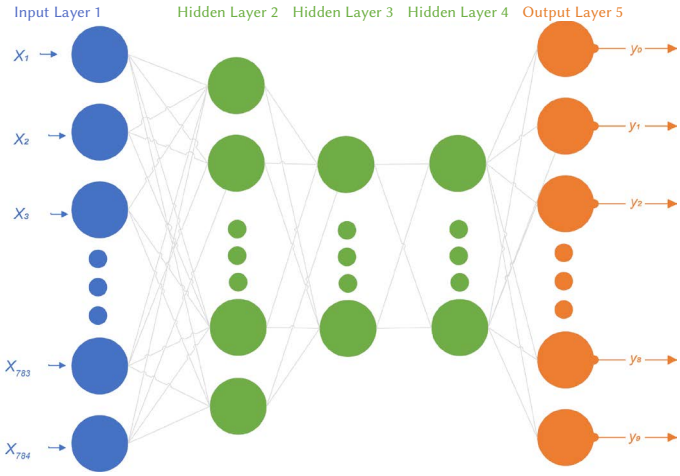


Fig. 10. Three Hidden Layer Multilayer Perceptron Structure Chart.

LeNet-5 and can input 224*224 pixel color images. Unlike LeNet-5, which adopts average pooling, the stride is smaller than the size of the mask, and 2 < 3 can repeat the inspection of features, avoiding important features being discarded during pooling, resulting in better feature calculation results. The sixth, seventh, and eighth layers are fully connected layers. Although the E13B samples in this study were not large images, this method was used as a control group to highlight the comparison of computational power and accuracy.

4. Hybrid CNN-SVM

Ahlawat and Choudhary [36] proposed the hybrid CNN-SVM architecture. As this architecture was proposed in recent years, we incorporated this architecture as a control group in our study, in addition to using MLP, LeNet-5, and Alexnet. This architecture uses CNN to extract the features of the image after two convolutions. The input, a 28 x 28 single-channel image, is taken through the first layer using a 5 x 5 filter for convolution operations. The output from this stage is six 24 x 24 feature maps, which are then input to the second layer. Alternatively, the 5 x 5 filter can be used for convolution operations to obtain sixteen 24 x 24 feature maps. A total of 576 neurons were flattened after convolution as a classification feature for the SVM.

B. Comparison Results

As shown in Table III, 24,000 characters were extracted from the 1,079 check specimens. The number of correct characters recognized in this study was 24,000, whereas the number of incorrect characters

recognized was zero. Table IV shows the results of the two-layer MLP. The recognition rate of the control group at k = 2, 5, and 10 was calculated to be 96.29% on average, while the experimental group recognition rate was 100%. This shows the superiority of the proposed method compared to the two-layer MLP model, which is trained by the artificial neural network in accordance with the accuracy of character recognition.

TABLE III. Treatment Group Accuracy Results

Test Model	Tests	Errors	Accuracy
Proposed Method	24,000	0	100%

TABLE IV. Two Layer MLP Accuracy Result

k	Total Tests	Number of Errors	Accuracy
2	24,000	63	99.74%
5	24,000	1816	92.43%
10	24,000	794	96.69%
		Average:	96.29%

Table V shows the results of the three-layer MLP. The recognition rate of the control group at k = 2, 5, and 10 was calculated to be 99.64% on an average, while the experimental group recognition rate was 100%. This shows the superiority of the proposed method compared to the three-layer MLP model, which is trained by the artificial neural network in accordance with the accuracy of character recognition.

TABLE V. Three Layer MLP Accuracy Result

k	Total Tests	Number of Errors	Accuracy
2	24,000	73	99.70%
5	24,000	125	99.48%
10	24,000	59	99.75%
		Average:	99.64%

Table VI shows the results of LeNet-5. The recognition rate of the control group at k = 2, 5, and 10 was calculated to be 99.83% on an average, while the experimental group recognition rate was 100%. This shows the superiority of the proposed method compared to the LeNet-5 model, which is trained by the artificial neural network in accordance with the accuracy of character recognition.

TABLE VI. LeNet-5 Accuracy Result

k	Total Tests	Number of Errors	Accuracy
2	24,000	120	99.5%
5	24,000	3	99.99%
10	24,000	2	99.99%
		Average:	99.83%

As shown in Table VII, the recognition rate of the control group at k = 2, 5, and 10 was calculated to be 99.99% on an average, while the experimental group recognition rate was 100%. This shows the superiority of the proposed method compared to the Alexnet model, which is trained by the artificial neural network in accordance with the accuracy of character recognition.

TABLE VII. AlexNet Accuracy Result

k	Total Tests	Number of Errors	Accuracy
2	24,000	2	99.99%
5	24,000	4	99.98%
10	24,000	3	99.99%
		Average:	99.99%

Table VIII shows the results of the CNN-SVM. The recognition rate of the control group at k = 2, 5, and 10 was calculated to be 99.99% on an average, while the experimental group recognition rate was 100%. This shows the superiority of our method compared to the CNN-SVM hybrid model, which is trained by the artificial neural network in accordance with the accuracy of character recognition.

TABLE VIII. CNN-SVM Accuracy Result

k	Total Tests	Number of Errors	Accuracy
2	24,000	3	99.99%
5	24,000	2	99.99%
10	24,000	2	99.99%
		Average:	99.99%

Table IX and Table X show a comparison between the average character recognition time of the treatment and control groups. The average recognition time per character for the treatment group was 0.03 ms, and the average character recognition time for the fastest control group was 1.60 ms per character. The speed of character recognition in the treatment group was significantly higher than that of the control group.

TABLE IX. Elapsed Time for Recognition of This Study

Test Model	Elapsed time Per Character (ms)
Proposed Method	0.03

TABLE X. Elapsed Time for Recognition of the Focus Group

Test Model	*k* = 2	*k* = 5	*k* = 10	Average
Two Layer MLP	1.54	1.58	1.67	1.60
Three Layer MLP	1.95	1.77	1.78	1.83
LeNet-5	1.79	2.73	4.08	2.87
AlexNet	66.23	63.71	68.00	65.98
CNN-SVM	1.25	2.73	4.72	2.9

V. Conclusion

Because E13B is a font for machine reading, traditional feature recognition can be used for stable and convenient font recognition. This approach reduces reliance on hardware, enhances the judgment of exceptions, and reduces the computational cost of the recognition algorithm. Moreover, the method proposed in this study is suitable for embedded platforms or thin clients. A small sample can yield a reasonably high accuracy rate for fixed fonts; the only drawback is that the training process has a long lead time, including real-time corrections of exceptional cases. The aim of this study is to determine the accuracy and speed of character recognition of E13B fonts. The so-called adequate sample eliminates the standard stamps or signature-related targets in the check, and only retrieves E13B. Additionally, filtering noise and partitioning contents were not within the scope of this study and were thus excluded. One limitation of the proposed method is that if the last decision point in the decision tree does not match the feature combination shown in Fig. 7, the output value will be x. However, as x is not a number between 0 and 9, such an output will cause errors during runtime. Using decision trees and comprehensive features has the advantage of speed, but this approach cannot output any confidential information. If there is an exception in the decision tree, it will output x. The addition of exception handling is therefore necessary but can be disadvantageous. In future research, this comprehensive feature can be used for shallow and low-dimensional deep learning. The trained model will be able to perform basic functions of accurate identification with high output accuracy. It will also be faster than general high-dimensional deep networks. The decision tree used in this study also constitutes the identification core of a dual classifier. Such an approach mitigates the drawback of decision tree results that cannot be classified correctly under exceptional conditions and enables better identification.

Acknowledgement

Plustek Inc. provided the samples for this study at no cost. We thank Bob Lin, General Manager at Plustek Inc., for his constant support, and ADView Technology for providing an Nvidia GPU for E13B model training, making this study possible. In addition, the identification framework of this study has obtained the Republic of China Patent No. M617631. The Plustek iKnow application software and SDK have been imported, which can be used by other software developers for authorization. This work was supported in part by the Ministry of Science and Technology, Taiwan, R.O.C., under the grant ID: MOST 110-2222-E-992 -006 -.

References

[1] L. Deng, "The mnist database of handwritten digit images for machine learning research [best of the web]," *IEEE signal processing magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[2] X. Corporation, "Generic micr fundamentals guide," Xerox Corporation, 2012.

[3] A. Choudhary, S. Ahlawat, R. Rishi, "A binarization feature extraction approach to ocr: Mlp vs. rbf," in *International Conference on Distributed Computing and Internet Technology*, 2014, pp. 341–346, Springer.

[4] I. B. Cruz, A. Díaz Sardiñas, R. Bello Pérez, Y. Sardiñas Oliva, "Learning optimization in a mlp neural network applied to ocr," in *Mexican International Conference on Artificial Intelligence*, 2002, pp. 292–300, Springer.

[5] A. Choudhary, R. Rishi, S. Ahlawat, "Off-line handwritten character recognition using features extracted from binarization technique," *Aasri Procedia*, vol. 4, pp. 306–312, 2013.

[6] A. F. Agarap, "An architecture combining convolutional neural network and support vector machine for image classification," *arXiv preprint arXiv:1712.03541*, 2017.

[7] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, "Gradient- based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[8] Z. Zhong, L. Jin, Z. Xie, "High performance offline handwritten chinese character recognition using googlenet and directional feature maps," in *2015 13th International Conference on Document Analysis and Recognition*, 2015, pp. 846–850, IEEE.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[10] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[11] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[12] N. Sharma, B. Kumar, V. Singh, "Recognition of off- line hand printed english characters, numerals and special symbols," in *2014 5th International Conference- Confluence The Next Generation Information Technology Summit*, 2014, pp. 640–645, IEEE.

[13] I. O. for Standardization, "Information processing — magnetic ink character recognition — part 1: Print specifications for e13b," International Organization for Standardization, 2018.

[14] Y. Yang, X. Lijia, C. Chen, "English character recognition based on feature combination," *Procedia Engineering*, vol. 24, pp. 159–164, 2011.

[15] M. Rani, Y. K. Meena, "An efficient feature extraction method for handwritten character recognition," in *International Conference on Swarm, Evolutionary, and Memetic Computing*, 2011, pp. 302–309, Springer.

[16] S. B. Moussa, A. Zahour, A. Benabdelhafid, A. M. Alimi, "New features using fractal multi-dimensions for generalized arabic font recognition," *Pattern Recognition Letters*, vol. 31, no. 5, pp. 361–371, 2010.

[17] H. Bay, T. Tuytelaars, L. V. Gool, "Surf: Speeded up robust features," in *European conference on computer vision*, 2006, pp. 404–417, Springer.

[18] L. Wang, S. Bi, X. Lu, Y. Gu, C. Zhai, "Deformation measurement of high-speed rotating drone blades based on digital image correlation combined with ring projection transform and orientation codes," *Measurement*, vol. 148, p. 106899, 2019.

[19] K. K. Shreyas, S. Rajeev, K. Panetta, S. S. Agaian, "Fingerprint authentication using geometric features," in *2017 IEEE International Symposium on Technologies for Homeland Security*, 2017, pp. 1–7, IEEE.

[20] T. Kobayashi, "Bfo meets hog: feature extraction based on histograms of oriented pdf gradients for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 747– 754.

[21] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[22] J.-M. Park, C. G. Looney, H.-C. Chen, "Fast connected component labeling algorithm using a divide and conquer technique," *Computers and Their Applications*, vol. 4, no. 20, p. 0, 2000.

[23] F. Kimura, M. Shridhar, "Handwritten numerical recognition based on multiple algorithms," *Pattern recognition*, vol. 24, no. 10, pp. 969–983, 1991.

[24] P. Singh, S. Budhiraja, "Feature extraction and classification techniques in ocr systems for handwritten gurmukhi script–a survey," *International Journal of Engineering Research and Applications*, vol. 1, no. 4, pp. 1736–1739, 2011.

[25] R. Verma, D. J. Ali, "A-survey of feature extraction and classification techniques in ocr systems," *International Journal of Computer Applications & Information Technology*, vol. 1, no. 3, pp. 1–3, 2012.

[26] D. Varshni, K. Thakral, L. Agarwal, R. Nijhawan, A. Mittal, "Pneumonia detection using cnn based feature extraction," in *2019 IEEE international conference on electrical, computer and communication technologies*, 2019, pp. 1–7, IEEE.

[27] A. Yang, X. Yang, W. Wu, H. Liu, Y. Zhuansun, "Research on feature extraction of tumor image based on convolutional neural network," *IEEE access*, vol. 7, pp. 24204–24213, 2019.

[28] G. S. Lehal, "Optical character recognition of gurmukhi script using multiple classifiers," in *Proceedings of the international workshop on multilingual OCR*, 2009, pp. 1– 9.

[29] T. Kobayashi, A. Hidaka, T. Kurita, "Selection of histograms of oriented gradients features for pedestrian detection," in *International conference on neural information processing*, 2007, pp. 598–607, Springer.

[30] S. Singh, A. Aggarwal, R. Dhir, "Use of gabor filters for recognition of handwritten gurmukhi character," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, no. 5, 2012.

[31] A. Shawon, M. J.-U. Rahman, F. Mahmud, M. A. Zaman, "Bangla handwritten digit recognition using deep cnn for large and unbiased dataset," in *2018 International Conference on Bangla Speech and Language Processing*, 2018, pp. 1–6, IEEE.

[32] V. Rajinikanth, S. Kadry, R. González-Crespo, E. Verdú, "A study on RGB image multi-thresholding using kapur/tsallis entropy and moth-flame algorithm," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 5, no. 2, 163-171, 2021.

[33] S. Acharya, A. K. Pant, P. K. Gyawali, "Deep learning based large scale handwritten devanagari character recognition," in *2015 9th International conference on software, knowledge, information management and applications*, 2015, pp. 1–6, IEEE.

[34] I. Ramadhan, P. Sukarno, M. A. Nugroho, "Comparative analysis of k-nearest neighbor and decision tree in detecting distributed denial of service," in *2020 8th International Conference on Information and Communication Technology*, 2020, pp. 1–4, IEEE.

[35] T. A. Assegie, P. S. Nair, "Handwritten digits recognition with decision tree classification: a machine learning approach," *International Journal of Electrical and Computer Engineering*, vol. 9, no. 5, pp. 4446–4451, 2019.

[36] S. Ahlawat, A. Choudhary, "Hybrid cnn-svm classifier for handwritten digit recognition," *Procedia Computer Science*, vol. 167, pp. 2554–2560, 2020.

[37] A. A. Barbhuiya, R. K. Karsh, R. Jain, "Cnn based feature extraction and classification for sign language," *Multimedia Tools and Applications*, vol. 80, no. 2, pp. 3051–3069, 2021.

[38] V. Dogra, S. Verma, N. Jhanjhi, U. Ghosh, D.-N. Le, *et al.*, "A comparative analysis of machine learning models for banking news extraction by multiclass classification with imbalanced datasets of financial news: Challenges and solutions.," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 3, 2022.

[39] M. Khari, A. K. Garg, R. G. Crespo, E. Verdú, "Gesture recognition of

RGB and RGB-D static images using convolutional neural networks.," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 7, pp. 22–27, 2019.

[40] J. D. Rodriguez, A. Perez, J. A. Lozano, "Sensitivity analysis of k-fold cross validation in prediction error estimation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 3, pp. 569–575, 2009.

Chung-Hsing Chen

C.H Chen received his master's degree at the Department of Information Management, National Sun Yat-Sen University, in 2006. Currently, he is the Director of Research and Development Department of Plustek Inc. His current research interests mainly include, network applications, embedded systems and AI image recognition.

Ko-Wei Huang

Ko-Wei Huang received his PhD from the Institute of Computer and Communication Engineering, Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan, in 2015. He is currently an Associate Professor at the Department of Electrical Engineering, National Kaohsiung University of Science and Technology, Taiwan. His current research interests mainly include data mining, deep learning, evolutionary computing, and medical image processing.

Cosine Similarity Based Hierarchical Skeleton and Cross Indexing for Large Scale Image Retrieval Using Mapreduce Framework

Zhong Qianwen

Electronic and Communication Engineering, Shangdong University, Shandong, Weihai, Huancui District (China)

Received 9 September 2020 | Accepted 23 September 2021 | Published 31 January 2023

Abstract

The imaging data in various fields like industries, institutions, medical, and so on has grown exponentially in recent years. An innovative software solution is required for the efficient management of image data. The MapReduce framework is used for large-scale image data processing. Various cross-indexing techniques are developed to transform the image into binary sequences but retrieving the image from the reducer on the feature vector results in a major challenge. Image retrieval using large-scale image databases attained major attention, where cross-indexing plays a key role in the research community. Therefore, in this research, a new method for image retrieval, named Cosine Similarity-based hierarchical skeleton and cross-indexing, is proposed to perform the retrieval process in the MapReduce framework effectively. The feature vector of the images is converted to binary sequences. The Most Significant Bit (MSB) of the binary code is used to store the images in the mapper using the cross-indexing model. The image retrieval process is achieved through the reducer based on the tanimoto similarity measure. The binary sequence for the query image is calculated based on the feature vector. The MSB bit of the binary code is matched with the MSB code of the images in the mapper to achieve the retrieval process. The proposed method effectively achieved better performance through the cross-indexing model with the usage of the feature vector. The performance of the proposed method is compared with the existing techniques using the UK bench dataset. The proposed method attains the values of 0.784, 0.729, 0.75, 31.23, 17.84secfor F1-score, precision, recall, computational cost, and computational time with the query set-1 by considering four mappers.

Keywords

Cosine Similarity, Cross Indexing, Hierarchical Skeleton, Image Retrieval, Tanimoto Similarity Measure.

DOI: 10.9781/ijimai.2023.01.008

I. Introduction

In the multimedia environment, the large-scale information processing system attracted significant attention in the research area due to the growth of information, such as videos and images [1]. However, to search an image in the huge volume of the dataset with high accuracy based on the semantic features has become a significant problem. A single image can convey more details than the usage of the number of words, such that image retrieval is considered an important topic in the research area for the past few decades [2]. One of the most challenging issues is helping the users retrieve their estimated images from the huge database [3]. Image retrieval is the process of finding the suitable image with the appropriate content or feature based on the image set or image content description. Due to the issues in human cognition and the subjectivity of image content, defining the universal and efficient image retrieval process becomes a challenging task. Therefore, the image retrieval process has become an important area in the information retrieval and the computer vision field in recent years [4]. Two different stages are used: cognitive load-based complexity and the second complexity classification to address the complexity of imager retrieval [5]. Moreover, generating digital images is rapidly increased in the advancement of network and multimedia technology. Due to the huge amount of data information, retrieving it securely and rapidly requires considerable effort [6], [7]. Secure big data transmission should require efficient sharing in all environmental conditions [8].

Due to the enormous amount of digitized image creation, the image retrieval process has become a complex issue because encryption is required while transmitting the image [9], [10]. The image sharing between different entities is done by encrypting the image [11].The Content-based image retrieval (CBIR) model is developed to perform the retrieval process, which effectively retrieves the suitable images according to the low level features, like texture, shape, and color [12]. The main aim of the CBIR is to retrieve the images from the large database with the most related visual information. It is required to analyze the image content to perform the content-based retrieval. Hence, similarity measurement and the feature representation factor are crucial elements in CBIR, but there is a challenging issue in CBIR called semantic gap [1]. CBIR is effectively used to manage the huge

* Corresponding author.
E-mail address: zhongqianwen77@gmail.com

volume of image database [13]; hence it acts as a possible solution in image retrieval. Furthermore, quick access to the large database needs an efficient and effective computing model. Therefore, the Hadoop framework is considered a suitable distributed computing scheme in the retrieval process based on the MapReduce framework. The MapReduce framework is widely used as the parallel device in the computing environment that processes the data based on peta byte and terabyte scales. Facebook, Google, and Amazon are the largest users in the MapReduce computing model, allowing the processes of the data in a distributed manner with intensive computing over different machines [14]. Due to the presence of semantic gap, the accuracy of the CBIR model was not adequate, such that the gap exists between the visual low level features, like colors and textures, and the high level image concepts are usually used by the user in the searching process [12], [15].Compressing the image before the transmission will help improve the accuracy of the image [16].

Due to the enormous growth in large-scale data technology, the application of image processing shows the characteristic features of large-scale processing technology in the image retrieval model. In general, some of the recent works are developed to acquire the fast searching time based on the large scale processing technology, such that the authors in [2] and [17] revealed to enhance the CBIR model in Hadoop. In [4], a new method is developed that focuses on the enhanced parallel K-Means algorithm based on the MapReduce scheme [18]. However, Artificial intelligence (AI) is considered a machine learning technique that attracted significant attention in the past decades [1], [19], [20]. The AI model's key objective is to allow computers to handle real-world tasks and simulate human intelligence. Minimizing the semantic gap in the CBIR model is essential in image retrieval. Some efforts are taken to reduce the gap in the machine learning methods. Especially, the deep learning methods attained significant progress in the recent years in image retrieval [21], namely deep brief network [22], deep Boltzmann machine [23], deep neural network [24], Region-based Convolutional Neural Network (R-CNN) [25], and so on. Among these methods, the deep convolutional neural network (DCNN) attained a great achievement in computer vision, like image classification, object recognition [26], and image segmentation [1], [27].

This research focuses on a new method for image retrieval based on the proposed CS-based hierarchical skeleton and cross-indexing. The features extracted using the Speeded-Up Robust Features (SURF) and CS-based hierarchical skeleton are converted to the binary sequences to perform the retrieval process effectively. The performance of the proposed method is increased with the representation of binary sequences rather than computing it through the decimal value. The cross-indexing increases retrieval performance by storing the image in the mappers using the feature vector. The feature vector enables the binarization of features to enhance the efficiency of the image retrieval process. The images retrieved from the mappers are responded to the user through the reducer in the MapReduce framework.

The contribution of the paper is image retrieval using MapReduce Framework with CS-based hierarchical skeleton and cross-indexing. The CS-based hierarchical skeleton and cross-indexing effectively enable the retrieval process by achieving better performance through the cross-indexing model. The conversion of features from the decimal to the binary sequences increases the performance of feature vector, which effectively makes the cross indexing model store the images in the mappers. The tanimoto similarity measure provides robust and accurate results in the perspective of retrieving the images.

The paper is organized as follows: Section II describes the motivation of the image retrieval model. Section III elaborates the proposed CS-based hierarchical skeleton and cross-indexing for image retrieval using the MapReduce framework. Section IV describes the results and discussion of the proposed image retrieval method, and finally, section V concludes the paper.

II. Motivation

In this section, the image retrieval method's motivation is discussed using various existing retrieval techniques, which motivates the researchers to develop a new model based on the CS-based hierarchical skeleton and cross-indexing.

A. Literature Survey

Various existing image retrieval methods are surveyed in this section: Bai, C *et al.* [1] developed a DCNN model for retrieving the large-scale images. It used the features of convolutional layers and attained a better extraction process. It was highly suitable for a large volume image database for mapping the high-dimensional feature vectors into the binary codes. The performance of the online retrieval processing time was very less. Sakr, N.A *et al.* [2] introduced a Chain Clustering Binary search tree (CC-BST) algorithm to model the visual statements to represent the features of an image. It was an effective solution in the retrieval mechanism for high-dimensional features. It offered a significant enhancement in the time cost factor than other competitive systems. However, it utilized more time to perform the retrieval process. Gao, X *et al.* [3] introduced a progressive image retrieval model to guarantee image quality. This model was parallelized using the MapReduce framework and attained enhanced performance with the MapReduce framework. It failed to use the adaptive learning methods. Cao, J *et al.* [4] modeled a parallel *k*-Means algorithm for selecting the cluster center in the image retrieval process. It maximized retrieval accuracy and reduced the overhead of retrieval time. To expand the node number in the Hadoop distributed platform, and the adjustment of relevant parameters used for retrieval in the system was not considered.

Xibing, S *et al.* [28] developed a MapReduce-based remote sensing image retrieval method to increase the efficiency of the retrieval process. It accurately retrieved the remote sensing image and attained better retrieval efficiency and accuracy. However, the node's information processing capacity was not effectively used to enhance the data efficiency at each node. Meng, Z [29] developed a remote sensing image retrieval algorithm. It effectively extracted the texture and the color features. It increased the retrieval accuracy and the efficiency of image retrieval. It failed to balance the system load. Mezzoudj, S *et al.* [18] developed a Content-Based Image Retrieval System using Spark (CBIR-S). It attained better performance with the spark. It failed to use the efficient model in terms of accuracy with larger datasets and clusters. Li, X *et al.* [30] introduced a parallel data processing model. It attained better performance in the single node system at both the capability and the retrieval speed of dealing with the huge volume of data. The accuracy of the searching process was effective only for few features.

B. Challenges

Some of the challenges associated with the image retrieval methods are explained as follows:

- Due to the various manners of viewing the image among the computer and human, there exists a gap in the CBIR approach. As no direct link is established between the low and high-level features, the semantic gap arises. However, it is a complex issue to solve the gap problem [1].
- With the enormous growth in the large-scale multimedia technology on the internet, especially images, constructing the CBIR system in the large scale environment becomes a challenging issue, as it used a large execution time [18].

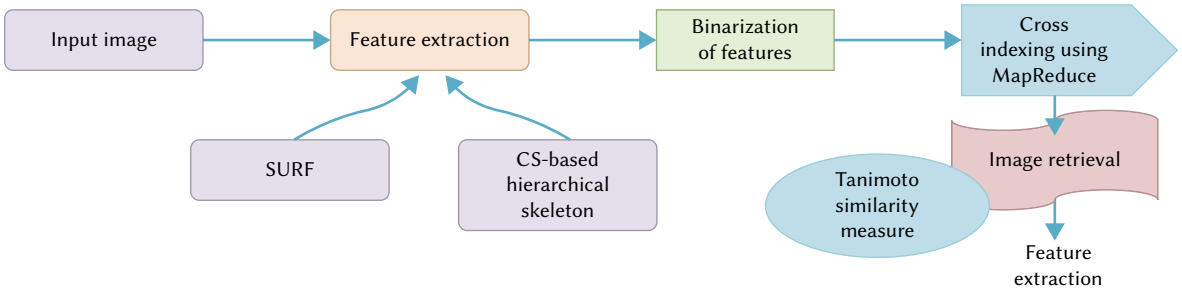


Fig. 1. Schematic diagram of the proposed CS-based hierarchical skeleton and cross-indexing.

- When a new image query is given to retrieve the specified object from the database, matching the image with the large pair of image sets poses a significant challenge [31].
- The explosive growth in the images also faces numerous challenges. One of the key challenges is helping the user retrieve their expected contents from the large volume database [3].
- The CCBST model developed in [2] was highly suitable to retrieve the images present at the top level based on the user query, but categorizing the features remains a crucial challenge in computer-based applications.

III. Proposed Cosine Similarity-based Hierarchical Skeleton and Cross-indexing for Image Retrieval Using MapReduce Framework

Image retrieval is a major consequence in the research domain to specify images based on the feature vector. Even though various image retrieval methods are available, defining the feature-based image representation poses a major challenge. Hence, a new method named CS-based hierarchical skeleton and cross-indexing is proposed in this research to solve the above issues. Initially, the input image is passed to the feature extraction stage, where the features are effectively extracted using the SURF and CS-based hierarchical skeleton. After extracting the features, the binarization of the feature vector is performed on the extracted features. Then, the binarization of the feature vector is given as the input to cross-indexing. The cross-indexing is performed in the mapper based on the binary sequences of the feature vector. Finally, the retrieval process is carried out through the reducer in the MapReduce framework based on the tanimoto similarity measure. Fig. 1 shows the block diagram of the proposed cross-indexing method.

A. Get the Input Image

At first, the input image used to perform the retrieval process using the MapReduce framework is collected from the dataset. Equation (1) refers to the database α with n number of images.

$$\alpha = \{M_1, M_2, \ldots M_j, \ldots M_n\} \tag{1}$$

where, α represents the database, and M denotes the images, M_n represents the total number of images.

B. Feature Extraction

Feature extraction is the process to reduce the dimensionality value through which the original input image is reduced to a manageable form. Feature extraction is highly important in the image retrieval mechanism to minimize the number of resources required to process the image without losing the relevant or important information. Feature extraction helps to minimize redundant values, which facilitates the speed of the retrieval process. The input image M_j is selected to perform the feature extraction process using SURF and CS-based hierarchical skeleton.

i) ***SURF***: The input image M_j is applied to the feature extraction module named as SURF, where the features from the image are effectively extracted and is denoted as f_1. SURF [32] uses the determinants of hessian matrices for locating the significant points of images based on the location and scale. Here, the dominant orientation is determined by computing the sum of all the responses that lies in the sliding orientation window with an angle of 60 degrees. SURF effectively extracts the robust features locally through the hessian matrix and the distribution-based descriptor. It uses the hessian detector to subtract the pyramid layers. Hessian detector identifies the interesting points or interest features for modifying the viewpoints. Here, the scale space is divided into various octaves and levels to achieve the scale invariance by examining the interesting points. The scale points are used to construct the pyramid levels based on the sub-sampling and Gaussian kernels. Non-maximal suppression of hessian matrices is the heart of SURF, which approximates the kernel with the rectangular box named as box filter. SURF is computationally faster and simpler without losing the performance. It is robust and stable because it has Hessian based detector. Moreover, surf uses only 64 dimensional vector. Equation (2) refers to the features extracted using the SURF based on the hessian matrix.

$$f_1 = X(y,z) = \begin{bmatrix} U_{aa}(y,z) & U_{ab}(y,z) \\ U_{ab}(y,z) & U_{bb}(y,z) \end{bmatrix} \tag{2}$$

where, $X(y, z)$ denotes the Hessian matrix, $U_{aa}(y, z)$ and $U_{ab}(y, z)$ denotes the convolution of the image with the second-order derivative of Gaussian $\lambda(z)$. Equations (3) and (4) define the terms $U_{aa}(y, z)$ and $U_{ab}(y, z)$.

$$U_{aa}(y,z) = J(p) * \frac{\partial^2 g}{\partial p^2}\lambda(z) \tag{3}$$

$$U_{ab}(y,z) = J(p) * \frac{\partial^2 g}{\partial p^2}\lambda(z) \tag{4}$$

$J(p)$ denotes the integral image and $p = (p, g)^T$ is used to store the sum of all the pixels in the rectangular area. Here, f_1 it represents the features extracted using SURF. Accordingly, the features extracted using SURF are specified in the matrix form as depicted in Fig. 2.

2	10	12	8	7
4	8	3	15	20
25	30	31	17	35
4	9	6	21	37
40	39	18	15	25

Fig. 2. Features extracted using SURF.

The output obtained from the SURF is in the form of a matrix and the pixel values are specified in the decimal format.

ii) ***CS-based hierarchical skeleton***: The input image M_j is applied to the CS-based hierarchical skeleton feature extraction module in order to extract the features, which is further used to achieve the image retrieval process. Moreover, the features extracted by the CS-based hierarchical skeleton feature extraction module are represented as f_2. CS-based hierarchical skeleton [33] is highly effective in extracting the features from the image. The CS-based hierarchical skeleton is derived by inheriting CS features with the hierarchical skeleton. The major advantage behind the CS-based hierarchical skeleton is that it provides additional information to the image retrieval process through which the accuracy of the retrieval mechanism is enhanced. It does not require any extra computational cost and effectively captures the topological and geometric features at various levels based on the skeleton pruning. The CS-based hierarchical skeleton takes the benefit of skeleton pruning, which eliminates the skeleton branches of insignificant parts of the shape. Equation (5) refers to the features extracted using the CS-based hierarchical skeleton.

$$f_2 = A(q_1, q_2) = A_1(q_1, q_2) + A_2(q_1, q_2) \tag{5}$$

Equations (6) and (7) define the terms $A_1(q_1, q_2)$ and $A_2(q_1, q_2)$.

$$A_1(q_1, q_2) = \frac{\mu(q_1,q_2)k(q_1)k(q_2)}{k(q_1)+k(q_2)} \tag{6}$$

$$A_2(q_1, q_2) = k(q_1).k(q_2)\,cos\big(\mu(q_1, q_2)\big) \tag{7}$$

where, $\mu(q_1, q_2)$ denotes the angle of corner location of q_1 and q_2, k denotes the length function, and f_2 represents the features extracted using CS-based hierarchical skeleton. The above equation (6) is modified with the CS measure based on the Cosine of two non-zero vectors using Euclidean dot product. Accordingly, the features extracted using a CS-based hierarchical skeleton are specified in the matrix form as represented in Fig. 3.

0	1	0	0	1
1	1	0	0	1
1	0	1	1	1
0	1	1	1	0
1	0	1	0	1

Fig. 3. Features extracted using CS-based hierarchical skeleton.

The output features obtained from the CS-based hierarchical skeleton are in the form of a matrix, and the pixel values are specified in the binary format.

C. Binarization of Features

The binarization process is required in the image retrieval process to enhance the retrieval accuracy in the MapReduce framework. Binarization is a tool in the image retrieval process for specifying the interested region image and its background. It is converting the pixel values of the image into the binary equivalent vector form to differentiate the interesting regions. The binarization process is performed by applying the 3×3 non-overlapping windowing mechanism to the feature matrices. In the binarization process, the pixel values are separated into two binary forms as either '0' or '1' based on the values of the Hessian matrix. The output features obtained from the SURF and CS-based hierarchical skeleton are allowed to perform the binarization process to generate the binary vector for the pixel values. The output features extracted from the SURF are in the form of a matrix with decimal pixel values. Hence, it is required to convert the matrix from the decimal to the binary format. The binarization of features will generate the binary vector by applying the 3×3 non-overlapping window to the binary form matrix. Here, the output obtained from the SURF is the decimal value matrix, while the output obtained from the CS-based hierarchical skeleton is the binary form matrix.

Let us first convert the output features obtained using SURF from the decimal form matrix into the binary form matrix by averaging the neighborhood pixels. Let us consider the first decimal pixel value '2' from Fig.2, and form a 3×3 matrix using the pixel's neighboring pixel value '2', which is depicted in Fig. 4.

0	0	0
0	2	10
0	4	8

Fig. 4. A 3×3 matrix formation for pixel value '2'.

Let us select the pixel value '2' and take the average of neighborhood pixel values to generate a decimal form matrix. For the pixel value '2', the value obtained by averaging the neighborhood pixels is 2.75, compared with the original pixel value '2'. When the original pixel value is smaller than the pixel value obtained by averaging the neighborhood pixels, then the binary value placed in the binary form matrix is '0'. If the original pixel value is greater than the pixel value obtained by averaging the neighborhood pixels, then the binary value '1' is placed in the binary form matrix. Similarly, the entire pixel values present in the matrix of Fig. 2 are converted into the binary form matrix as depicted in Fig. 5.

0	1	1	1	1
1	0	1	0	1
1	0	0	1	0
0	1	1	0	0
1	0	0	1	1

Fig. 5. Binary representation of the features extracted using SURF.

The binary form of the features extracted from the CS-based hierarchical skeleton and SURF is represented in Fig.3, and Fig.5, respectively. From the binary matrix, the binary vector of the pixel values is computed by applying the 3×3 non-overlapping windowing process.

i) ***Compute the binary vector for SURF features***: The binary matrix generated for the features obtained using SURF is represented in Fig. 5. The binary matrix value is converted into the binary vector by applying the 3×3 non-overlapping windowing process. Let us form a 3×3 binary matrix including all the pixel values from Fig. 5 and apply the 3×3 non-overlapping windowing process to generate the binary vector, which is shown in Fig. 6.

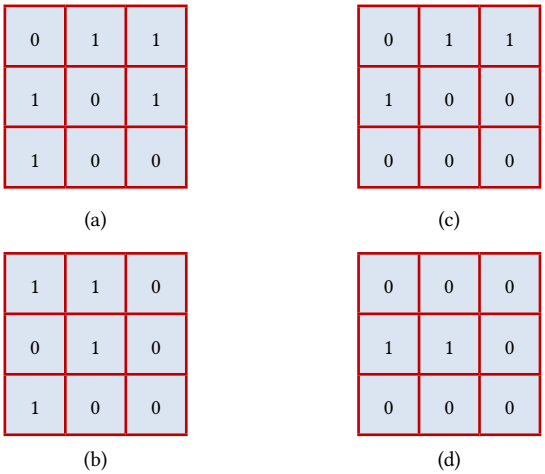


Fig. 6. 3×3 binary matrix formation for SURF features.

The binary matrix representation for the SURF-based features is partitioned into various 3×3 binary matrices and is represented in Fig. 6 a), Fig. 6 b), Fig. 6 c), and Fig. 6 d), respectively. The binary values are grouped as 011101100, 110010100, 011100000, and 000110000 and are converted into the decimal form. The binary values present in each of the 3×3 matrix is converted into the decimal format and is specified as 236 for Fig.6 a), 404 for Fig.6 b), 224 for Fig. 6 c), and 48 for Fig. 6 d), respectively. Therefore, the binarization of the feature vector is formed using the decimal values and is represented in Fig. 7.

Fig. 7. Binarization of feature vector based on SURF features.

ii) ***Compute the binary vector for CS-based hierarchical skeleton features***: The binary matrix generated for the features obtained using the CS-based hierarchical skeleton is depicted in Fig. 3. Here, the binary vector for the pixel values is generated by applying the 3×3 non-overlapping windowing process to the binary matrix. Let us form the 3×3 binary matrix including all the pixel values from Fig. 3 and apply the 3×3 non-overlapping windowing process to generate the binary vector, which is represented in Fig. 8.

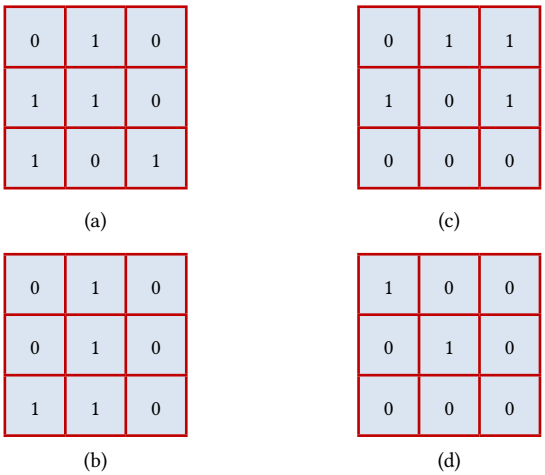


Fig. 8. 3×3 binary matrix formation for CS-based hierarchical skeleton features.

The binary matrix representation for CS-based hierarchical skeleton features is partitioned into various 3×3 binary matrices and is represented in Fig. 8 a), Fig. 8 b), Fig. 8 c), and Fig. 8 d), respectively. The binary values are grouped as 010110101, 010010110, 011101000, and 100010000 and are converted into the decimal form. The binary values present in each of the 3×3 matrix is converted into the decimal format and is specified as 181 for Fig. 8 a), 150 for Fig. 8 b), 232 for Fig. 8 c), and 272 for Fig. 8 d), respectively. Therefore, the binarization of feature vector is formed using the decimal values and is represented in Fig. 9.

Fig. 9. Binarization of feature vector based on CS-based hierarchical skeleton features.

Finally, the resultant binarization of feature vector obtained based on the SURF features is added with the resultant binarization of feature vector obtained using the CS-based hierarchical skeleton features in order to perform the cross-indexing process using MapReduce framework. Fig. 10 portrays the binarization of the feature vector.

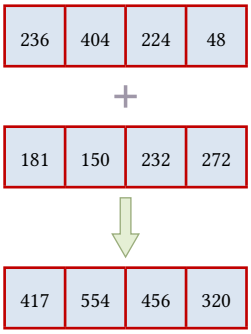


Fig. 10. Binarization of feature vector.

By adding the feature vector of SURF-based features with the feature vector of the CS-based hierarchical skeleton, the feature vector's final output is generated based on the decimal representation. Finally, the decimal form of the feature vector is converted into the binary format and is denoted as P. P specifies the binary equivalent of the decimal value H; $(H = 417 + 554 + 456 + 320)$.

D. Proposed Cross Indexing Based on the Feature Vector

Cross indexing is the process of transforming the features into binary codes to simplify the retrieval process. The features of the images are indexed with the binary value to make the retrieval process more effective. Most of the existing techniques failed to perform the mapping process to save the images in the mapper. The cross-indexing model [34] executes the binarization of the feature vector to make the image retrieval process highly robust and effective. The output of the binarized feature vector is used to perform cross-indexing based on the MapReduce framework. The feature vector for the input image M_j is denoted as P, which specifies the binary equivalent of H. The MSB of P is considered to record the image in the mappers. Based on the MSB of the features, the input images are placed in the mappers simultaneously.

The proposed CS-based hierarchical skeleton and cross-indexing effectively perform the image retrieval process based on the MSB code of features. Hence, it is required to transform the feature vector in the respective binary equivalent. Saving the images in the mappers based on the MSB code helps increase the retrieval process's efficiency, which further increases the overall system performance. When the MSB bits are selected to two bits as, 00, 01, 10, and 11, it requires four different mappers, namely 00 in mapper 1, 01 in mapper 2, 10 in mapper 3, and 11 in mapper 4, to save the images. When the MSB bits are selected as 000, 001, 010, 011, 100, 101, 110, and 111 by considering three bits, then it requires eight different mappers, such as 000 in mapper 1, 001 in mapper 2, 010 in mapper 3, 011 in mapper 4, 100 in mapper 5, 101 in

mapper 6, 110 in mapper 7, and 111 in mapper 8 for saving the images. Hence, based on the MSB bits of the binary feature vector, the images are stored in the mappers. For example, if the input image and their MSB bits of the image are given as 10, then the image will be stored in mapper 3. Similarly, the MSB bits for all the incoming images are verified based on the MSB feature code. The images are stored in the mapper through the MapReduce framework.

E. Image Retrieval Using MapReduce Framework

Image retrieval is the process of extracting the images based on the feature vector. The binary sequence for each image is computed and stored in the mapper based on the MSB code of the features. The retrieval process is carried out using the reducer in the MapReduce framework. The MapReduce framework consists of two different tasks, namely mapper and reducer. The images are stored in the mapper based on the MSB bits of the feature vector, whereas the retrieval process is performed using the reducer. Based on the binary sequence, the images are stored in different mappers. Similarly, there exist various reducers through which the retrieval process can be achieved. The number of reducers is equivalent to the number of mappers present in the image retrieval process. When the user sends the request to retrieve the mapper's image, the query image will be transformed to the vector form by considering the binary sequences. The process of converting the query image into the binary sequence is done using the tanimoto measure. Equation (8) refers to the tanimoto measure.

$$S = \frac{\sum_{t=1}^{m} X_t R_t}{\sum_{t=1}^{m} X_t^2 + \sum_{t=1}^{m} R_t^2 - \sum_{t=1}^{m} X_t R_t} \quad (8)$$

where, S represents the tanimoto measure, X_t and R_t represents the feature vector of query image and mapper image, respectively. For each user query image, the binary sequence of the feature vector is calculated based on the tanimoto similarity measure. The MSB code of the binary sequences is selected as either two-bit or three-bit sequence, and the MSB code is verified with the mapper. As the images are saved in the mapper based on the MSB code, it must convert the query image into a binary representation. When the MSB code of the query image is verified with the mapper's MSB code, then the image is retrieved through the reducer in the MapReduce framework. When the query image is matched with the mapper 3, then the image in mapper in mapper 3 is retrieved through reducer 3. If the query image is matched with the image present in mapper 4, then the matched image is retrieved through reducer 4, respectively. Fig. 11 portrays the image retrieval process based on the feature vector using the MapReduce framework.

In Fig. 11, the user sends the query image to the mapper in order to retrieve the matched image result. The binary sequence for the user's query image is calculated in the MapReduce framework based on the feature vector. However, the MSB code of the binary sequence is 01. The request is to the mapper 2, where the matched results are gathered and retrieved through reducer 2 using the tanimoto similarity measure.

IV. Results and Discussion

The proposed method is evaluated with respect to the existing methods based on precision, F-measure, and recall parameters. The analysis is done by employing two images taken from the UK bench dataset.

A. Experimental Setup

The experimentation is done in MATLAB tool using PC with 4 GB RAM, Windows 10 OS, and Intel I3 processor. The proposed method employs two different images to analyze the performance.

1. Description of Datasets

The UK bench dataset [35] is designed by Nister and Stewenius. The dataset is a collection of several research domains, which include computer vision and image retrieval. The dataset is published in 2006 with open source_media collection.

2. Performance Metrics

The Proposed CS-based hierarchical skeleton system's performance, the metrics like Precision, Recall, F1-score, computational cost, and computational time have been considered.

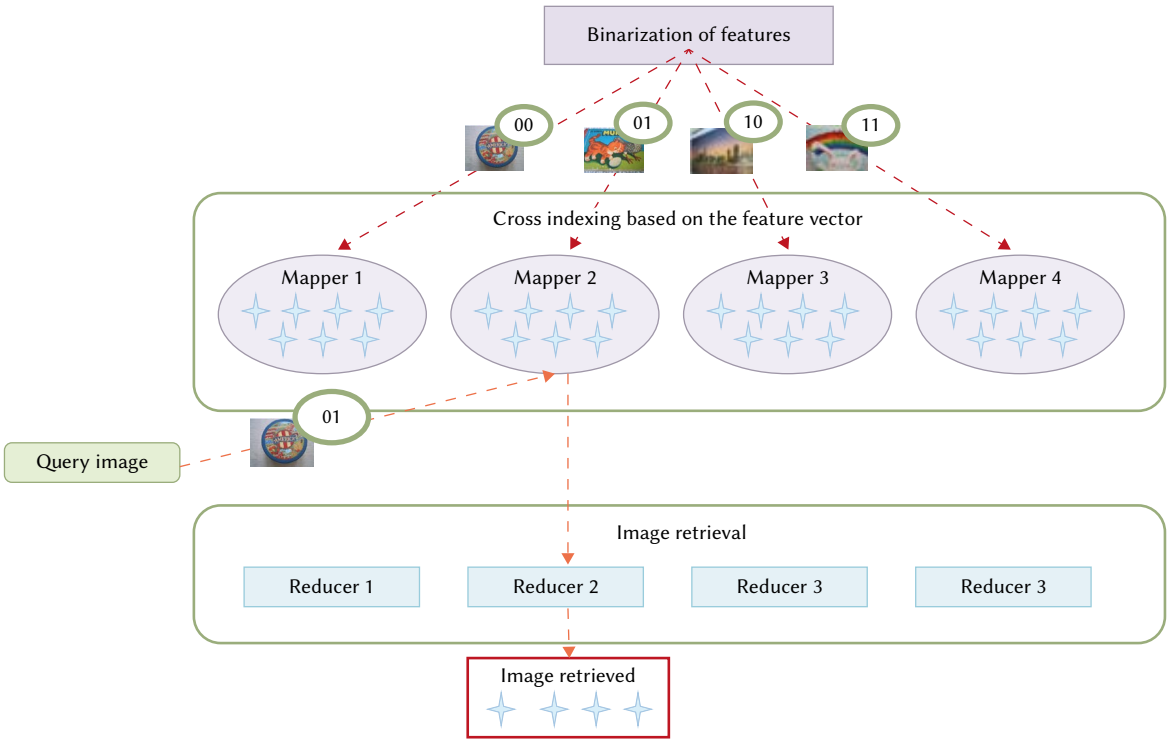


Fig. 11. Image retrieval based on the feature vector using MapReduce framework.

Fig. 12. Experimental results of proposed CS based hierarchical skeleton using a) Original image b) Extracted SURF feature c) Hierarchical skeleton image d) Query image e) Retrieved images from four reducers.

3. Comparative Methods

The analysis is made using comparative methods like Deep convolution neural network (DCNN) [1], Progressive image retrieval [3], MKSIFT+ Cross indexing [34], and Proposed CS-based hierarchical skeleton.

B. Experimental Results

This section elaborates the analysis of performance using images taken from UK bench dataset. Fig. 12 represents the experimental results of the proposed method using two different input images. Fig. 12a) represents two input image obtained from the UK bench dataset. Fig. 12b) represents the image in which the SURF features are extracted using the provided input image. Fig. 12c) portrays the hierarchical skeleton feature extracted from the input image. The query image is provided for retrieving required images, which is represented in Fig. 12d). Finally, Fig. 12e) illustrates the retrieval images obtained from the four sets of reducers.

C. Comparative Analysis

The comparison between proposed CS based hierarchical skeleton and existing techniques is done based on Recall, F-measure, and Precision using two images. The methods are analyzed using four and eight mappers using two query sets.

1. Analysis Using Four Mappers

The analysis of methods using four mappers with two query sets is elaborated in the subsections. The effectiveness of proposed method is evaluated using precision, recall and F1-Score parameters.

a) Based on Query Set-1

Fig. 13 presents the analysis of methods using four mappers considering query set-1 in terms of F1-score, precision, and recalls parameters. The analysis based on F1-Score using query set-1 is illustrated in Fig. 13a). When the number of retrieval is 2, the corresponding F1-Score values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 0.664, 0.719, 0.720, and 0.778, respectively. Likewise, for 4 retrievals, the corresponding F1-Score values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 0.721, 0.686, 0.727, and 0.784, respectively. The analysis based on precision parameter using query set-1 is illustrated in Fig. 13b). When the number of retrieval is 2, the corresponding precision values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 0.742, 0.764, 0.776, and 0.816, respectively. Likewise, for 4 retrievals, the corresponding precision values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 0.646, 0.650, 0.660, and 0.729, respectively. The analysis based on recall parameter using query set-1 is illustrated in Fig. 13c). When the number of retrieval is 2, the corresponding recall values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 0.55, 0.6, 0.6, and 0.65, respectively. Likewise, for 4 retrievals, the corresponding recall values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 0.7, 0.65, 0.7, and 0.75, respectively. The analysis based on recall parameter

using query set-1 is illustrated in Fig. 13d). When the number of retrieval is 2, the corresponding computational cost values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 34.870, 32.985, 33.931, and 32.772, respectively. Likewise, for 4 retrievals, the corresponding recall values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 49.185, 46.814, 48.585, and 46.285, respectively. The analysis based on recall parameter using query set-1 is illustrated in Fig. 13e). When the number of retrieval is 2, the corresponding computational time values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 30.059, 27.238, 28.403, and 26.189, respectively. Likewise, for 4 retrievals, the corresponding recall values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 34.597, 32.512, 33.909, and 29.9830, respectively.

b) Based on Query Set-2

Fig. 14 presents the analysis of methods using four mappers considering query set-2 in terms of F1-score, precision, and recalls parameters. The analysis based on F1-Score using query set-1 is illustrated in Fig. 14a). When the number of retrieval is 2, the corresponding F1-Score values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 0.648, 0.638, 0.648, and 0.707, respectively. Likewise, for 4 retrievals, the corresponding F1-Score values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 0.580, 0.562, 0.579, and 0.634, respectively. The analysis based on precision parameter using query set-1 is illustrated in Fig. 14b). When the number of retrieval is 2, the corresponding precision values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 0.680, 0.626, 0.682, and 0.735, respectively. Likewise, for 4 retrievals, the corresponding precision values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 0.575, 0.527, 0.570, and 0.635, respectively. The analysis based on recall parameter using query set-1 is illustrated in Fig. 14c). When the number of is 2, the corresponding recall values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 0.484, 0.517, 0.551, and 0.556, respectively. Likewise, for 4 retrievals, the corresponding recall values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 0.591, 0.595, 0.596, and 0.727, respectively. The analysis based on computational cost parameter using query set-1 is illustrated in Fig. 14d). When the number of is 2, the corresponding precision values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 41.615, 33.598, 32.079, and 31.238, respectively. Likewise, for 4 retrievals, the corresponding precision values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 49.680, 42.376, 42.356, and 37.742, respectively. The analysis based on computational time parameter using query set-1 is illustrated in Fig. 14e). When the number of retrieval is 2, the corresponding precision values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 21.540, 16.656, 18.112, and 15.838, respectively. Likewise, for 4 retrievals, the corresponding precision values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 22.942, 21.019, 22.481, and 19.505, respectively.

2. Analysis Using Eight Mappers

The analysis of methods using eight mappers with two query sets is elaborated in subsections. The efficiency of proposed method is evaluated with precision, recall and F1-Score measures.

a) Based on Query Set-1

Fig. 15 presents the analysis of methods using eight mappers considering query set-1 in terms of F1-score, precision, and recalls parameters. The analysis based on F1-Score using query set-1 is illustrated in Fig. 15a). When the number of retrieval is 2, the corresponding F1-Score values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 0.619, 0.675, 0.681, and 0.758, respectively. Likewise, for 4 retrievals, the corresponding F1-Score values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 0.676, 0.648, 0.686, and 0.735, respectively. The analysis based on precision parameter using query set-1 is illustrated in Fig. 15b). When the number of retrieval is 2, the corresponding precision values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 0.670, 0.611, 0.650, and 0.734, respectively. Likewise, for 4 retrievals, the corresponding precision values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 0.557, 0.491, 0.550, and 0.618, respectively. The analysis based on recall parameter using query set-1 is illustrated in Fig. 15c). When the number of retrieval is 2, the corresponding recall values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 0.519, 0.570, 0.597, and 0.629, respectively. Likewise, for 4 retrievals, the corresponding recall values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 0.693, 0.620, 0.658, and 0.705, respectively. The analysis based on computational cost parameter using query set-1 is illustrated in Fig. 15d). When the number of retrieval is 2, the corresponding precision values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 41.615, 33.598, 32.079, and 31.238, respectively. Likewise, for 4 retrievals, the corresponding precision values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 49.680, 42.376, 42.356, and 37.742, respectively. The analysis based on computational time parameter using query set-1 is illustrated in Fig. 15e). When the number of retrieval is 2, the corresponding precision values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 28.804, 27.858, 28.516, and 27.510, respectively. Likewise, for 4 retrievals, the corresponding precision values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 34.171, 32.572, 33.308, and 31.160, respectively.

b) Based on Query Set-2

Fig. 16 presents the analysis of methods using eight mappers considering query set-2 in terms of F1-score, precision, and recalls parameters. The analysis based on F1-Score using query set-1 is illustrated in Fig. 16a). When the number of retrieval is 2, the corresponding F1-Score values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 0.648, 0.638, 0.648, and 0.707, respectively. Likewise, for 4 retrievals, the corresponding F1-Score values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical

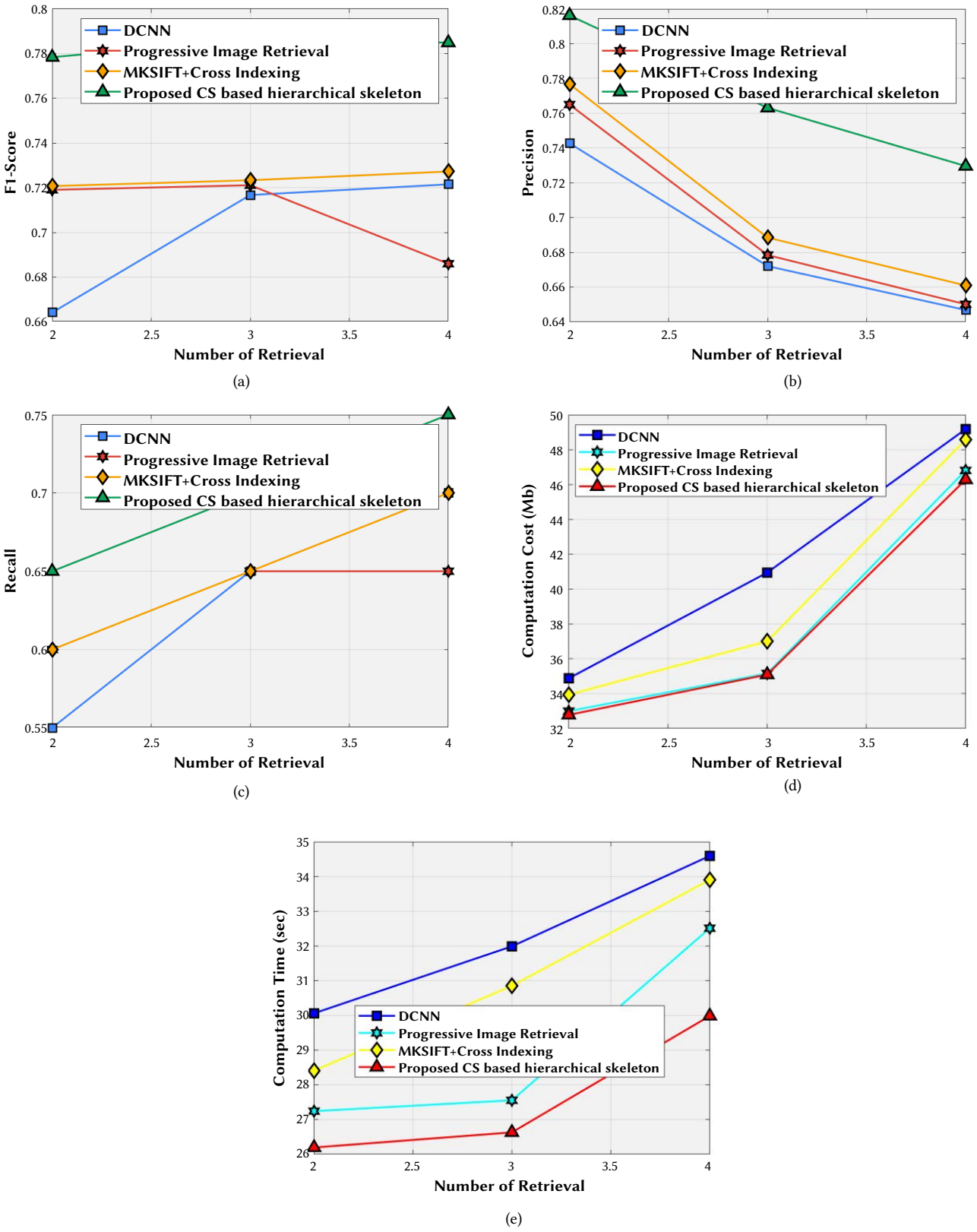


Fig. 13. Comparative analysis of methods using query set-1 with a) F1-Score b) Precision c) Recalld) Computational cost(Mb) and e) Computational Time (sec).

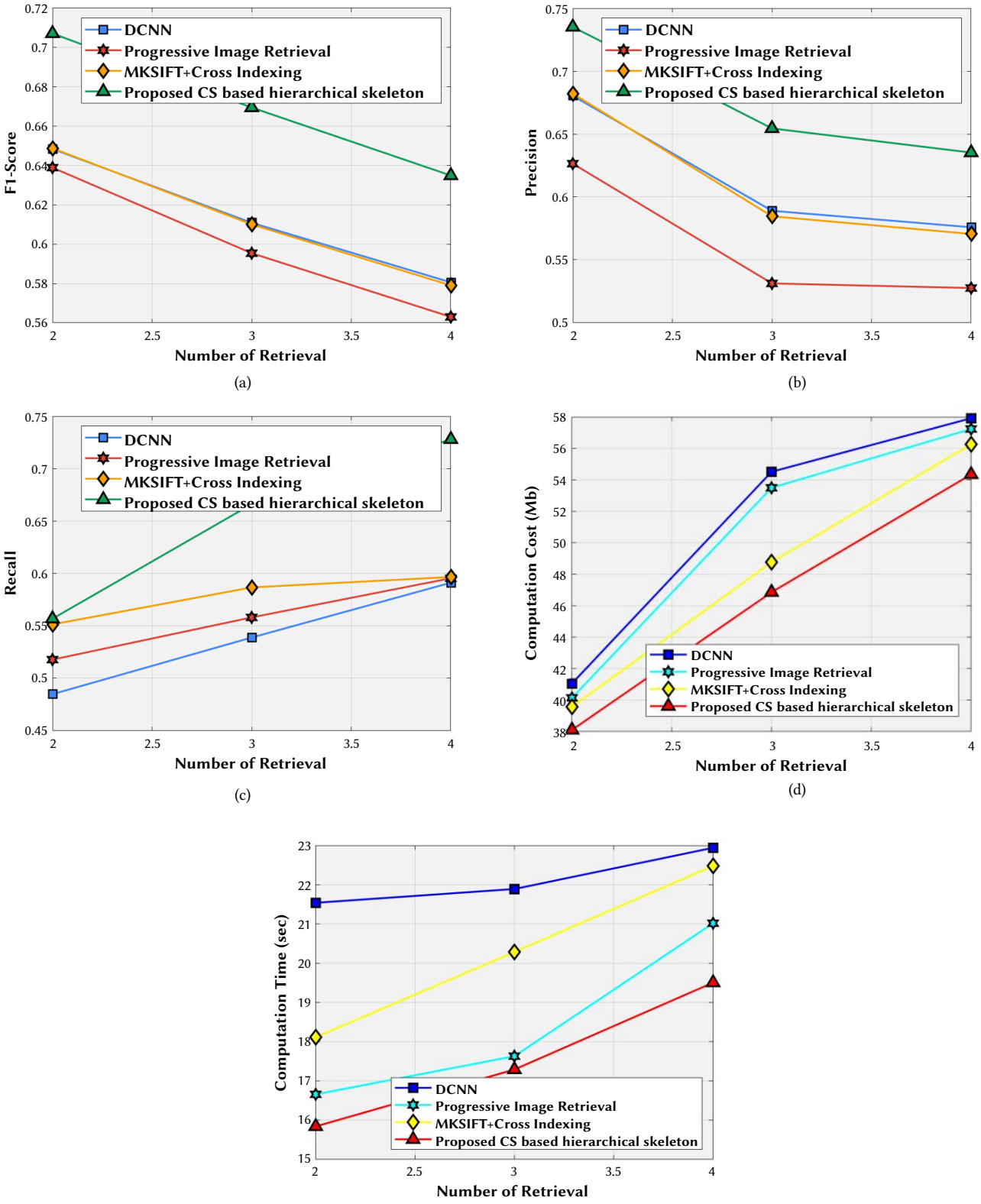


Fig. 14. Comparative analysis of methods using query set-2 with a) F1-Score b) Precision c) Recall d) Computational cost (Mb) and e) Computational Time (sec).

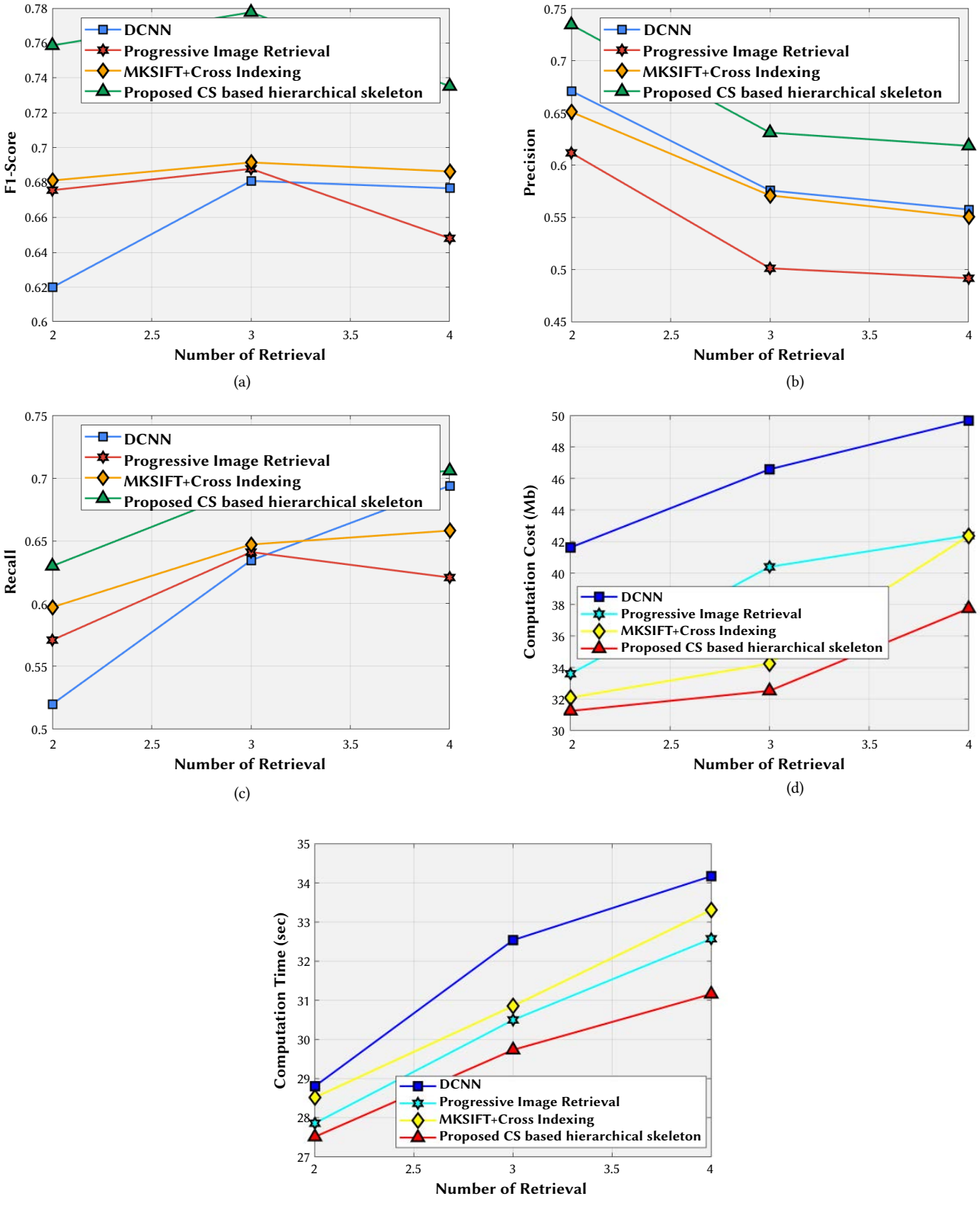


Fig. 15. Comparative analysis of methods using query set-1 with a) F1-Score b) Precision c) Recall d) Computational cost (Mb) and e) Computational Time (sec).

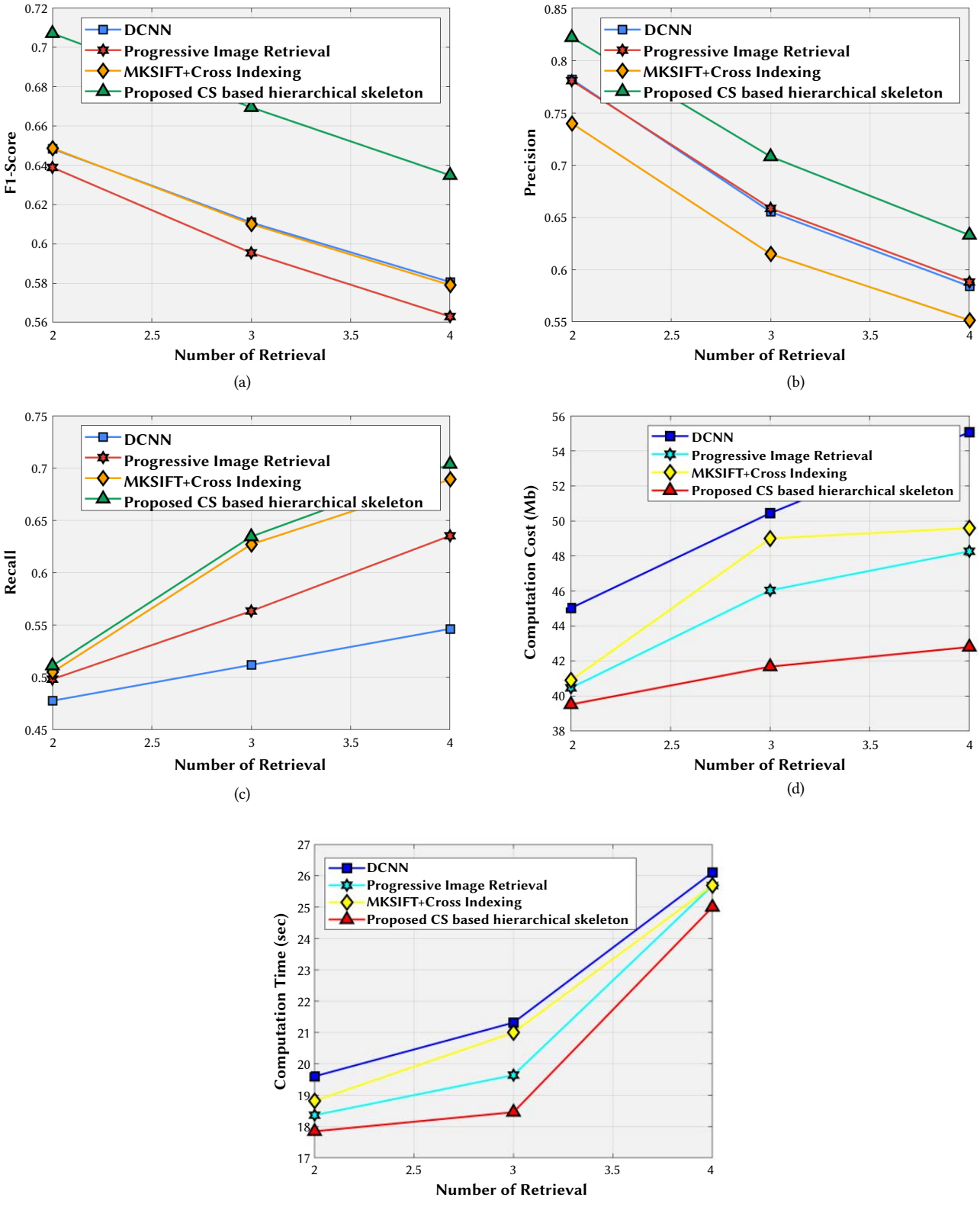


Fig. 16. Comparative analysis of methods using query set-2 with a) F1-Score b) Precision c) Recall d) Computational cost (Mb) and e) Computational Time (sec).

TABLE I. Comparative Analysis

Mappers	Query set	Metrics	DCNN	Progressive image retrieval	MKSIFT + Cross indexing	Proposed CS-based hierarchical skeleton
Using 4 mappers	Query set-1	F1-score	0.721	0.686	0.727	**0.784**
		Precision	0.646	0.65	0.66	**0.729**
		Recall	0.7	0.65	0.7	**0.75**
		CC	34.87	32.98	33.93	**32.77**
		CT	30.05	27.23	28.4	**26.18**
	Query set-2	F1-score	0.58	0.562	0.579	**0.634**
		Precision	0.575	0.527	0.57	**0.635**
		Recall	0.591	0.595	0.596	**0.727**
		CC	41.04	40.13	39.56	**38.09**
		CT	21.54	16.65	18.11	**15.83**
Using 8 mappers	Query set-1	F1-score	0.676	0.648	0.686	**0.735**
		Precision	0.557	0.491	0.55	**0.618**
		Recall	0.693	0.62	0.658	**0.705**
		CC	41.61	33.59	32.07	**31.23**
		CT	28.804	27.858	28.51	**27.51**
	Query set-2	F1-score	0.58	0.562	0.579	**0.634**
		Precision	0.584	0.588	0.552	**0.633**
		Recall	0.546	0.635	0.689	**0.703**
		CC	45.019	40.46	40.89	**39.519**
		CT	19.598	18.36	18.818	**17.844**

skeleton are 0.580, 0.562, 0.579, and 0.634, respectively. The analysis based on precision parameter using query set-1 is illustrated in Fig. 16b). When the number of retrieval is 2, the corresponding precision values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 0.782, 0.780, 0.740, and 0.822, respectively. Likewise, for 4 retrievals, the corresponding precision values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 0.584, 0.588, 0.552, and 0.633, respectively. The analysis based on recall parameter using query set-1 is illustrated in Fig. 16c). When the number of retrieval is 2, the corresponding recall values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 0.477, 0.498, 0.504, and 0.511, respectively. Likewise, for 4 retrievals, the corresponding recall values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 0.546, 0.635, 0.689, and 0.703, respectively. The analysis based on computational cost parameter using query set-1 is illustrated in Fig. 16d). When the number of retrieval is 2, the corresponding precision values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 45.019, 40.466, 40.899, and 39.519, respectively. Likewise, for 4 retrievals, the corresponding precision values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 55.060, 48.264, 49.594, and 42.798, respectively. The analysis based on computational time parameter using query set-1 is illustrated in Fig. 16e). When the number of retrieval is 2, the corresponding precision values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 19.598, 18.360, 18.818, and 17.844, respectively. Likewise, for 4 retrievals, the corresponding precision values computed by existing DCNN, Progressive image retrieval, MKSIFT + Cross indexing, and Proposed CS-based hierarchical skeleton are 26.106, 25.686, 25.692, and 25.000, respectively.

D. Comparative Discussion

Table I illustrates the analysis of methods using four and eight mappers along with two query sets. The analysis is done by adapting the maximal performance attained by each methods using precision, F1-Score and recall parameters. Considering four mappers, the proposed method attained better performance with maximal F1-Score of 0.784, maximal precision of 0.729 and maximal recall of 0.75, minimal computational cost of 32.77, and minimal computational time of 26.18 for query set-1. Using eight mappers, the proposed method attained maximal F1-Score of 0.735, and maximal recall of 0.705 for query set-1. The maximal precision is obtained in query set-2 with precision value of 0.633. Also, using 8 mappers, the proposed method has the minimal computational cost and computational time of 31.23 and 17, 84 sec, respectively for query set-1. From the analysis, it is noted that the best performance is achieved by proposed method, which shows its effectiveness in image retrieval.

The proposed CS-based Hierarchical Skeleton and Cross Indexing for Large Scale Image Retrieval Using Mapreduce Framework has the better results than the existing methodologies. This is happened because the proposed method has the benefit like low complexity for sparse vector. The cross indexing in image retrieval mechanism helps to improve the accuracy of image retrieval.

V. Conclusion

In this research, a new image retrieval method using CS-based hierarchical skeleton and cross indexing is proposed to perform the image retrieval mechanism based on the feature vector. The proposed method effectively achieves better retrieval performance through the cross indexing model. The features extracted using SURF and CS-based hierarchical skeleton are transformed into the binary sequences. The feature vector with the representation of binary code ensures the effectiveness of retrieval process based on the MSB code. The images stored in the mapper are retrieved based on the binary sequences using the tanimoto similarity measure. The retrieval process is carried

out through the reducer in the MapReduce framework based on the MSB code of the binary sequences. The proposed image retrieval is effectively operated with the binary sequence rather than considering the decimal value. The proposed attained better performance with the values of 0.784, 0.729, 0.75, 31.23, and 17.84sec for F1-score, precision, recall, computational cost and computational time with the query set-1 by considering 4 mappers. The proposed image retrieval is widely useful in large scale image processing field for image retrieval in effective and efficient manner on cloud environments. Moreover it is helpful in the medical field for the diagnosis aids. In future, the performance of image retrieval mechanism will be increased by some other model for saving the image in the mapper.

References

[1] C.Bai, L. Huang, X. Pan, J. Zheng, S. Chen, "Optimization of deep convolutional neural network for large scale image retrieval," *Neurocomputing*, vol. 303, pp. 60-67, 2018.https://doi.org/10.1016/j.neucom.2018.04.034

[2] N.A. Sakr, A.I. Eldesouky, H. Arafat, "An efficient fast-response content-based image retrieval framework for big data," *Computers & Electrical Engineering*, vol. 54, pp.522-538, 2016. https://doi.org/10.1016/j.compeleceng.2016.04.015

[3] X. Gao, X. Shi, G. Zhang,J. Lin, M. Liao, K.C. Li, C. Li, "Progressive Image Retrieval With Quality Guarantee Under MapReduce Framework," *IEEE Access*, vol. 6, pp.44685-44697, 2018.https://doi.org/10.1109/ACCESS.2018.2842796

[4] J. Cao, M. Wang, H. Shi, G. Hu, Y. Tian, "A new approach for large-scale scene image retrieval based on improved parallel-means algorithm in mapreduce environment," *Mathematical Problems in Engineering*, 2016. https://doi.org/10.1155/2016/3593975

[5] H. Wang, Z. Li, Y. Li, B.B. Gupta, C. Choi, "Visual saliency guided complex image retrieval," *Pattern Recognition Letters*, Volume 130, pp64-72, 2020. https://doi.org/10.1016/j.patrec.2018.08.010

[6] M. A. Alazeez, H. B. Abdalla, G. Li, J.Lin, "NoSQL Injection: Data Security on Web Vulnerability," *International Journal of Security and Its Applications*, vol. 10, no. 9, pp. 55-64, 2016. http://dx.doi.org/10.14257/ijsia.2016.10.9.07

[7] H. B. Abdalla, J. Lin, G. Li, and M. Gilani, "NoSQL: Confidential on Data Security and Data Management by using a Mobile Application," *International Journal of Information and Electronics Engineering*, vol. 6, no. 2, pp. 84-88, 2016.https://doi.org/10.18178/IJIEE.2016.6.2.600

[8] C. Stergiou, K. E. Psannis, B. B.Gupta, "IoT-based Big Data secure management in the Fog over a 6G Wireless Network,"*IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 5164 - 5171, 2021. https://doi.org/10.1109/JIOT.2020.3033131

[9] C. Yu, J. Li, X. Li, X. Ren, B. B. Gupta, "Four-image encryption scheme based on quaternion Fresnel transform, chaos and computer generated hologram,"*Multimedia Tools and Applications*, vol. 77, no. 19, pp. 4585–4608, February 2018.https://doi.org/10.1007/s11042-017-4637-6

[10] C. Yu, X. Li, S. Xu, and J. Li"Computer Generated Hologram-Based Image Cryptosystem with Multiple Chaotic Systems,"*Wireless Networks*, vol. 27, pp. 3507–3521, 2021.https://doi.org/10.1007/s11276-019-02223-z

[11] A.A.E. Latif, B.A.E.Atty, M.S.Hossain, A. Rahman, A. Alamri, and B. B. Gupta, "Efficient Quantum Information Hiding for Remote Medical Image Sharing,"*IEEE Access*, vol. 6, pp. 21075 – 21083,2018.https://doi.org/10.1109/ACCESS.2018.2820603

[12] W. Guo, N.K. Alham, Y. Liu, M. Li, M. Qi, "A resource aware MapReduce based parallel SVM for large scale image classifications," *Neural Processing Letters*, vol. 44, no. 1, pp.161-184, 2016.https://doi.org/10.1007/s11063-015-9472-z.

[13] V. Srivastava, S. Gupta, G. Chaudhary, A. Balodi, M. Khari, and V. G. Díaz, "An Enhanced Texture-Based Feature Extraction Approach for Classification of Biomedical Images of CT-Scan of Lungs," *International Journal of Interactive Multimedia and Artificial Intelligence* (Special Issue on Current Trends in Intelligent Multimedia Processing Systems), vol. 6, no. 7, pp. 18-25, 2021. http://doi.org/10.9781/ijimai.2020.11.003

[14] N. Lohar, D. Chavan, S. Arade, A. Jadhav, D. Chikmurge, "Content Based Image Retrieval System over Hadoop Using MapReduce,"*International Journal of Scientific Research in Science, Engineering and Technology*, Vol.2, no. 1, pp. 123-125, 2016.

[15] A.W.Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 22, no. 12, pp.1349-1380, 2000.https://doi.org/10.1109/34.895972

[16] M. Alsmirat, F. Alalem, M. A.Ayyoub, Y. Jararweh, B. Gupta, "Impact of digital fingerprint image quality on the fingerprint recognition accuracy,"*Multimedia Tools and Applications*, vol.78(4), pp. 3649–3688, February 2019.https://doi.org/10.1007/s11042-017-5537-5

[17] M. Lagiewka, M. Korytkowski, R.Scherer, R, "Distributed image retrieval with colour and keypoint features," *Journal of Information and Telecommunication*, vol. 3, no. 4, pp.430-445, 2019.https://doi.org/10.1080/24751839.2019.1620023

[18] S. Mezzoudj, R. Seghir, Y. Saadna,"A Parallel Content-Based Image Retrieval System Using Spark and Tachyon Frameworks," *Journal of King Saud University-Computer and Information Sciences*, vol. 33, no. 2, pp. 141-149 2021https://doi.org/10.1016/j.jksuci.2019.01.003

[19] C. Yan, H. Xie, S. Liu, J. Yin, Y. Zhang, Q. Dai, "Effective Uyghur language text detection in complex background images for traffic prompt identification," *IEEE transactions on intelligent transportation systems*, vol. 19, no. 1, pp.220-229, 2018.https://doi.org/10.1109/TITS.2017.2749977

[20] C. Yan, Y. Zhang, J. Xu, F. Dai, J. Zhang, Q. Dai, F. Wu, "Efficient parallel framework for HEVC motion estimation on many-core processors," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 12, pp.2077-2089, 2014.https://doi.org/10.1109/TCSVT.2014.2335852

[21] L. Deng, "A tutorial survey of architectures, algorithms, and applications for deep learning," *APSIPA Transactions on Signal and Information Processing*, vol. 3, 2014. https://doi.org/10.1017/atsip.2013.9

[22] G.E. Hinton, S. Osindero, Y.W.The, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp.1527-1554, 2006.https://doi.org/10.1162/neco.2006.18.7.1527

[23] M. Norouzi, M. Ranjbar, G. Mori, "Stacks of convolutional restricted boltzmann machines for shift-invariant feature learning," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2735-2742, June 2009. https://doi.org/10.1109/CVPRW.2009.5206577

[24] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu,F.E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp.11-26, 2017.https://doi.org/10.1016/j.neucom.2016.12.038

[25] D. Li, L. Deng, B.B. Gupta, H. Wang, C. Choi, "A Novel CNN based Security Guaranteed Image Watermarking Generation Scenario for Smart City Applications,"*Information Sciences*, vol. 479, pp. 432-447, 2019. https://doi.org/10.1016/j.ins.2018.02.060

[26] A. Sharif Razavian, H.Azizpour, J. Sullivan, S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," *In Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 806-813, 2014. https://doi.org/10.1109/CVPRW.2014.131

[27] R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580-587, 2014.https://doi.org/10.1109/CVPR.2014.81

[28] S. Xibing, W. Rong, Y. Yi, "MapReduce Based Remote Sensing Image Retrieval Algorithm," *International Journal of Database Theory and Application*, vol. 9, no. 8, pp.1-12, 2016.http://dx.doi.org/10.14257/ijdta.2016.9.8.01

[29] Z. Meng, "Remote Sensing Image Retrieval Algorithm based on MapReduce and Characteristic Information," *International Journal of Simulation, Systems, Science and Technology*, vol. 17, no. 3, 2016. https://doi.org/10.5013/IJSSST.a.17.03.07

[30] X. Li, D. He, J.Y. Li, "Parallel image search application based on online hashing hierarchical ranking," *Cluster Computing*, vol. 22, no. 1, pp.1469-1478, 2019.https://doi.org/10.1007/s10586-018-1922-8

[31] C. Schmid, R. Mohr, "Local grayvalue invariants for image retrieval," *IEEE transactions on pattern analysis and machine intelligence*, vol. 19, no. 5, pp.530-535, 1997.https://doi.org/10.1109/34.589215

[32] U. Hany, L. Akter, "Speeded-Up Robust Feature extraction and matching for fingerprint recognition," IEEE International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), pp. 1-7, May 2015.https://doi.org/10.1109/ICEEICT.2015.7307439

[33] C. Yang, O. Tiebe, K. Shirahama, M. Grzegorzek, "Object matching with

hierarchical skeletons," *Pattern Recognition*, vol. 55, pp.183-197, 2016. https://doi.org/10.1016/j.patcog.2016.01.022

[34] B. Mathan Kumar, R. PushpaLakshmi, "Multiple kernel scale invariant feature transform and cross indexing for image search and retrieval," *The Imaging Science Journal*, vol.66, no.2, pp.84-97, 2018.https://doi.org/10.1080/13682199.2017.1378285

[35] Nister and Stewenius, "UK bench dataset", "https://archive.org/details/ukbench," accessed on September 10, 2019.

Zhong Qianwen

Zhong Qianwen is doing Electronic and Communication Engineering in Shangdong University, China, Shandong, Weihai, Huancui District, Shandong University. His area of interest is image processing, big data classification, Artificial Intelligence and Network and security. He has published many papers in various reputed journals.

A Greedy Randomized Adaptive Search With Probabilistic Learning for solving the Uncapacitated Plant Cycle Location Problem

Israel López-Plata[1]*, Christopher Expósito-Izquierdo[1], Eduardo Lalla-Ruiz[2], Belén Melián-Batista[1], J. Marcos Moreno-Vega[1]

[1] Universidad de La Laguna, La Laguna (Spain)
[2] University of Twente, Enschede (Netherlands)

Received 17 December 2020 | Accepted 18 December 2021 | Published 25 April 2022

Abstract

In this paper, we address the Uncapacitated Plant Cycle Location Problem. It is a location-routing problem aimed at determining a subset of locations to set up plants dedicated to serving customers. We propose a mathematical formulation to model the problem. The high computational burden required by the formulation when tackling large scenarios encourages us to develop a Greedy Randomized Adaptive Search Procedure with Probabilistic Learning Model. Its rationale is to divide the problem into two interconnected sub-problems. The computational results indicate the high performance of our proposal in terms of the quality of reported solutions and computational time. Specifically, we have overcome the best approach from the literature on a wide range of scenarios.

Keywords

Greedy Randomized Adaptive Search Procedure, Probabilistic Learning Model, Uncapacitated Plant Cycle Location Problem.

DOI: 10.9781/ijimai.2022.04.003

I. Introduction

Location-routing problems are a family of hard combinatorial optimization problems found in the field of distribution network design. The objective is to open a subset of depots in potential locations with the aim of fulfilling the demand of customers by means of a fleet of vehicles. Traditionally, these optimization problems have been tackled separately. However, the evolution of computers, the emergence of new optimization techniques, and the necessity of holistic solutions for new problem applications have aroused a renewed interest in their joint solution.

Nowadays, in city and last-mile logistics, freight transportation stakeholders and service providers have to regularly redesign and improve their logistics processes to satisfy customers requirements while reducing infrastructure and transportation costs. The design of transportation networks in the context of less-thana- truck deliveries leads to the definition of the Uncapacitated Plant Cycle Location Problem (UPCLP).

The UPCLP is an NP-Hard optimization problem whose main goal is to select a subset of locations from a bigger set of potential locations where establish plants to serve a determined set of customers. The number of plants is unknown in advance, but it is important to remark that due to the operations required to create a plant, both the set up and the assignment of a customer to it have a specific cost. Each plant has one vehicle to serve all its customers following a determined route that also have an associated cost. A solution for the UPCLP solves two different subproblems:

- Obtaining the set of locations where open the plants to serve all customers, minimizing the cost to open the plants and assigning every customer to a determined plant.
- Determining the routes followed by the vehicles to serve its assigned customers with the less possible cost.

In the strong sense, this problem combines two well-known optimization problems: the Uncapacitated Facility Location Problem [1] and the Multi-Depot Travelling Salesman Problem [2]. Section III explains the UPCLP in detail.

The applications for the UPCLP are those related to the location of plants where the service or freights distributed to customers are not affected by plant or vehicle capacity constraints. Related applications can be found in humanitarian logistics [3], telecommunications [4], [5], distribution system design [6], postal delivery [7], [8], among others.

The main goals of the present paper are described as follows:

- Proposing an optimization model for the UPCLP.
- Developing a metaheuristic approach based on the paradigm of the Greedy Randomized Adaptive Search Procedure (GRASP) that incorporates a probability distribution for selecting locations within the UPCLP. Its goal is to obtain faster solutions than the optimization model, and on the other hand, provide feasible solutions for larger scenarios that may appear in practical cases and where the optimization model is unable to provide a solution.

* Corresponding author.
E-mail address: ilopezpl@ull.edu.es

- Assessing the performance of the GRASP in comparison with the best approach so far and our optimization model on problem instances from the literature. Additionally, with the aim of evaluating our metaheuristic on structured scenarios, a set of instances adapted from TSPLIB [9] to simulate largescale scenarios are also tackled.

The remainder of this work is organized as follows. Section II reviews the literature related to the UPCLP. Section III explains the detail of the UPCLP. Afterwards, Section IV presents an optimization model of the UPCLP. Section V describes a GRASP with Probabilistic Learning Model to solve the UPCLP from an approximate standpoint. Section VI discusses the applicability of the optimization model and a metaheuristic approach to realistic scenarios and checks their performances in comparison with a previous approach. Finally, Section VII extracts the main conclusions from the work and suggests several lines for further research.

II. Literature Review

The interest in location-routing problems by the scientific community has increased over the last years. One of the seminal papers in the field can be attributed to Watson-Gandy and Dohrn in 1973, as indicated in [10]. The main reason for the joint treatment of decisions concerning the location of plants and vehicle routing is found in that solving them independently gives rise to only suboptimal results in most cases. The suboptimality of the results has been demonstrated in a multitude of works published in the literature. This is the case of the two-phase tabu search proposed in [11].

Evidence of the increasing interest is the fact that numerous surveys dedicated to the analysis and classification of works related to location-routing problems have been published so far. Some outstanding examples are the papers [12], [10], [13], [14], and [15]. New variants and applications of location-routing problems are also discussed in [16].

The interest in this kind of problem has arisen from several practical fields. This is the case of waste management. The work [17] aims to identify the best place to open treatment centers and how to route the hazardous waste derived from industrial activity to disposal centers in a region of Turkey. In this case, the optimization goal is to minimize the transportation risk and the total cost associated with the fixed annual cost of opening a treatment technology and a disposal facility. The authors propose a mixed-integer programming model that incorporates constraints to handle mass balance or the minimum amount of waste required for technology. Furthermore, [18] introduces an improved metaheuristic with a specially-designed directed local search procedure to solve a general two-echelon multi-objective location routing problem in waste collection. In particular, two-echelon transportations must first collect waste generated in demand nodes to collection centers. Then, the waste must be transported while considering flow constraints and capacity constraints. [19] uses a K-Means clustering combined with an Ant Colony Optimization (ACO) to find the shortest routes between 2 nodes in a network of IoT devices optimizing the Quality of Service of the network. This paper divides the whole network into clusters depending of the types of subnetworks, which optimizes the routes creation. Lastly, [20] focuses on the capacitated location routing problem, where each depot has a fixed opening cost and a known capacity to satisfy the customers' demand. Also, the vehicles have capacities and travelling costs. Thus, the optimization criterion is to minimize the total cost, composed of depot, travelling, and vehicle costs. The problem is solved by means of a a hybrid genetic algorithm that explores unfeasible solutions and presents a high competitive performance in comparison with other approaches found in the literature in terms of solution quality and time efficiency.

Other practical applications of location-routing problems are telecommunication network design, electric vehicle transportation, good distribution or airline topologies, among others. For example, [21] addresses the problem of designing synchronous digital hierarchy rings in the context of mobile communications access networks. The problem consists in finding the number and type of the base station controllers to locate at each potential site and, on the other hand, in defining synchronous digital hierarchy rings such that each base transceiver station is in exactly one ring. The authors propose a mixed integer programming model and a heuristic method to solve the problem in realworld instances. [22] seeks to find the number and location of electric vehicle battery swap stations with an optimal route plan based on stochastic customer demands. The problem is solved by means of a hybrid variable neighbourhood search algorithm that combines a binary particle swarm optimization. [23] presents a transportation location routing problem in which the goal is to satisfy the demand of clients from a set of plants with maximum capacity and through intermediate eligible points called city distribution centers, which are sites dedicated to receive products from the plants and deliver them to the clients. The objectives are minimizing the total operation cost of the system and maintain balance in the vehicle operator's workload. Lastly, [24] includes a set of hubs to improve the routes followed by planes on iranian airspace, using a Multi-objective Genetic Algorithm to set the best places to locate these hubs.

Due to their performance, metaheuristics have become attractive alternatives to address location routing problems. Representative examples of these techniques have been proposed so far. Some of them are variable neighborhood search [25], multiple ant colony optimization algorithm [26], Simulated Annealing [27], Particle Swarm Optimization [28], hybrid PSO with Path Relinking [29], Tabu Search [30], GRASP with Path Relinking [31], and clustering analysis [32].

In spite of the existence of a wide corpus of papers in the literature about location-routing problems, the works briefly described in the following are of special interest in this paper. [33] introduced the Uncapacitated Plant Cycle Location Problem and proposed a preliminary version of the technique presented in the paper at hand. [34] presents a strategic problem that can be seen as a generalization of the location-routing problem in which the Capacitated Facility Location Problem and the Multi-Depot Vehicle Routing Problem are combined. The mentioned problem considers costs derived from vehicle usage, vehicle and location capacities, and customer demands. Similar multiroute capacitated approaches have been recently considered in [35], [36], and [37]. Moreover, [38] presents a multiobjective application of location-routing problems to home-to-work bus service. On the other hand, [39] proposes a MIP model and a Branch-and-Cut algorithm to solve several two-level network design problems. Furthermore, [40] and [41] address the PCLP with maximum service capacity constraints associated with the plants to set up. In the first paper, the authors propose a Branch and Cut, even guarantees the optimality of the reported solutions, it requires extremely large computational times (more than 1 hour) in a multitude of cases. In the latter paper, a tabu search is proposed in which an initial solution is obtained from an optimization model. Lastly, [42] proposes a metaheuristic approach based upon the Honey Bees Mating Optimization algorithm for solving the UPCLP. The computational results indicate the algorithm provides highquality solutions in reasonable computational times.

III. Uncapacitated Plant Cycle Location Problem

The Uncapacitated Plant Cycle Location Problem (UPCLP) is a deterministic optimization problem that seeks to select a subset of locations to set up plants with the aim of serving customers geographically distributed on a two-dimensional scenario.

Input data of the UPCLP is a set of m potential discrete locations (*e.g.*, places with the required technical equipment, safety places, etc.), denoted as M, in which to place plants (*e.g.*, industrial infrastructures, hubs, health-care services, warehouses, cross-docking centers, etc.) to serve a well-known set of n customers, denoted as N. Each available location, $j \in M$, could have at most one plant. This way, the set of plants is denoted as $P \subseteq M$. The number of plants set up at the available locations is $k \leq m$, but unknown in advance. That is, $|P| = k$. In this regard, setting up a plant at location $j \in M$ incurs a fixed cost, denoted as $o_j \geq 0$, which indicates, according to the application field, the opening cost, time required to establish a medical camp, etc.

In the UPCLP, each plant can serve an unlimited number of customers whereas each customer must be served directly by exactly one of the plants (*i.e.*, single-echelon approach). However, assigning a customer $i \in N$ to a plant at location $j \in M$ gives rise to a fixed cost, denoted as $c_{ij} \geq 0$, which indicates the cost of providing service to the customer. Without loss of generality, it is assumed that a given customer can be assigned to a plant at any location. The plant where a customer $i \in N$ is assigned to is denoted as $\sigma(i) \in P$. The set of customers served from a given location $j \in M$ is denoted as N_j, where $N = \cup_{j \in M} N_j$ and $N_j \cap N_{j'} = \emptyset, \forall j, j' \in M$. It should be noted that $N_j = H$ whenever no plant is set up at location $j \in M$. Furthermore, the set of customers assigned to a given plant must be served following a delivery route. In this regard, the travel cost between two customers or locations, $i, j \in N \cup M$, is symmetric and denoted as $d_{ij} > 0$, where $d_{ij} = d_{ji}$. All the travel costs satisfy the triangle inequality [43].

The previous description of the UPCLP indicates the following decisions have to be made: (i) selecting a subset of locations in which to set up plants, (ii) determining which each plant serves a subset of customers, and (iii) building vehicle routes to serve the customers (*i.e.*, the sequence in which those customers associated with each plant are going to be served).

Fig. 1. illustrates an example of the UPCLP composed of $m = 5$ locations and $n = 25$ customers. In this case, $k = 2$ plants have been set up. One of the plants serves customers 1, 10, 11, 19, 7, 8, 18, 5, 17, 14, 15, 2, 13, and 6, whereas the other plant serves the remaining customers.

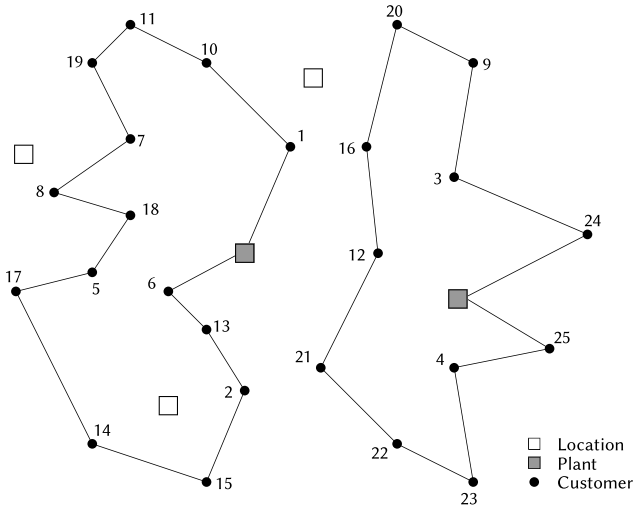


Fig. 1. Example of the Uncapacitated Plant Cycle Location Problem composed of $m = 5$ locations and $n = 25$ customers.

Finally, it worth mentioning that, whenever the UPCLP consists of only one location in which to place plants to serve the customers (*i.e.*, $m = k = 1$) and all the assignment costs are equal (*i.e.*, $c_{ij} = r; \forall i \in N; j \in M$, where r is a non-negative value), the UPCLP reduces to the Travelling Salesman Problem [44]. Consequently, in the strong sense, the UPCLP belongs to the NP-hard class of optimization problem.

IV. Optimization Model

In this section, we present a Mixed-Integer Linear Programming (MILP) model aimed at solving the Uncapacitated Plant Cycle Location Problem (UPCLP). With this goal in mind, in the following we firstly introduce the families of variables used in the model:

u_i	Integer variable associated with customer $i \in N$
x_{ij}	1 if the edge (i, j) is included in the solution, where $i, j \in M \cup N$. 0, otherwise
y_j	1 if a plant is set up at location $j \in M$. 0, otherwise
z_{ij}	1 if customer $i \in N$ is assigned to a plant set up at location $j \in M$. 0, otherwise

The objective function of the MILP is to minimize the costs derived from (i) setting up plants, (ii) assigning customers to the plants, and (iii) routing the customers, as shown in (1):

$$\min \sum_{j \in M} o_j \cdot y_j + \sum_{i \in N} \sum_{j \in M} c_{ij} \cdot z_{ij} + \sum_{i \in M \cup N} \sum_{j \in M \cup N} d_{ij} \cdot x_{ij} \tag{1}$$

Each customer is assigned to exactly one plant, as seen in (2):

$$\sum_{j \in M} z_{ij} = 1, \forall i \in N \tag{2}$$

Each plant can serve an unlimited number of customers, as shown in (3):

$$\sum_{i \in N} z_{ij} \leq n_{limit} \cdot y_j, \forall j \in M \tag{3}$$

where parameter n_{limit} has to be equal to or larger than n for modeling the uncapacitated version of this problem with regards to facilities. For switching to the capacitated version, then n_{limit} has to be less than n (leading to the PCLP). Finally, note that in our case, this constraint is redundant and can be omitted.

Degree constraints aimed at ensuring that each customer has previous (4) and next (5) nodes in its route:

$$\sum_{\substack{i' \in M \cup N \\ i' \neq i}} x_{ii'} = 1, \forall i \in N \tag{4}$$

$$\sum_{\substack{i' \in M \cup N \\ i' \neq i}} x_{i'i} = 1, \forall i \in N \tag{5}$$

Degree constraints aimed at ensuring that each location has previous (6) and next (7) nodes only if a plant has been set up:

$$\sum_{i \in N} x_{ij} = y_j, \forall j \in M \tag{6}$$

$$\sum_{i \in N} x_{ji} = y_j, \forall j \in M \tag{7}$$

Subtour elimination constraints in which all the customers can be served along the same route can be seen on equations (8) and (9):

$$u_i - u_{i'} + n \cdot x_{i'i} \leq n - 1, \forall i, i' \in N, i \neq i' \tag{8}$$

$$1 \leq u_i \leq n, \forall i \in N \tag{9}$$

The edge $(i; j)$ can be used if and only if customer $i \in N$ is assigned to a plant set up at location $j \in M$, as seen in (10):

$$x_{ij} \leq z_{ij}, \forall i \in N, \forall j \in M \tag{10}$$

Constraint (11) restricts if customers $i, i' \in N$ are assigned to plants

set up at different locations, $j, j' \in M$, then they cannot be in the same route:

$$x_{i'i} + z_{ij} + z_{i'j'} \leq 2, \forall i, i' \in N, i \neq i', \forall j, j' \in M, j \neq j' \quad (11)$$

Finally, the domain of the decision variables is defined on equations (12), (13) and (14):

$$x_{ij} \in \{0,1\} \quad \forall i, j \in M \cup N \quad (12)$$

$$y_j \in \{0,1\} \quad \forall j \in M \quad (13)$$

$$z_{ij} \in \{0,1\} \quad \forall i \in N, j \in M \quad (14)$$

V. Greedy Randomized Adaptive Search Procedure With Probabilistic Learning Model

A Greedy Randomized Adaptive Search Procedure [45] with Probabilistic Learning Model (GRASP-PLM) is here presented to solve the Uncapacitated Plant Cycle Location Problem (UPCLP). In general terms, a GRASP is an iterative metaheuristic based upon two main components: a constructive phase aimed at building feasible solutions and an intensification phase dedicated to improving the quality of the found solutions. The high performance of the GRASP when tackling a wide range of heterogeneous combinatorial problems from the literature encourages us to consider it as a promising candidate to solve the UPCLP. The GRASP-PLM includes a joint probability distribution that allows selecting a subset of locations to set up plants.

The rationale behind our GRASP-PLM is to split the UPCLP into the following two interconnected sub-problems to be solved consecutively:

1. *High-level Problem* (*HP*). Determining the subset of locations in which to set up plants to serve the customers (*i.e.*, $P \subseteq M$).
2. *Low-level Problem* (*LP*). Given the plants, assigning the customers to the plants and determining the delivery routes to serve them.

The pseudocode of our GRASP-PLM is depicted in Algorithm 1. The first step is to obtain a solution of the HP, s_{HP} , by sampling the PLM (line 3). This process is described in Section A. Once the plants have been set up, the assignments and routes of the customers are determined by means of a constructive phase (line 4). This phase gives rise to a feasible solution of the UPCLP, denoted as s. A local optimum, s_{local}, is achieve from s. The PLM is updated in those cases in which the best solution found by the search is improved (lines 6-9). Lastly, the search is finished when a certain stop criterion is met (lines 2-10).

Algorithm 1. Pseudocode of the GRASP-PLM for the Uncapacitated Plant Cycle Location Problem

1: $s_{best} \rightarrow \emptyset$
2: **while** (stop criterion is not met) **do**
3: $s_{HP} \leftarrow$ Get high-level solution from the probabilistic learning model
4: $s \leftarrow$ Assign customers and determine routes associated with the plants in s_{HP}
5: $s_{local} \leftarrow$ Apply local search to s
6: **if** ($f(s_{local})$ $f(s_{best})$) **then**
7: $s_{best} \leftarrow s_{local}$
8: Update probabilistic learning model with s_{best}
9: **end if**
10: **end while**
11: Return s_{best}

A. High-Level Problem

The High-level Problem (HP) seeks to determine a subset of locations to set up plants aimed at serving the customers. The main decisions to make at this point are to (i) determine the number of those plants to set up and (ii) select a non-empty set of locations to set up plants. With these goals in mind, we propose to use a Probabilistic Learning Model (PLM). It is composed of the following vectors of probabilities:

1. v_1. It is a vector of m elements, where $v_1(i)$ is the probability of opening i plants. This probability is formally defined as follows:

$$v_1(i) = \frac{f_1(i)}{\sum_{j=1}^{n} f_1(j)}, \forall i = 1,2, \ldots, m \quad (15)$$

where $f_1(i)$ is the number of times i plants have been open in a previous high-quality solution found during the search. Initially, $f_1(i) = 1, \forall i = 1, 2, \ldots, m$.

2. v_2. It is a vector of m elements, where $v_2(i)$ is the probability of setting up a plant at location i. This probability is formally defined as follows:

$$v_2(i) = \frac{f_2(i)}{\sum_{j=1}^{n} f_2(j)}, \forall i = 1,2, \ldots, m \quad (16)$$

where $f_2(i)$ is the number of times a plant has been set up at location $i \in M$ in a previous high-quality solution found during the search. Initially, $f_2(i) = 1, \forall i = 1, 2, \ldots, m$.

A two-step process is carried out to sample solutions from the PLM. Firstly, a random probability is generated, denoted as $p_1 \in [0 \ldots 1]$. This probability allows to determine the number of plants to set up, $1 \leq k \leq m$, as follows:

$$k = \arg max_{i=1,2,\ldots,m} \left\{ \sum_{j=1}^{i} v_1(j) \mid \sum_{j=1}^{i} v_1(j) \leq p_1 \right\} \quad (17)$$

Finally, once the number of plants is known, a set P composed of k locations must be defined according to the probabilities in v_2. The pseudocode of this process is depicted in Algorithm 2. At each step, a probability $p_2 \in [0 \ldots 1]$ is generated. The location with the maximum cumulative probability no greater than p_2 is selected. The process finishes when k different locations have been selected. It should be noted that P constitutes a solution of the HP.

Algorithm 2. Pseudocode of selection of locations in which to set up plants

1: $P \leftarrow \emptyset$
2: **while** ($|P| < k$) **do**
3: $p_2 \leftarrow$ Generate random probability
4: $l \leftarrow \arg max_{i=1,2,\ldots,m} \{\sum_{j=1}^{i} v_2(j) \mid \sum_{j=1}^{i} v_2(j) \leq p_2\}$
5: $P \leftarrow P \cup (l)$
6: end while
7: Return P

As indicated in Algorithm 1, the PLM is updated every time a new best solution, s_{best}, is found during the search. This means that, if $|P| = k$ in s_{best}, the following operations are carried out:

$$f_1(k) = f_1(k) + 1 \quad (18)$$

$$f_2(l) = f_2(l) + 1, \forall l \in P \quad (19)$$

This way, the influence of selecting k plants and the relevant locations are increased for the following sampling process according to equations (15) and (16), respectively.

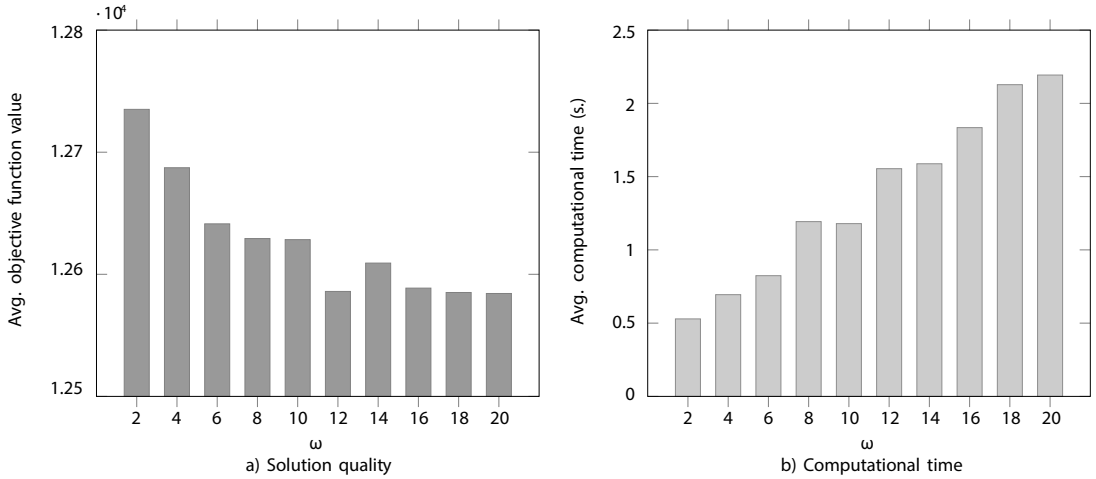


Fig. 2. Average objective function value and average computational time of the GRASP-PLM on a subset of problem instances.

B. Low-Level Problem

The solutions of the HP are not feasible solutions of the UPCLP. In order to overcome this fact, we must still determine the assignments of customers to plants and the routes to serve them. For this purpose, we use a GRASP.

The constructive phase of our GRASP builds a feasible solution of the UPCLP iteratively. At each step, one non-assigned customer is considered to be part of the solution. This customer can be assigned to one of the k plants defined by the high-level solution previously. Additionally, given a customer and a plant to serve it, this customer can be routed between each pair of consecutive nodes in the route of that plant.

A greedy function, $g : (a, b, c, p) \longrightarrow \mathbb{R}$, evaluates the impact on the objective function value of including the customer $b \in N$ between the consecutive customers or locations $a, c \in M \cup N$ in the route of the plant $p \in P$. That is:

$$g(a, b, c, p) = d_{ab} + d_{bc} - d_{ac} + c_{bp} \tag{20}$$

This impact on the objective function value originates in adding edge (a, b) and (b, c), assigning customer $b \in N$ to plant $p \in P$, and removing edge (a, c).

All the possible positions in which non-assigned customers can be placed in the routes of the selected plants are evaluated according to $g(\cdot,\cdot,\cdot,\cdot)$. These possibilities are ordered incrementally based on their impacts on the objective function value. The $\omega > 0$ best possibilities constitute a Restricted Candidate List (RCL). The value of parameter ω is set by the user. At each step, one candidate is extracted from the RCL according to the roulette wheel selection and the involved non-assigned customer is included in the solution under construction. The process finishes when all the customers have been assigned and routed. Consequently, a feasible solution of the UPCLP is, at this point, obtained.

The optimality of the solutions reported by the constructive phase is not guaranteed. The reason is found in using a greedy but myopic function to evaluate the impact of including non-assigned customers into the solutions under construction.

We propose an intensification phase based upon local search to explore the current region of the search space. With this goal in mind, we consider a single one-point movement to explore the neighbourhood of each solution obtained after applying the constructive phase. Given a feasible solution of the UPCLP, the one-point movement relocates a customer into a new position, in the same, different, or new route. We evaluate the impact on the objective function value of applying the one-point movement to relocate each customer into each possible target position. Particularly, removing a customer $i \in N$ from its current position in a route is computed as follows:

$$\mu(i) = d_{p(i)s(i)} - \left(d_{p(i)i} + d_{is(i)} - c_{i\sigma(i)} + o_{\sigma(i)} \cdot \phi(i, 1)\right) \tag{21}$$

where $p(i)$ and $s(i)$ denote the previous and next nodes of customer i in its route, whereas $\phi(i, \alpha)$ is a binary variable that takes value 1 if and only if the route of i contains exactly α customers (*i.e.*, $|\mathrm{N}_{\sigma(i)}| = \alpha$). It should be noted that, a plant can be removed from the solution when its route contains only one customer and this is relocated. Similarly, relocating a customer $i \in N$ before another node $j \in M \cup N$ is computed as follows:

$$\gamma(i, j) = \begin{cases} 2 \cdot d_{ij} + c_{ij} + o_j, & \text{if } j \in M \text{ and } \emptyset\,(j, 0) = 1 \\ d_{p(j)i} + d_{ij} + c_{ij} - d_{p(j)j}, & \text{if } j \in M \text{ and } \emptyset\,(j, 0) = 0 \\ d_{p(j)i} + d_{ij} + c_{i\sigma(j)} - d_{p(j)j}, & \text{otherwise} \end{cases} \tag{22}$$

The first case corresponds to those scenarios in which customer i is included in a new route starting from a plant set up at location $j \in M$. The remaining cases refer to those environments in which i is placed before a plant or another customer, respectively.

According to equations (21) and (22), the impact on the objective function value of relocating a customer $i \in N$ from its current position to the previous position of node $j \in M \cup N$ is computed as follows:

$$h(i, j) = \mu(i) + \gamma(i, j) \tag{23}$$

Finally, it is worth mentioning that the customers are randomly selected to be relocated. The travel costs between a customer $i \in N$ to relocate and the remaining nodes are sorted in increasing order so that we first evaluate relocating i before those nodes at minimum travel cost. In addition, at each step, the best improving neighbour solution is chosen.

VI. Computational Experiments

This section is dedicated to assessing the optimization model's performances introduced in Section IV and the Greedy Randomized Adaptive Search Procedure with Probabilistic Learning Model (GRASP-PLM) presented in Section V. In this regard, all the computational experiments presented hereunder have been conducted over the benchmark suite proposed in [42] and in-stances adapted from the TSPLIB [9]. All the problem instances are published to be freely used by the research community[1]. The mathematical model has been executed with CPLEX 12.3, set to all-default. Our proposed

[1] https://sites.google.com/site/gciports/plantcycle

optimization technique has been im-plemented in Java Standard Edition 7. In all cases, we have used a computer equipped with an Intel i7-3.50 GHz and 16 GB of RAM and performed 10 executions of each problem instance.

A. Parameter Setting

We have carried out a parameter setting before applying the GRASP-PLM. The parameters whose values must be determined are the size of the Restricted Candidate List (RCL), denoted as ω, and the number of iterations to perform.

Fig. 2. shows the average objective function value of 10 executions over a subset of problem instances with different sizes and the average computational time required by our GRASP-PLM when varying ω from 2 up to 20.

As can be checked, there is a strong tendency to improve the quality of the solutions reported by the GRASP-PLM when increasing the number of elements included in the RCL. However, increasing the value of ω gives rise to require larger computational times. The reason is that considering a large number of elements allows the search to have a relevant diversity, but it is harder to build the RCL at each step. This is because every solution must be evaluated before including it in the RCL, with its corresponding computational cost. In order to obtain a good balance between quality and computational time, in the remainder of this paper, we have executed our GRASP-PLM with $\omega = 12$.

With the aim to determine if there are significant differences between the groups of solutions obtained with different values of ω, the Friedman test [46] is applied to the average objective function value of these solutions. The significance level of this test is 0.05, which indicates that there are statistically significant differences among the solutions under analysis. Fig. 3. shows the interquartile range returned by the Friedman test for every possible values of ω. Depending on its interquartile range, every group of solutions is classified on different groups, identified by letters. Groups of solutions classified with the same letter do not have significant differences between them. As can be observed on this graphic, results with $\omega \in [12, 20]$ belong to group f, which implies that their solutions do not have significant differences. This consolidates the decision of using $\omega = 12$ in the subsequent computational experiments.

Moreover, when assessing the number of iterations, we have evaluated the average objective function value for the different groups of instances for 1000 iterations and $\omega = 12$. Fig. 4., Fig. 5., and Fig. 6. show the results when $n = 10, 25, 100$, respectively. As can be checked, when the iterations are increased, the average quality in terms of objective function value increases. However, it should be noted that the performance improvement is accompanied by a linear increase of the computational time. In the following experiments, we have selected 100 as a number of iterations to perform for each instance on the basis of maintaining a suitable and competitive performance in terms of computational time with the other approach reported in the literature [42].

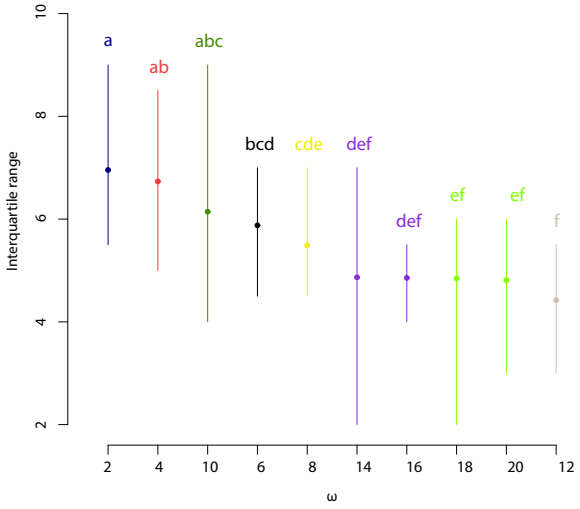


Fig. 3. Results of Friedman's test with different values of RCL size (i.e., ω).

B. Performance Evaluation

Tables I and II show the results obtained by the optimization model introduced in Section IV, the Honey Bees Mating Optimization algorithm (HBMO) proposed in [42], and our GRASP-PLM on a wide range of small-, medium-, and large-size instances proposed in [42]. In this case, column *Instances* reports the characteristics of the problem instances under analysis. For each problem instance, the number of customers, n, the number of locations, m, and the cost to set up a plant, o_p, are shown. In Table I, the computational results for the small- and medium-size problem instances are reported. These instances have a number of locations ranging from $m = 5$ up to $m = 25$, the number of customers ranges from $n = 10$ up to $n = 25$, whereas the cost to set up plants ranges from $o_p = 1$ up to $o_p = 1000$. On the other hand, Table II shows the results for the large-size problem instances. These instances have a number of locations ranging from $m = 50$ up to $m = 100$, the number of customers ranges from $n = 50$ up to $n = 100$, whereas the cost to set up plants ranges from $o_p = 1$ up to $o_p = 1000$. Each entry of the tables corresponds to a group of 5 problem instances. Hence, the average values are reported for each case.

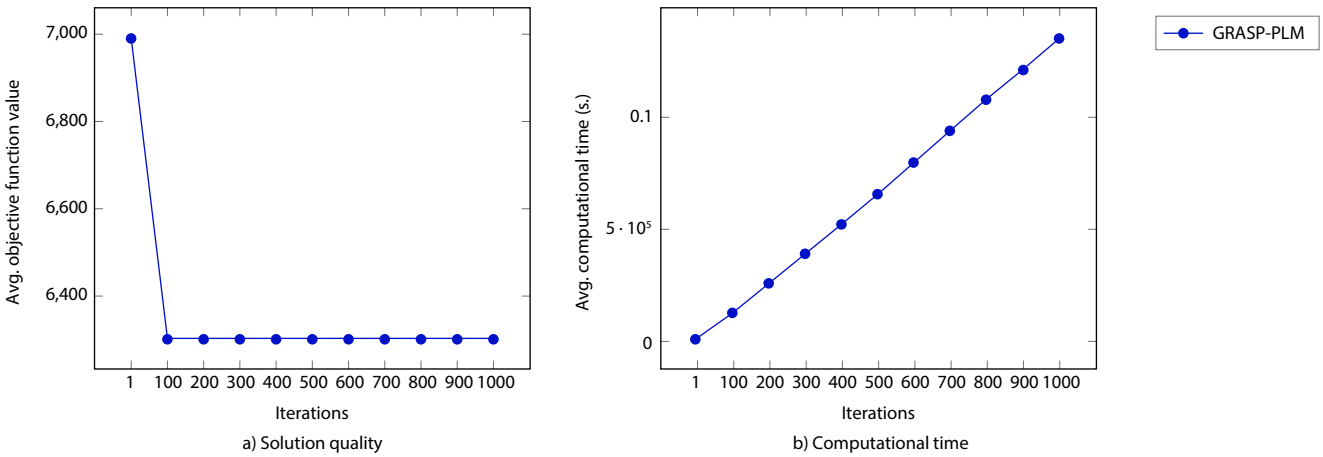


Fig. 4. Average objective function value and average computational time of the GRASP-PLM over the subset of problem instances with $n = 10$.

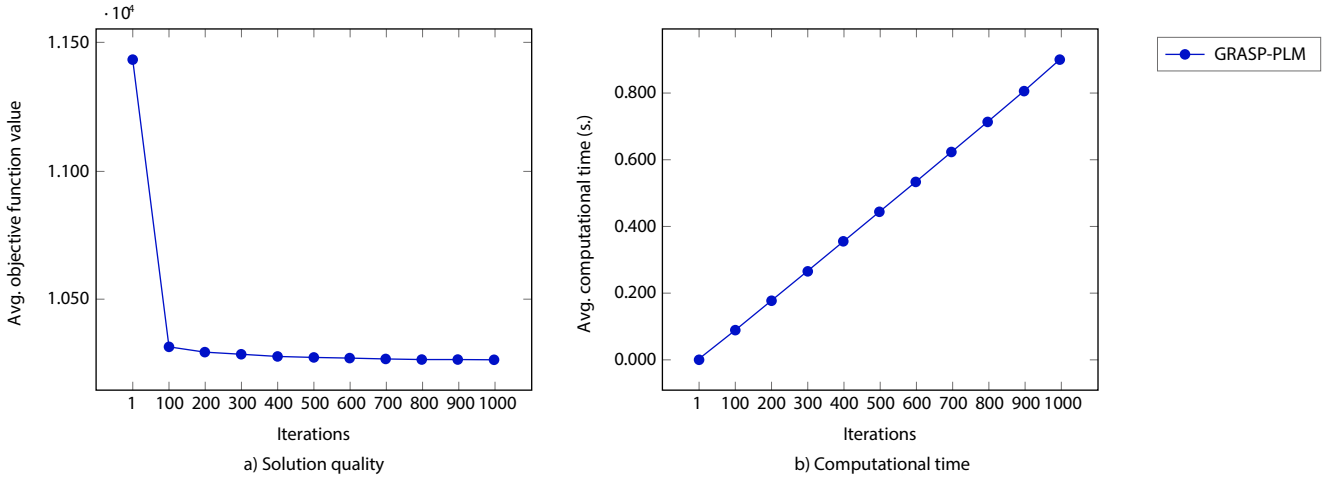


Fig. 5. Average objective function value and average computational time of the GRASP-PLM over the subset of problem instances with $n = 25$.

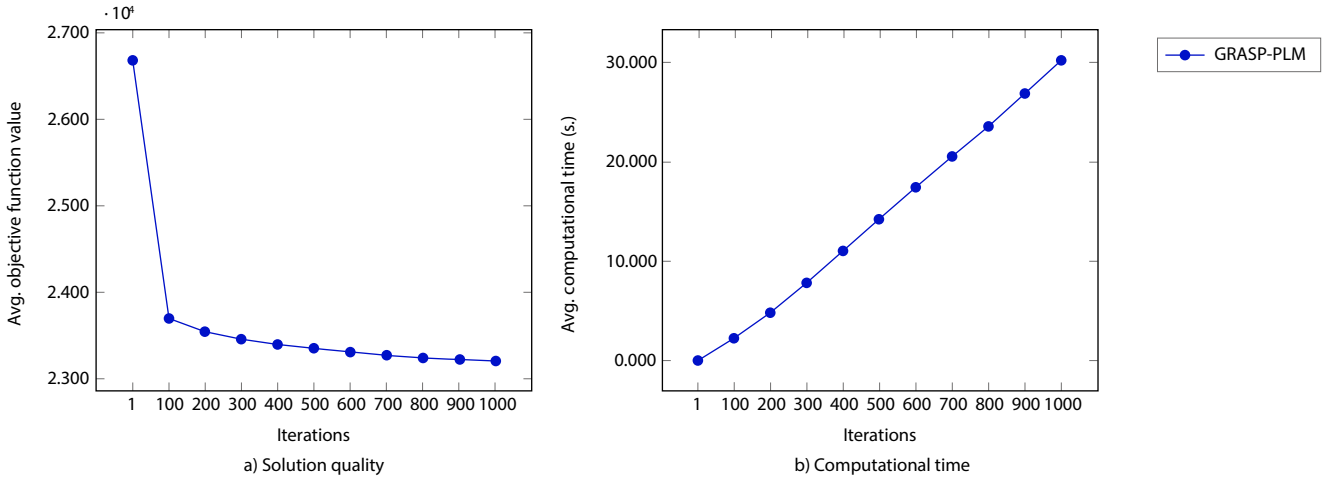


Fig. 6. Average objective function value and average computational time of the GRASP-PLM over the subset of problem instances with $n = 100$.

TABLE I. Computational Results for the Small- and Medium-Size Problem Instances

Instances			CPLEX		HBMO						GRASP-PLM					
					Deviation (%)			t (s.)			Deviation (%)			t (s.)		
n	m	O_p	Opt.	t (s.)	Min.	Avg.	Max.	Min.	Avg.	Max.	Min.	Avg.	Max.	Min.	Avg.	Max.
10	5	1	6003.0	0.21	0.32	0.46	1.00	0.33	0.35	0.38	**0.00**	**0.04**	**0.04**	< 0.01	0.01	0.01
		250	6105.6	0.13	0.00	0.88	1.38	0.35	0.39	0.42	**0.00**	**0.00**	**0.00**	< 0.01	0.01	0.01
		500	6362.2	0.15	0.00	0.36	0.90	0.50	0.52	0.53	**0.00**	**0.00**	**0.00**	< 0.01	0.01	0.01
		964	6735.8	0.19	0.00	0.52	1.90	0.44	0.48	0.61	**0.00**	**0.00**	**0.00**	< 0.01	0.01	0.01
		1000	7956.6	0.20	0.35	0.35	0.35	0.52	0.56	0.60	**0.00**	**0.02**	**0.20**	< 0.01	0.01	0.01
10	10	1	4291.2	0.23	0.69	0.77	1.07	0.27	0.27	0.28	**0.00**	**0.00**	**0.00**	0.01	0.01	0.02
		13	5106.2	0.40	0.99	2.52	3.74	0.27	0.30	0.33	**0.00**	**0.00**	**0.00**	0.01	0.01	0.02
		250	5785.2	1.26	0.00	1.48	4.54	0.31	0.35	0.38	**0.00**	**0.00**	**0.00**	0.01	0.01	0.02
		500	6655.4	0.79	0.00	1.56	4.36	0.30	0.45	0.71	**0.00**	**0.00**	**0.00**	0.01	0.01	0.02
		1000	8042.2	0.74	0.01	1.34	3.02	0.34	0.44	0.59	**0.00**	**0.01**	**0.08**	0.01	0.01	0.02
25	10	1	8616.0	41.84	0.23	0.58	1.31	0.79	0.85	0.93	**0.00**	**0.00**	**0.03**	0.05	0.06	0.06
		250	10206.2	89.28	0.57	1.35	1.89	0.58	0.67	0.81	**0.00**	**0.04**	**0.16**	0.05	0.06	0.06
		500	12048.2	101.83	1.14	2.85	4.27	0.62	0.81	1.14	**0.00**	**0.39**	**0.91**	0.05	0.06	0.06
		508	11823.8	422.96	0.83	2.47	5.21	1.16	1.78	2.46	**0.00**	**0.11**	**0.33**	0.05	0.06	0.07
		1000	13524.8	114.33	1.65	4.83	7.68	1.21	1.82	2.88	**0.26**	**1.54**	**3.56**	0.05	0.05	0.06
25	25	1	6042.4	20.95	0.33	0.66	0.87	0.49	0.51	0.54	**0.00**	**0.00**	**0.00**	0.10	0.10	0.11
		70	7501.8	84.99	1.21	1.71	2.42	0.45	0.48	0.52	**0.00**	**0.00**	**0.00**	0.09	0.10	0.10
		250	9610.6	75.62	4.94	6.41	7.38	0.45	0.49	0.53	**0.00**	**0.14**	**0.41**	0.09	0.10	0.11
		500	9989.4	98.23	6.06	10.42	14.23	0.48	0.59	0.81	**0.02**	**0.34**	**1.02**	0.09	0.10	0.10
		1000	13096.2	1970.58	2.34	8.73	17.94	0.72	1.38	2.44	**1.71**	**3.47**	**5.41**	0.09	0.09	0.10

TABLE II. Computational Results for the Large-Size Problem Instances

Instances			HBMO						GRASP-PLM								
			Objetive function value			t (s.)			Objetive function value			Deviation (%)			t (s.)		
n	m	O_p	Min.	Avg.	Max.	Min.	Avg.	Max.	Min.	Avg.	Max.	Min.	Avg.	Max.	Min.	Avg.	Max.
50	50	1	9091.40	9149.40	9208.60	0.94	1.00	1.07	**9038.20**	**9046.60**	**9055.80**	**-0.59**	**-1.12**	**-1.66**	0.58	0.60	0.62
		250	14852.20	15071.80	15333.00	0.89	0.94	0.99	**13415.60**	**13498.14**	**13607.60**	**-9.70**	**-10.48**	**-11.30**	0.56	0.58	0.59
		454	19277.60	20717.64	22103.20	0.85	0.94	1.11	**16770.40**	**16991.82**	**17205.20**	**-12.96**	**-17.95**	**-22.10**	0.53	0.55	0.57
		500	19510.60	20606.24	21977.20	0.92	1.06	1.24	**16653.60**	**16962.30**	**17270.40**	**-14.56**	**-17.65**	**-21.33**	0.53	0.55	0.58
		1000	20792.20	23732.60	26991.40	1.10	2.36	4.68	**20391.60**	**21086.86**	**21981.00**	**-4.20**	**-11.98**	**-19.96**	0.52	0.55	0.57
100	50	1	16384.60	16457.72	16541.00	2.17	2.27	2.38	**15908.80**	**15935.32**	**15968.00**	**-2.87**	**-3.15**	**-3.45**	2.49	2.52	2.55
		250	24173.00	24686.96	25271.20	2.25	2.44	2.75	**21545.00**	**21782.50**	**21950.60**	**-10.62**	**-11.56**	**-12.97**	2.18	2.26	2.32
		500	31267.40	32430.72	33239.60	2.25	2.63	3.44	**26571.80**	**27178.06**	**27610.40**	**-15.03**	**-16.21**	**-16.95**	2.12	2.18	2.24
		964	35864.00	37897.32	40118.40	3.23	6.23	11.24	**31073.60**	**32194.08**	**33181.20**	**-13.30**	**-15.03**	**-17.21**	2.04	2.07	2.13
		1000	36165.20	38626.20	41159.60	3.41	5.42	8.86	**31509.40**	**32783.08**	**33630.60**	**-12.65**	**-14.94**	**-18.08**	2.04	2.10	2.15
100	100	1	13024.00	13124.72	13226.40	2.58	2.68	2.76	**12783.20**	**12804.74**	**12835.40**	**-1.84**	**-2.43**	**-2.95**	4.02	4.07	4.11
		13	13023.20	13097.48	13161.80	2.61	2.68	2.76	**12856.20**	**12884.52**	**12930.20**	**-1.28**	**-1.62**	**-1.75**	4.04	4.09	4.13
		250	24269.00	24911.24	25434.60	2.47	2.58	2.70	**21164.80**	**21417.90**	**21599.40**	**-12.79**	**-14.03**	**-15.08**	3.62	3.74	3.84
		500	31830.00	33783.60	36248.20	2.44	2.64	2.87	**26217.40**	**26825.74**	**27344.60**	**-17.59**	**-20.57**	**-24.55**	3.50	3.63	3.74
		1000	37435.60	40444.72	44159.60	3.16	4.46	6.19	**32248.20**	**33400.64**	**34472.20**	**-13.63**	**-17.17**	**-21.51**	3.41	3.54	3.69

TABLE III. Computational Results for Problem Instances Adapted From the TSPLIB [9]

Instances				CPLEX			GRASP-PLM								
							Objetive function value			Deviation (%)			t (s.)		
name	n	m	O_p	Obj.	Gap (%)	t (s.)	Min.	Avg.	Max.	Min.	Avg.	Max.	Min.	Avg.	Max.
burma14	11	3	1	57.0	0.00	0.13	57.00	57.00	57.00	0.00	0.00	0.00	0.05	0.08	0.31
			250	323.0	0.00	0.17	323.00	323.00	323.00	0.00	0.00	0.00	0.05	0.08	0.23
			500	569.0	0.00	0.11	569.00	569.00	569.00	0.00	0.00	0.00	0.04	0.05	0.06
			1000	1070.0	0.00	0.07	1070.00	1070.00	1070.00	0.00	0.00	0.00	0.05	0.06	0.07
ulises22	1	5	1	209.0	0.00	9.10	215.00	215.00	215.00	2.87	2.87	2.87	0.14	0.15	0.15
			250	520.0	0.00	13.48	520.00	520.00	520.00	0.00	0.00	0.00	0.12	0.13	0.14
			500	786.0	0.00	3.92	786.00	786.10	787.00	0.00	0.01	0.13	0.12	0.13	0.13
			1000	1345.0	0.00	5.05	1345.00	1345.00	1345.00	0.00	0.00	0.00	0.12	0.13	0.14
dantzig42	32	10	1	2519.0	26.03	3600.00	2243.00	2243.00	2243.00	-10.96	-10.96	-10.96	0.67	0.70	0.85
			250	3331.0	10.60	3600.00	3195.00	3284.20	3357.00	-4.08	-1.40	0.78	0.60	0.64	0.66
			500	3574.0	0.00	1809.00	3600.00	3633.50	3723.00	0.73	1.66	4.17	0.60	0.63	0.67
			1000	4536.0	0.00	1294.07	4536.00	4551.30	4571.00	0.00	0.34	0.77	0.56	0.61	0.85
hk48	36	12	1	33158.0	20.60	3600.00	33833.00	33956.00	34053.00	2.04	2.41	2.70	0.92	1.01	1.15
			250	41256.0	30.85	3600.00	36405.00	36441.50	36482.00	-11.76	-11.67	-11.57	0.91	0.94	0.96
			500	42324.0	31.64	3600.00	35588.00	35596.40	35609.00	-15.92	-15.90	-15.87	0.86	0.90	0.94
			1000	41308.0	27.41	3600.00	37827.00	38205.10	38378.00	-8.43	-7.51	-7.09	0.90	0.95	1.01
lin105	79	26	1	—	—	—	59713.00	60758.70	61868.00	—	—	—	7.72	7.96	8.26
			250	—	—	—	67029.00	67176.80	67332.00	—	—	—	7.96	8.52	10.00
			500	—	—	—	65138.00	66055.60	67283.00	—	—	—	7.64	8.02	8.53
			1000	—	—	—	67328.00	68483.50	69016.00	—	—	—	7.63	7.93	8.51
pr152	114	38	1	—	—	—	442586.00	459030.10	471043.00	—	—	—	23.77	24.46	25.95
			250	—	—	—	432127.00	447690.30	463756.00	—	—	—	23.99	25.05	25.91
			500	—	—	—	477308.00	496260.80	505313.00	—	—	—	25.15	25.39	25.72
			1000	—	—	—	461980.00	474970.20	490027.00	—	—	—	24.17	24.61	24.98

Column *CPLEX* in Table I reports the objective function value (*Opt.*) and computational time (t (*s.*)), measured in seconds, required by CPLEX when solving the optimization model. Columns *HBMO* and *GRASP-PLM* show the results obtained by the approximate techniques. In each case, the deviation (*Deviation* (%)) in terms of objective function value in comparison with the solutions obtained by CPLEX and the computational times (t (*s.*)) are shown. The minimum (*Min.*), average (*Avg.*), and maximum (*Max.*) deviations are reported in both cases. Similarly, the minimum (*Min.*), average (*Avg.*), and maximum (*Max.*) computational times used by the techniques are shown.

As can be checked in Table I, GRASP-PLM outperforms HBMO in terms of quality of the solutions and computational time. It should be noted that the deviations corresponding to the worst solutions reported by GRASP-PLM are still better than the average deviations reported by HBMO. Concerning the computational time, GRASP-PLM exhibits a competitive performance in comparison with HBMO and CPLEX. In this regard, GRASP-PLM maintains a stable temporal performance, requiring at most about 0.11 seconds on average. This computational advantage added to the relevant robustness shown by GRASP-PLM in terms of average deviations and difference between the best and the worst average deviations, makes our algorithm a competitive and suitable approach when tackling scenarios of this size.

Table II shows the computational results for the large-size problem instances. In this case, due to the fact that CPLEX is not able to provide a feasible solution, only the results obtained by the approximate approaches are reported. Namely, columns *HBMO* and *GRASP-PLM*. The objective function value (*Objective function value*) and computational time (t (*s.*)) are provided for each one. These columns include the minimum (*Min.*), average (*Avg.*), and maximum (*Max.*) computed values based upon the 10 executions. Moreover, we also provide the deviation (*Deviation* (*%*)) in terms of objective function value calculated in comparison with those objective function values provided by HBMO.

The computational results reported in Table II indicate that GRASP-PLM clearly improves HBMO on the basis of the quality of the solutions found. It should be noted that GRASP-PLM presents an average improvement of about 20% for a group of problem instances. Even though the computational times required by both methods are quite similar, it should be highlighted that GRASP-PLM presents a stable performance in terms of the difference between minimum and maximum computational times. On the other hand, as can be checked, HBMO reports the worst performance in this aspect. Hence, at the light of these results, it concludes that GRASP-PLM is also a competitive approach for large-size scenarios.

With the goal of checking if there are significant differences between the results obtained by the GRASP-PLM in comparison with those returned by the HBMO, the *paired-sample Wilcoxon test* is applied [47]. This test is performed with the results from the experiments summarized in Tables I and II, with a significance level of 0.05. This test concludes that there are significant differences between the solutions reported by both algorithms.

The performance of the GRASP-PLM has also been checked on representative problem instances adapted from other problems. In this case, the instances included in the TSPLIB [9] for the well-known Travelling Salesman Problem [44]. Concretely, a subset of points is randomly selected to be potential locations in which to set up plants, whereas the remaining points are customers.

The computational results for the instances adapted from TSPLIB are reported in Table III. As can be checked, as long as the size of the instances increases the performance of CPLEX in terms of quality of the solutions is compromised. Specifically, CPLEX is not able to provide a feasible solution for the largest instances (lin105 and pr152) within a time limit of 3600 seconds. Nevertheless, GRASP-PLM provides a feasible solution in all the cases. For the problem instances where both, CPLEX and GRASP-PLM, provide feasible solutions, we report the deviation of GRASP-PLM with respect to the best solution provided by CPLEX. In this regard, although GRASP-PLM is not able to provide the best solution in some cases, its temporal performance greatly outperforms CPLEX and maintains a similar performance regardless the variation of the value of o_p.

VII. Conclusions and Further Research

In this paper, a mathematical model and a metaheuristic approach based on Greedy Randomized Adaptive Search Procedure with a Probabilistic Learning Model (GRASP-PLM) for solving the Uncapacitated Plant Cycle Location Problem (UPCLP) have been studied. In order to evaluate their performances, extensive computational experiments over a wide range of problem instances from a benchmark suite proposed in the related literature is performed. Thus, these problem instances have been solved using both, the mathematical model implemented in a general purpose solver (i.e., CPLEX) and our GRASP-PLM. Moreover, we also report a comparison for these problem instances with an approximate approach published in the literature based on the Honey Bees Mating Optimization algorithm (HBMO) [42], which returns good quality solutions in short computational times when solving the UPCLP. Finally, a new set of instances from the well-known TSPLIB has been adapted in order to evaluate the performance of our approaches in different structured instances.

The computational results show that our GRASP-PLM exhibits a competitive performance in terms of computational times for the small- and medium-size problem instances in comparison with the computational times required by CPLEX and HBMO. Specifically, GRASP-PLM outperforms HBMO on the basis of the average objective function value. This improvement becomes even more substantial when tackling large-size problem instances, where CPLEX is not even able to provide a feasible solution. Unlike CPLEX, GRASP-PLM provides high-quality solutions through short computational times. It also reports a stable and slight increase in computational time when the problem size increases.

Considering the computational results provided in this work, we can claim that GRASP-PLM is an advisable algorithm for tackling the UPCLP in practical environments, being especially suitable in large-scale scenarios. It provides high-quality solutions by means of short computational times, in the range of a few seconds. Another outstanding characteristic exhibited by GRASP-PLM for this problem is that the variance of time and quality is quite stable in terms of the difference between the minimum and maximum values.

For future work, we intend to extend this model and algorithm to emergency scenarios. In these scenarios, we usually have several types of locations and vehicles and we have to schedule them for providing care to the victims as soon as possible. Lastly, the UPCLP can be studied assigning priorities to the customers, which can be applied also on emergency scenarios.

Acknowledgements

This work has been partially funded by the Spanish Ministry of Economy and Competitiveness with FEDER funds (project code PID2019-104410RB-I00/AEI/10.13039/501100011033).

References

[1] G. Cornuéjols, G. Nemhauser, and L. Wolsey, "The uncapacitated facility location problem," tech. rep., Cornell University Operations Research and Industrial Engineering, 1983.

[2] W. Ho, P. Ji, and P. K. Dey, "A multi-depot travelling salesman problem and its iterative and integrated approaches," *International Journal of Operational Research*, vol. 1, no. 4, pp. 382–395, 2006.

[3] S. V. Ukkusuri and W. F. Yushimito, "Location routing approach for the humanitarian prepositioning problem," *Transportation research record*, vol. 2089, no. 1, pp. 18–25, 2008.

[4] S. Schwarze, E. Lalla-Ruiz, and S. Voß, "Modeling the capacitated p-cable trench problem with facility costs," *Central European journal of operations research*, pp. 1–23, 2020.

[5] E. Lalla-Ruiz and S. Voß, "A popmusic approach for the multi-depot cumulative capacitated vehicle routing problem," *Optimization Letters*, vol. 14, no. 3, pp. 671–691, 2020.

[6] R. T. Berger, C. R. Coullard, and M. S. Daskin, "Location-routing problems with distance constraints," *Transportation Science*, vol. 41, no. 1, pp. 29–43, 2007.

[7] S. Çetiner, C. Sepil, and H. Süral, "Hubbing and routing in postal delivery systems," *Annals of Operations research*, vol. 181, no. 1, pp. 109–124, 2010.

[8] A. Bruns, A. Klose, and P. Stähly, "Restructuring of swiss parcel delivery services," *OR-Spektrum*, vol. 22, no. 2, pp. 285–302, 2000.

[9] G. Reinelt, "Tsplib–a traveling salesman problem library," *ORSA journal on computing*, vol. 3, no. 4, pp. 376–384, 1991.

[10] H. Min, V. Jayaraman, and R. Srivastava, "Combined location-routing problems: A synthesis and future research directions," *European Journal of Operational Research*, vol. 108, no. 1, pp. 1 – 15, 1998.

[11] D. Tuzun and L. Burke, "A two-phase tabu search approach to the location routing problem," *European Journal of Operational Research*, vol. 116, no. 1, pp. 87 – 99, 1999.

[12] A. Balakrishnan, J. Ward, and R. Wong, "Integrated facility location and vehicle routing models: Recent work and future prospects," *American Journal of Mathematical and Management Sciences*, vol. 7, no. 1-2, pp. 35–61, 1987.

[13] G. Nagy and S. Salhi, "Location-routing: Issues, models and methods," *European Journal of Operational Research*, vol. 177, no. 2, pp. 649 – 672, 2007.

[14] C. Prodhon and C. Prins, "A survey of recent research on location-routing problems," *European Journal of Operational Research*, vol. 238, no. 1, pp. 1 – 17, 2014.

[15] M. Schneider and M. Drexl, "A survey of the standard location-routing problem," *Annals of Operations Research*, vol. 259, no. 1-2, pp. 389–414, 2017.

[16] M. Drexl and M. Schneider, "A survey of variants and extensions of the location-routing problem," *European Journal of Operational Research*, vol. 241, no. 2, pp. 283 – 308, 2015.

[17] S. Alumur and B. Y. Kara, "A new model for the hazardous waste location-routing problem," *Computers & Operations Research*, vol. 34, no. 5, pp. 1406–1423, 2007.

[18] X. Yu, Y. Zhou, and X.-F. Liu, "The two-echelon multi-objective location routing problem inspired by realistic waste collection applications: The composable model and a metaheuristic algorithm," *Applied Soft Computing*, p. 106477, 2020.

[19] S. Kumar, V. Kumar-Solanki, S. K. Choudhary, A. Selamat, and R. González-Crespo, "Comparative study on ant colony optimization (aco) and k-means clustering approaches for jobs scheduling and energy optimization model in internet of things (iot)," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 1, pp. 107–116, 2020.

[20] M. Rabbani, R. Heidari, H. Farrokhi-Asl, and N. Rahimi, "Using metaheuristic algorithms to solve a multi-objective industrial hazardous waste location-routing problem considering incompatible waste types," *Journal of Cleaner Production*, vol. 170, pp. 227–241, 2018.

[21] A. Billionnet, S. Elloumi, and L. G. Djerbi, "Designing radio-mobile access networks based on synchronous digital hierarchy rings," *Computers & operations research*, vol. 32, no. 2, pp. 379–394, 2005.

[22] S. Zhang, M. Chen, and W. Zhang, "A novel location-routing problem in electric vehicle transportation with stochastic demands," *Journal of Cleaner Production*, vol. 221, pp. 567–581, 2019.

[23] I. A. Martínez-Salazar, J. Molina, F. Ángel-Bello, T. Gómez, and R. Caballero, "Solving a bi-objective transportation location routing problem by metaheuristic algorithms," *European Journal of Operational Research*, vol. 234, no. 1, pp. 25–36, 2014.

[24] M. Borhani, K. Akbari, A. Matkan, and M. Tanasan, "A multicriteria optimization for flight route networks in large-scale airlines using intelligent spatial information," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 1, pp. 123 – 131, 2019.

[25] B. Jarboui, H. Derbel, S. Hanafi, and N. Mladenović, "Variable neighborhood search for location routing," *Computers & Operations Research*, vol. 40, no. 1, pp. 47–57, 2013.

[26] C.-J. Ting and C.-H. Chen, "A multiple ant colony optimization algorithm for the capacitated location routing problem," *International Journal of Production Economics*, vol. 141, no. 1, pp. 34–44, 2013.

[27] T.-H. Wu, C. Low, and J.-W. Bai, "Heuristic solutions to multi-depot location-routing problems," *Computers & Operations Research*, vol. 29, no. 10, pp. 1393–1415, 2002.

[28] P. Yang and B. Li-Jun, "An integrated optimization problem in logistics and the pso solution," in *2006 International Conference on Service Systems and Service Management*, vol. 2, pp. 965–970, IEEE, 2006.

[29] Y. Marinakis and M. Marinaki, "A particle swarm optimization algorithm with path relinking for the location routing problem," *Journal of Mathematical Modelling and Algorithms*, vol. 7, no. 1, pp. 59–78, 2008.

[30] R. Caballero, M. González, F. M. Guerrero, J. Molina, and Paralera, "Solving a multiobjective location routing problem with a metaheuristic based on tabu search. application to a real case in andalusia," *European Journal of Operational Research*, vol. 177, no. 3, pp. 1751–1763, 2007.

[31] C. Prins, C. Prodhon, and R. W. Calvo, "Solving the capacitated location-routing problem by a grasp complemented by a learning process and a path relinking," *4OR*, vol. 4, no. 3, pp. 221–238, 2006.

[32] S. Barreto, C. Ferreira, J. Paixao, and B. S. Santos, "Using clustering analysis in a capacitated location-routing problem," *European Journal of Operational Research*, vol. 179, no. 3, pp. 968–977, 2007.

[33] I. López-Plata, C. Expósito-Izquierdo, E. Lalla-Ruiz,B. Melián-Batista, and J. M. Moreno-Vega, "A greedy randomized adaptive search procedure for solving the uncapacitated plant cycle problem," in *International Conference on Computer Aided Systems Theory*, pp. 263–270, Springer, 2015.

[34] J. Escobar, R. Linfati, M. Baldoquin, and P. Toth, "A granular variable tabu neighborhood search for the capacitated location-routing problem," *Transportation Research Part B: Methodological*, vol. 67, no. 0, pp. 344 – 356, 2014.

[35] R. Baldacci, A. Mingozzi, and R. Calvo, "An exact method for the capacitated location-routing problem.," *Operations Research*, vol. 59, no. 5, pp. 1284–1296, 2011.

[36] J. Belenguer, E. Benavent, C. Prins, C. Prodhon, and R. Calvo, "A branch-and-cut method for the capacitated location-routing problem," *Computers & Operations Research*, vol. 38, no. 6, pp. 931 – 941, 2011.

[37] C. Contardo, J.F. Cordeau, and B. Gendron, "An exact algorithm based on cut-and-column generation for the capacitated location-routing problem," *INFORMS Journal on Computing*, vol. 26, pp. 88–102, 2014.

[38] A. Perugia, L. Moccia, J. Cordeau, and G. Laporte, "Designing a home-to-work bus service in a metropolitan area," *Transportation Research Part B: Methodological*, vol. 45, no. 10, pp. 1710 – 1726, 2011.

[39] I. Rodríguez-Martín, J.-J. Salazar-González, and H. Yaman, "A branch-and-cut algorithm for two-level survivable network design problems," *Computers & Operations Research*, vol. 67, pp. 102–112, 2016.

[40] M. Labbé, I. Rodríguez-Martín, and J. Salazar-Rodríguez, "A branch-and-cut algorithm for the plant-cycle location problem," *Journal of the Operational Research Society*, vol. 55, no. 5, pp. 513 – 520, 2004.

[41] M. Albareda-Sambola, J. Díaz, and E. Fernández, "A compact model and tight bounds for a combined location-routing problem," *Computers & Operations Research*, vol. 32, no. 3, pp. 407 – 428, 2005.

[42] B. Melián-Batista, J. Moreno-Vega, N. Vaswani, and R. Yumar, "A nature inspired approach for the uncapacitated plant cycle location problem," in *Nature Inspired Cooperative Strategies for Optimization (NICSO 2008)* (N. Krasnogor, M. B. Melián-Batista, J. A. Moreno-Pérez, J. M. Moreno- Vega, and D. A. Pelta, eds.), vol. 236 of *Studies in Computational Intelligence*, pp. 49–60, Springer Berlin Heidelberg, 2009.

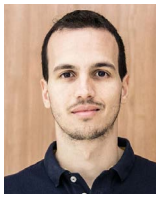
[43] C. Fleming, S. Griffis, and J. Bell, "The effects of triangle inequality on the vehicle routing problem," *European Journal of Operational Research*, vol. 224, no. 1, pp. 1 – 7, 2013.

[44] G. Gutin and A. Punnen, *The Traveling Salesman Problem and Its Variations*. Springer, 1 edition ed., 2002.

[45] T. Feo and M. Resende, "Greedy randomized adaptive search procedures," *Journal of Global Optimization*, vol. 6, no. 2, pp. 109–133, 1995.

[46] W. W. Daniel, *Applied nonparametric statistics*. Boston: PWS-Kent Publishing Company, 1990.

[47] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.



Israel López-Plata

He received the Computer Science Engineering degree in 2009 and the Ph.D. in Computer Science in 2020, both from the University of La Laguna, Spain. He has 12 year experience in software development related with logistics and Artificial Intelligence. Currently he is working as professor on Department of Computer and Systems Engineering, University of La Laguna. His research is focused on optimization applied to the resolution of logistic problems, in addition to other Artificial Intelligence techniques.

Christopher Expósito-Izquierdo

Christopher Expósito-Izquierdo is engineering in Computer Science and has obtained his Ph.D. on the optimization of the transshipment operations in maritime container terminals at the University of La Laguna. He is nowadays lecturer at the Department of Computer and Systems Engineering, University of La Laguna, in which his activity is mainly dedicated to the analysis and optimisation of logistic operations and transportation, especially focused on maritime container terminals.

Eduardo Lalla-Ruiz

Eduardo Lalla-Ruiz is an assistant professor at the Department of High Tech and Entrepreneurship in the Industrial Engineering and Business Information Systems (IEBIS) section at the University of Twente (The Netherlands). He holds a bachelor degree in industrial engineering, an advanced engineering degree in industrial automation, and a master of advanced studies in computer sci-ence and artificial intelligence from the University of La Laguna (Spain). At that same institution, Eduardo completed his Ph.D. and was awarded an extraordinary doctoral distinction. After finishing his Ph.D., Eduardo joined the Institute of Information Systems (University of Hamburg, Germany) as a researcher and lecturer. There, Eduardo became a fellow of the prestigious Alexander von Humboldt Research Foundation. Eduardo has published several research papers in high-quality journals, collaborated in international research projects, served as editor, guest editor, and referee in relevant journals, and participated in the organization of sessions and international conferences. Recently, he was awarded as runner up for the EURO excellence in practice award 2021 as well as the meritorious service award in Transportation Science (INFORMS). His research interests lie in the fields of operations research, logistics, artificial intelligence, and mathematical programming.

Belén Melián-Batista

She received the M.Sc. degree in mathematics and the Ph.D. in computer science from the University of La Laguna, Canary Islands, Spain, in 1999 and 2003, respectively. She is a Full Professor in Computer Science and Artificial Intelligence at the Department of Computer and Systems Engineering, University of La Laguna, Spain. She has published more than 30 scientific papers in leading scientific journals, and has been Advisor of 4 Ph.D. dissertations. She has been Principal Researcher in a variety of national research projects. Her research focuses on the design and implementation of metaheuristic techniques for solving combinatorial optimization problems in different application areas. Specifically, her current main interest is the application of these techniques to solve logistic and transportation problems.

J. Marcos Moreno-Vega

He received the M.Sc. degree in Mathematics in 1990 and the Ph.D. degree in Computer Science in 1997, both from the University of La Laguna, Canary Islands, Spain. Currently, he is a Full Professor at the Department of Computer and Systems Engineering, University of La Laguna, in the area of Computer Science and Artificial Intelligence. His research interests include soft computing, metaheuristics and data mining applied to logistic problems.

Resource and Process Management With a Decision Model Based on Fuzzy Logic

J. T. Fornerón Martínez[1], F. Agostini[2], D. L. La Red Martínez[2]

[1] Faculty of Applied Sciences, National University of Pilar, Pilar (Paraguay)
[2] Faculty of Exact and Natural Sciences and Surveying, Northeastern National University, Corrientes (Argentine)

Received 28 June 2020 | Accepted 3 February 2023 | Published 14 February 2023

Abstract

The allocation of the resources to be shared in the context of a distributed processing system needs to be coordinated through the mutual exclusion mechanism, which will decide the order in which the shared resources will be allocated to those processes that require them. This paper proposes an aggregation operator, which can be used by a module that manages the shared resources, whose function is to assign the resources to the processes according to their requirements (shared resources) and the status of the distributed nodes in which the processes operate (computational load), by using 2-tuple associated to linguistic labels.

Keywords

Aggregation Operators, Communication Between Groups Of Processes, Computing With Word, Fuzzy Logic, Mutual Exclusion, Operating Systems.

DOI: 10.9781/ijimai.2023.02.009

I. Introduction

Distributed systems, composed of multiple nodes and multiple processes, cooperatively performing a given function, require the use of decision models that allow the use of shared resources to groups of processes that require them, accessed through the mutual exclusion mode.

Solutions proposed for this problem are found in [1] and [2], where the main synchronization algorithms in distributed systems are described. [3] presents an efficient solution, also fault-tolerant, for the problem of distributed mutual exclusion. [4], [5] and [6] present algorithms for the management of mutual exclusion in computer networks. [7] focuses on the main algorithms for the management of distributed processes, distributed mutual exclusion and distributed global states.

Solutions proposed for this problem are found in [1] and [2], where the main synchronization algorithms in distributed systems are described. [3] presents an efficient solution, also fault-tolerant, for the problem of distributed mutual exclusion. [4], [5] and [6] present algorithms managing mutual exclusion in computer networks. [7] focuses on the main algorithms for the management of distributed processes, distributed mutual exclusion and distributed global states.

In distributed systems, the allocation of resources to processes must be done considering the priorities of the processes and the workload status of the computational nodes in which the processes are executed.

Besides, solutions that we could call classic for several types of distributed systems have been proposed in [8], [9], [10], [11] and [12]. Also, in [13] and [14] works focused on ensuring mutual exclusion are presented. [15] presents an interesting distributed solution based on permissions and [16] a solution based on process priorities. A solution using consensus in resource allocation is presented in [17].

There are practical situations in which problems must be solved by having vague and imprecise information. This means that information is not always evaluated accurately with quantitative values, but with qualitative values. This was solved by [18], by incorporating the concept of the linguistic variable and applying it to decision making as well as explained in [19]. Also [20] expressed that computing with word (CWW) is a methodology in which words are used instead of numbers for computation and reasoning and where fuzzy logic plays a fundamental role in CWW and vice versa.

As mentioned in [21], based on the concept of symbolic translation, Herrera and Martínez proposed the 2-tuple linguistic representation model, which expresses linguistic assessment information using the linguistic 2-tuple (s_i, α), where the semantic element s_i is a linguistic label from a predefined linguistic variable, S, and α is a numerical value that represents the symbolic translation. Also, Zhang et al. introduced interval-valued hesitant fuzzy soft sets by combining the interval-valued hesitant fuzzy set and soft set models and evaluated their operations. Tao et al. presented the 2-tuple linguistic soft set method, incorporating the 2-tuple linguistic term set and soft set, to solve complex group decision-making problems. Today, soft set methods are widely applied to solve real-life decision-making problems (e.g., Ali and Shabir; Chang; Deli and Cagman; Tang; Chang; Chang et al.). The use of various algorithms to counteract uncertainty and incomplete

* Corresponding author.
E-mail addresses: jforneron@aplicadas.edu.py (J. Fornerón), fagostini@conicet.gov.ar (F. Agostini), lrmdavid@exa.unne.edu.ar (D. La Red Martínez)

information when trying to solve multi-criteria decision-making problems (MCDM), where generally precise values and a single set of linguistic terms are insufficient to deal with the complexity of selection problems, where experts hesitate among sets of linguistic terms to determine the values of evaluation attributes in said problems.

Different words can have different meanings for different people. Answer about how can a Computing With Words (CWW) engine be validated, what Fuzzy Set models should be used or what choices should be made to keep the design of the CWW engine as simple as possible, are analyzed and founded in [22].

The new decision models for allocating shared resources could be executed in the context of a shared resource manager for the distributed system, which would receive the shared resource requirements of the processes running on the different distributed nodes, as well as the computational load state of the nodes.

It has been worked with fuzzy variables using linguistic labels and 2-tuple to avoid losing precision in computing with words.

A computational model has been presented in [23], called 2-tuple linguistic computational model, in which a parameter was incorporated to the basic linguistic representation to improve the accuracy of linguistic calculations.

The fuzzy linguistic approach, although with the limitations at the moment of being used in fusion processes on the linguistic values, is used successfully in the resolution of many problems and presents tools to improve the application of the fuzzy linguistic approach, in relation to the loss of information caused by the need to express the results in the initial expression domain, which is discrete through an approximate process and implies a lack of precision in the final results of the fusion of linguistic information. Linguistic information is expressed through a 2-tuple, composed of a linguistic term and a numerical value evaluated at [-0.5, 0.5], which allows to represent the information obtained in an aggregation process. Together with the 2-tuple representation, a computational technique for word computing (CWW) is developed [24].

A clear explanation about main CW concepts can be found in [25]: granules and linguistic variables. A granule is defined as a clump of objects (or points) which are drawn together by indistinguishability, similarity, proximity, or functionality.

An example of granularity is a system that is composed of several smaller subsystems and these smaller subsystems are in turn divided into even smaller ones. The decomposition of the whole into parts (granulation) is, in general, hierarchical in nature.

A linguistic variable is a variable whose values are not numbers but words or sentences in a natural or artificial language. The main purpose of using linguistic values (words or sentences) instead of numbers is that linguistic characterizations are, in general, less specific than numerical ones, but much closer to the way that humans express and use their knowledge.

Other models have been presented in [23], [26], [27], [28] and [29]. These works show different advantages of this formalism to represent linguistic information over classical models.

It is a continuous linguistic domain, where the linguistic calculation model is based on linguistic tuples and performs word computation processes easily and without loss of information, therefore, the results of the word computation processes are always expressed in the initial linguistic domain.

Due to these advantages, this model of linguistic representation will be used to achieve the development of a procedure for the fusion of linguistic and numerical information.

The 2-tuple model of linguistic representation represents linguistic information by means of a 2-tuple, (s, α). In this work it will be used to represent the load of the nodes, nodal preferences, and final priorities. An example of this can be seen in Table I.

The symbolic translation of a linguistic term $s_i \in S = \{s_o, ..., s_g\}$ consists of a numerical value $\alpha_i \in [-.5, .5)$ that supports the "information difference" between an information count β evaluated in [0, g] obtained after a symbolic aggregation operation (acting on the order index of the labels) and the closest value in $[0, ..., g]$ that indicates the index of the closest linguistic term in S (s_i).

TABLE I. 2-Tuple Linguistic Weighted Average Score

	Criteria					*Aggregate Weight*
	C1	***C2***	***C3***	***C4***	***C5***	
A1	(S5, 0.00)	(S5, 0.31)	(S6, 0.33)	(S5, 0.50)	(S5, 0.00)	
A2	(S5, 0.00)	(S6, 0.00)	(S2, 0.00)	(S5, 0.31)	(S6, 0.00)	(S5, -0.21)
A3	(S5, 0.31)	(S5, 0.50)	(S3, 0.00)	(S6, 0.00)	(S5, 0.00)	(S5, -0.31)
A4	(S5, 0.00)	(S5, 0.50)	(S5, 0.50)	(S5, 0.00)	(S5, 0.31)	(S5, -0.15)
A5	(S4, 0.00)	(S5, 0.50)	(S4, 0.00)	(S5, 0.50)	(S4, 0.00)	(S4, 0.19)

Distributed systems are used in multiple solutions around the world, smart container management, connected smart plants, banking networks, the world wide web, etc., as seen in Fig 1.

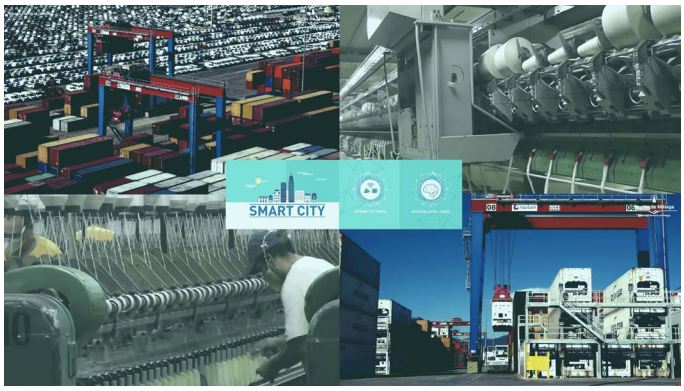


Fig. 1. Samples of the internet of things, smart cities, industry 4.0 (industrial robotics), etc.

In this paper, a new aggregation operator will be presented specifically for the problem. This falls under the category of OWA operators, more specifically Neat OWA. This will present an innovative method for shared resource management in distributed systems.

The structure of this document is as follows: section II gives guidelines about the premises and data structures to be used, section III describes the steps of the proposed aggregation operator, section IV explains details of an example of application of the proposed aggregation operator, section V presents the conclusions and future work and section VI mentions acknowledgement.

II. Data Structures to be Used

The following premises and data structures will be used.

These are groups of processes distributed in process nodes that access critical resources. These resources are shared in the form of distributed mutual exclusion and it must be decided, according to the demand for resources in the processes, what the priorities will be for assigning the resources to the processes that require them (in order to be assigned in the processes, only the available resources will be taken into account, that is, those that have not yet been assigned in certain

processes. All the premises, resources and processes running in the different nodes, groups, cardinals, criteria, and categories to evaluate the different weights and calculations required are those mentioned in [30].

III. Description of the Aggregation Operator

The proposed operator consists of the following steps:

A. Calculation of the current computational load of the nodes.

B. Establishment of the categories of computational load and the vectors of weights associated with them.

C. Calculation of the priorities or preferences of the processes considering the state of the node (they are calculated in each node for each process).

D. Expression of the calculated values in terms of 2-tuple using a set of linguistic labels.

E. Calculation of the priorities or preferences of the processes to access the shared resources available and determination of the order and to which process the resources will be allocated.

Each of the steps above is described below.

A. Calculation of the Current Computational Load of the Nodes

To obtain an indicator of the current computational load of each node, different criteria can be adopted; in this proposal the criteria will be the percentage of CPU usage, the percentage of memory usage and the percentage of use of input / output operation. The computational load of each node, the number of criteria to determine the load of the nodes, the criteria that apply and the calculation of the computational load of each node, are those mentioned in [30].

B. Establishment of the Categories of Computational Load and of the Vectors of Weights Associated Thereto

The current computational load categories of each node, the number of categories to determine the load of the nodes, the categories that apply, the vectors of weights associated with the current computational load categories of each node. In this proposal, the criteria will be those used in [30].

Establishment of vectors of weights (same for all nodes): weights = $\{w_{ij}\}$ con $i = 1, ..., a$ (categories number of computational load) y $j = 1, ..., e$ (maximum number of criteria).

C. Calculation of the Priorities or Preferences of the Processes Considering the Status of the Node (They Are Calculated in Each Node for Each Process and Could Be Called Nodal Priorities)

These priorities are calculated at each node for each resource request originated in each process; the calculation considers the corresponding weight vector according to the current load of the node and the vector of the values granted by the node according to the evaluation criteria of the request.

The valuation vectors that will be applied for each request of a resource by a process, according to the criteria established for the determination of the priority that in each case and moment will fix the node in which the request occurs, are the following: *valuations* $(r_{ij}\, p_{kl}) = \{cp_m\}$ con $i = 1, ..., n$ (node where the resource resides), $j = 1, ..., r$ (resource on node i), $k = 1, ..., n$ (node where the process resides), $l = 1, ..., p$ (process at node k) and $m = 1, ..., e$ (valuation criteria of the requirement priority). As can be seen in Table II.

TABLE II. Valuations Assigned to the Criteria to Calculate the Priority or Preference That Each Node Will Give to Each Requirement of Each Process According to the Node Load

Resources - Processes	Criteria				
$r_{11}\, p_{11}$	cp_1	...	cp_m	...	cp_e
...	...	...	...	...	...
$r_{ij}\, p_{kl}$	cp_1	...	cp_m	...	cp_e
...	...	...	...	...	...
$r_{nr}\, p_{np}$	cp_1	...	cp_m	...	cp_e

D. Expression of the Calculated Values in Terms of 2-Tuple Using a Set of Linguistic Labels

The valuations expressed in a linguistic format using the linguistic and semantic labels mentioned can be seen in Fig. 2, where in the abscissa are indicated the linguistic labels and in the ordinates the values of probability of belonging to them.

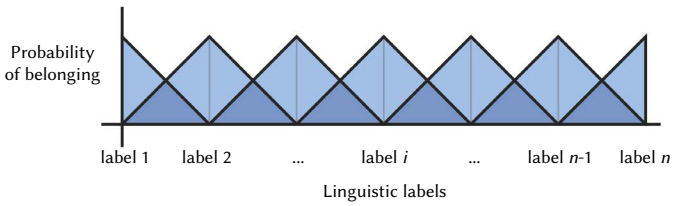


Fig. 2. Representation of the label set.

The next step is to transform these values into the 2-tuple format, considering the linguistic labels proposed above. Therefore, each criteria value will have to be compared with the average value of each label, the minimum difference of that comparison will be the appropriate label [24].

The first element of the 2-tuple will be the linguistic value of that label. The second element will be the difference between the value of the searched criteria and the average value of the selected label.

dm = the minimum difference between the cp_m differences and the most representative value of each language label.

Label valuations $(r_{ij}\, p_{kl})$ = 2-tuple = $T(label_m; dm_m)$ where the sub-index m corresponds to the linguistic labels defined above, as can be seen in Table III.

TABLE III. Valuations Assigned to the Criteria for Calculating the Priority or Preference That Each Node Will Give to Each Requirement of Each Process According to the 2-Tuple Load of the Node

Resources / Processes	2-tuple criteria				
$r_{11}\, p_{11}$	$T(label_1; dm_1)$	...	$T(label_m; dm_m)$	...	$T(label_e; dm_e)$
...	...	...	...	...	...
$r_{ij}\, p_{kl}$	$T(label_1; dm_1)$	...	$T(label_m; dm_m)$	...	$T(label_e; dm_e)$
...	...	...	...	...	...
$r_{nr}\, p_{np}$	$T(label_1; dm_1)$	...	$T(label_m; dm_m)$	...	$T(label_e; dm_e)$

To sum up, the nodal priority (to be calculated at the node where the request occurs) of a process to access a given resource (which can be at any node) is calculated by the scalar product of the mentioned vectors: nodal priority $(r_{ij}\, p_{kl}) = \Sigma\, w_{om} * T(label_m; dm_m) = T(label_n; dm_n) = NPT_{ijkl}$ (Nodal Priority Tuple) with o indicating the weights vector according to the load of the node, all other sub-index maintaining the meanings explained above. With m and n indicating the corresponding linguistic label within the adopted set defined above.

This nodal priority must be transformed into the 2-tuple format, considering the linguistic labels already mentioned. Therefore, it will be necessary to compare each nodal priority value with the average value of each label, the minimum difference of these comparisons will indicate the corresponding label.

E. Calculation of Process Priorities or Preferences to Access Available Shares. In Addition, Determining the Order in Which the Resources Will Be Allocated, and to Which Process Each Resource Will Be Allocated

Table IV is used to calculate the final priorities, in which the priorities or nodal preferences calculated in the previous stage are placed; in this table each row contains the information of the nodal priorities of the different processes to access a certain resource.

Next, it is necessary to calculate the vector of final weights that will be used in the process of aggregation to determine the order or priority of access to the resources.

final weights = $\{wf_{kl}\}$ con $k = 1, ..., n$ (number of nodes) and $l = 1, ..., p$ (Maximum number of processes per node), where np is the number of processes in the system and prg_i is the priority of the process group to which the process belongs (explained in the previous section).

TABLE IV. Nodal Priorities of the Processes to Access Each Resource in 2-Tuple

Resources	2-tuple				
r_{11}	NPT_{1111}	...	NPT_{11kl}	...	NPT_{11np}
...	...	...	...	...	...
r_{ij}	NPT_{ij11}	...	NPT_{ijkl}	...	NPT_{ijnp}
...	...	...	...	...	...
r_{nr}	NPT_{nr11}	...	NPT_{npkl}	...	NPT_{nrnp}

The next step is to normalize the newly obtained weights by dividing each by the sum of all of them.

Thus, a normalized weight vector (in the range of 0 to 1 inclusive) is obtained and with the restriction that the sum of the elements of the vector must give 1:

$\Sigma \{nwf_{kl}\} = 1$ with $k = 1, ..., n$ (number of nodes) and $l = 1,..., p$ (maximum number of processes per node).

The nodal priorities taken row by row for each resource will be scalar multiplied by the normalized final weight vector. In this way, it is possible to obtain each process's final global access priorities to each resource. It is indicated below how the order or priority with which the resources will be allocated is obtained and to which process each one will be assigned.

overall final priority $(r_{ij}\, p_{kl}) = NPT_{ijkl} = FGPT_{ijkl}$ (Final Global Priority Tuple) with r_{ij} indicating the resource j of node i, NPT_{ijkl} is the 2-tuple format, ij indicating the resource j of node i, kl the process l of node k and the product of the overall final priority of the process to access such resource, as can be seen in Table V.

TABLE V. Final Global Priority Tuple

Resources	Nodal Process Priorities				
r_{11}	$FGPT_{1111}$	...	$FGPT_{11kl}$	...	$FGPT_{11np}$
...	...	...	...	...	...
r_{ij}	$FGPT_{ij11}$	...	$FGPT_{ijkl}$	...	$FGPT_{ijnp}$
...	...	...	...	...	...
r_{nr}	$FGPT_{r11}$	...	$FGPT_{pkl}$	...	$FGPT_{rnp}$

The next step is to normalize Table V between extreme values. This will be done using the maximum, minimum and range values calculated from Table VI and represented in Table VII.

TABLE VI. Calculation of the Maximum, Minimum and Range Values

Label	Value
Maximum Value	Maximum ($FGPT_{ijkl}$)
Minimum value	Minimum ($FGPT_{ijkl}$)
Range	Maximum ($FGPT_{ijkl}$) – Minimum ($FGPT_{ijkl}$)

TABLE VII. Normalized Final Global Priority Tuple

Resources	2-tuple				
r_{11}	$NFGPT_{1111}$	...	$NFGPT_{11kl}$	...	$NFGPT_{11np}$
...	...	...	...	...	...
r_{ij}	$NFGPT_{ij11}$	...	$NFGPT_{ijkl}$	...	$NFGPT_{ijnp}$
...	...	...	...	...	...
r_{nr}	$NFGPT_{nr11}$	...	$NFGPT_{npkl}$	...	$NFGPT_{nrnp}$

The greater of these products made for the different processes in relation to the same resource will indicate which of the processes will have access to the resource.

The addition of all these products in relation to the same resource will indicate the priority that will have that resource to be assigned, in relation to the other resources that will also have to be assigned. This is what will be called Linguistic Distributed Systems Assignment Function (*LDSAF*). Refer to (1).

$$LDSAF(r_{ij}) = \Sigma\, NFGPT_{ijkl} = \text{resource allocation priority } r_{ij} \quad (1)$$

By calculating the *LDSAF* for all resources a 2-tuple vector will be obtained, and by ordering its elements from highest to lowest, the priority order of allocation of resources will be obtained. These must be normalized guaranteeing that the 2-tuples obtained are in the interval [0, 1]. The maximum, minimum and range values can be seen in Table VIII.

TABLE VIII. Valuations to Normalize the LDSAF

Label	Value
Maximum Value	Maximum ($LDSAF_{ijkl}$)
Minimum value	Minimum ($LDSAF_{ijkl}$)
Range	Maximum ($LDSAF_{ijkl}$) – Minimum ($LDSAF_{ijkl}$)

In addition, as already indicated, the largest of the products $NFGPT_{ijkl}$ for each resource will indicate the process to which the resource will be assigned.

This is what will be called Normalized Linguistic Distributed Systems Assignment Function (*NLDSAF*). Refer to (2).

$$NLDSAF(r_{ij}) = \Sigma\, (NFGPT_{ijkl}/(\text{maximum } (NFGPT_{ijkl}) - \text{minimum } (NFGPT_{ijkl}))) = r_{ij} \text{ resource allocation priority normalized between extreme values} \quad (2)$$

This can be seen in Table IX.

TABLE IX. Order or Final Priority of Assignment of Resources and Process to Which Is Allocated Each Resource in the First Iteration

Order of allocation of resources	*Process to which the resource will be assigned*
1°: *r*ij of the Max(*NLDSAF*(*r*ij))	p_{kl} of the Max(NFGPT$_{ijkl}$) for the selected r_{ij}
2°: *r*ij of the Max(*NLDSAF*(*r*ij)) for unallocated r_{ij}	p_{kl} of the Max(NFGPT$_{ijkl}$) for the selected r_{ij}
...	...
last: r_{ij} no assigned	p_{kl} of the Max(NFGPT$_{ijkl}$) for the selected r_{ij}

The next step is to repeat the procedure but removing the requests of already made allocations; it must be noted that the assigned resources will be available once they are released by the processes, and can therefore be allocated to other processes. Table VIII should be recalculated by omitting the resource allocations already done.

F. Considerations for Aggregation Operations

The characteristics of the aggregation operations described allow to consider that the proposed method belongs to the family of aggregation operators Neat-OWA, which are characterized by [31] and [32].

The values of the variables are expressed by sets of linguistic labels and 2-tuples [33], thus generalizing the model proposed in [30].

IV. Example and Discussion of Results

This section will explain in detail an example of application of the proposed aggregation operator. The distributed processing system, premises, resources, and processes running in the different nodes, groups, cardinals, criteria, and categories to evaluate the different loads and calculations needed, are those mentioned in [30], corresponding to steps A and B.

Calculation of the priorities or preferences of the processes taking the status of the node into account (they are calculated in each node for each process and could be called nodal priorities).

The valuation vectors are applied for each requirement of a resource made by a process, according to the criteria established for the determination of the priority that in each case and moment fixes the node in which the request occurs.

A. Expression of the Calculated Values in Terms of 2-Tuple Using a Set of Linguistic Labels

The valuations expressed in a linguistic format using the linguistic and semantic labels mentioned, with minimum, medium and maximum values, can be seen in Fig. 2 and Table X.

TABLE X. Proposals for Priority Assessment

EH: Extremely High	0.83	1.00	1.00
VH: Very High	0.67	0.83	1.00
H: High	0.50	0.67	0.83
M: Medium	0.33	0.50	0.67
L: Low	0.17	0.33	0.50
VL: Very Low	0.00	0.17	0.33
EL: Extremely Low	0.00	0.00	0.17

The next step is to transform these values into the 2-tuple format, considering the linguistic labels proposed above. Therefore, it will be necessary to compare each criterion value with the average value of each label, the minimum difference of this comparison will be the appropriate label.

The minimum difference between the differences of each criterion and the most representative value of each language label will be the most representative value.

The first element of the 2-tuple will be the linguistic value of that label, while the second element will be the difference between the value of the searched criterion and the average value of the selected label. This can be seen in Table XI, where "Process Priority" is represented by the methods considered traditional [2], [3] and [4].

As mentioned in the previous stage, each vector of evaluations of each requirement is scalar multiplied by the vector of weights corresponding to the current load category of the node to obtain the priority according to each criterion and the nodal priority granted to each requirement. This can be seen in Table XII.

TABLE XI. The Valuations Assigned to the Criteria to Calculate the Priority or Nodal Preference That Each Node Will Grant Each Requirement of Each Process According to the Node Load

Res./Proc.	Criteria				
	%CPU	...	Process Priority	...	%VM
$r_{11}p_{11}$	T(EB;0.0250)	...	T(EB;0.0800)	...	T(M;-0.0500)
$r_{12}p_{11}$	T(EB;0.0350)	...	T(EB;0.0300)	...	T(MB;0.0833)
$r_{21}p_{11}$	T(EB;0.0200)	...	T(MB;-0.0767)	...	T(MB;-0.0667)
$r_{22}p_{11}$	T(EB;0.0250)	...	T(EB;0.0800)	...	T(MB;0.0333)
$r_{23}p_{11}$	T(EB;0.0300)	...	T(MB;-0.0717)	...	T(B;0.0667)
$r_{24}p_{11}$	T(EB;0.0250)	...	T(EB;0.0600)	...	T(MB;-0.0667)
...	...	...	...	...	...
$r_{11}p_{13}$	T(EB;0.0350)	...	T(EB;0.0600)	...	T(M;-0.0500)
$r_{12}p_{13}$	T(EB;0.0400)	...	T(MB;-0.0767)	...	T(MB;0.0833)
$r_{13}p_{13}$	T(EB;0.0300)	...	T(MB;-0.0767)	...	T(MB;-0.0667)
$r_{21}p_{13}$	T(EB;0.0200)	...	T(EB;0.0500)	...	T(MB;0.0333)
$r_{22}p_{13}$	T(EB;0.0450)	...	T(EB;0.0500)	...	T(B;0.0667)
$r_{31}p_{13}$	T(EB;0.0350)	...	T(EB;0.0800)	...	T(MB;-0.0667)
$r_{32}p_{13}$	T(EB;0.0450)	...	T(EB;0.0400)	...	T(M;-0.0500)
$r_{33}p_{13}$	T(EB;0.0100)	...	T(MB;-0.0767)	...	T(MB;0.0833)
...	...	...	...	...	...
$r_{12}p_{23}$	T(MB;-0.0467)	...	T(MB;-0.0667)	...	T(EB;0.0300)
$r_{24}p_{23}$	T(EB;0.0400)	...	T(EB;0.0600)	...	T(EB;0.0700)
$r_{31}p_{23}$	T(EB;0.0200)	...	T(MB;-0.0267)	...	T(EB;0.0800)
$r_{32}p_{23}$	T(EB;0.0800)	...	T(EB;0.0800)	...	T(EB;0.0200)
$r_{33}p_{23}$	T(MB;-0.0467)	...	T(MB;0.0133)	...	T(EB;0.0200)
...	...	...	...	...	...
$r_{12}p_{34}$	T(B;-0.0333)	...	T(EB;0.0700)	...	T(MB;-0.0667)
$r_{13}p_{34}$	T(MB;0.0733)	...	T(EB;0.0800)	...	T(MB;-0.0067)
$r_{22}p_{34}$	T(MB;0.0133)	...	T(MB;-0.0767)	...	T(MB;-0.0067)
$r_{23}p_{34}$	T(MB;-0.0467)	...	T(MB;-0.0767)	...	T(EB;0.0800)
$r_{24}p_{34}$	T(MB;0.0133)	...	T(EB;0.0700)	...	T(MB;0.0133)
$r_{31}p_{34}$	T(MB;0.0133)	...	T(EB;0.0700)	...	T(MB;-0.0667)
$r_{32}p_{34}$	T(MB;0.0133)	...	T(EB;0.0600)	...	T(MB;-0.0667)
$r_{33}p_{34}$	T(B;-0.0333)	...	T(MB;-0.0767)	...	T(MB;-0.0667)
...	...	...	...	...	...
$r_{11}p_{37}$	T(MB;0.0133)	...	T(MB;-0.0767)	...	T(MB;-0.0467)
$r_{12}p_{37}$	T(B;-0.0633)	...	T(EB;0.0500)	...	T(MB;-0.0267)
$r_{21}p_{37}$	T(MB;0.0433)	...	T(EB;0.0600)	...	T(MB;-0.0467)
$r_{32}p_{37}$	T(B;-0.0633)	...	T(EB;0.0800)	...	T(EB;0.0600)
$r_{33}p_{37}$	T(MB;-0.0467)	...	T(EB;0.0800)	...	T(MB;-0.0067)

TABLE XII. The Valuations Assigned to the Criteria to Calculate the Priority or Nodal Preference That Each Node Will Grant Each Requirement of Each Process According to the Node Load

Resources/ Processes	Criteria				
	%CPU	...	Process Priority	...	%VM
$r_{11}p_{11}$	T(M;0.000)	...	T(VH;-0.0333)	...	T(VH;0.0667)
$r_{12}p_{11}$	T(H;0.0333)	...	T(L;-0.0333)	...	T(M;0.0000)
$r_{21}p_{11}$	T(L;0.0667)	...	T(VH;0.0667)	...	T(VL;0.0333)
$r_{22}p_{11}$	T(M;0.0000)	...	T(VH;-0.0333)	...	T(L;0.0667)
$r_{23}p_{11}$	T(H;-0.0667)	...	T(EH;-0.05)	...	T(VH;-0.0333)
$r_{24}p_{11}$	T(M;0.0000)	...	T(H;-0.0667)	...	T(VL;0.0333)
...	...	...	...	...	...
$r_{11}p_{13}$	T(H;0.0333)	...	T(H;-0.0667)	...	T(VH;-0.0333)
$r_{12}p_{13}$	T(VH;-0.0333)	...	T(VH;0.0667)	...	T(L;0.0667)
$r_{13}p_{13}$	T(H;-0.0667)	...	T(VH;0.0667)	...	T(VH;-0.0333)
$r_{21}p_{13}$	T(L;0.0667)	...	T(M;0.0000)	...	T(L;-0.0333)
$r_{22}p_{13}$	T(VH;0.0667)	...	T(M;0.0000)	...	T(L;-0.0333)
$r_{31}p_{13}$	T(H;0.0333)	...	T(VH;-0.0333)	...	T(H;-0.0667)
$r_{32}p_{13}$	T(VH;0.0667)	...	T(L;0.0667)	...	T(H;-0.0667)
$r_{33}p_{13}$	T(VL;0.0333)	...	T(VH;0.0667)	...	T(H;-0.0667)
...	...	...	...	...	...
$r_{12}p_{23}$	T(H;-0.0667)	...	T(M;0.0000)	...	T(L;-0.0333)
$r_{24}p_{23}$	T(VL;0.0333)	...	T(L;-0.0333)	...	T(H;0.0333)
$r_{31}p_{23}$	T(VL;-0.0667)	...	T(H;0.0333)	...	T(VH;-0.0333)
$r_{32}p_{23}$	T(L;0.0667)	...	T(L;0.0667)	...	T(VL;0.0333)
$r_{33}p_{23}$	T(H;-0.0667)	...	T(VH;0.0667)	...	T(VL;0.0333)
...	...	...	...	...	...
$r_{12}p_{34}$	T(EH;0.0000)	...	T(H;0.0333)	...	T(M;0.0000)
$r_{13}p_{34}$	T(EH;0.0000)	...	T(VH;-0.0333)	...	T(M;0.0000)
$r_{22}p_{34}$	T(VH;-0.0333)	...	T(VH;0.0667)	...	T(VH;-0.0333)
$r_{23}p_{34}$	T(H;-0.0667)	...	T(VH;0.0667)	...	T(VH;-0.0333)
$r_{24}p_{34}$	T(L;0.0667)	...	T(H;0.0333)	...	T(L;0.0667)
$r_{31}p_{34}$	T(H;-0.0667)	...	T(H;0.0333)	...	T(VH;0.0667)
$r_{32}p_{34}$	T(H;-0.0667)	...	T(H;-0.0667)	...	T(M;0.0000)
$r_{33}p_{34}$	T(H;-0.0667)	...	T(VH;0.0667)	...	T(M;0.0000)
...	...	...	...	...	...
$r_{11}p_{37}$	T(H;-0.0667)	...	T(VH;0.0667)	...	T(H;-0.0667)
$r_{12}p_{37}$	T(VH;0.0667)	...	T(M;0.0000)	...	T(H;0.0333)
$r_{21}p_{37}$	T(H;0.0333)	...	T(H;-0.0667)	...	T(H;-0.0667)
$r_{32}p_{37}$	T(VH;0.0667)	...	T(VH;-0.0333)	...	T(L;-0.0333)
$r_{33}p_{37}$	T(L;0.0667)	...	T(VH;-0.0333)	...	T(VH;-0.0333)

In summary, the nodal priority (to be calculated at the node where the request occurs) of a process to access a given resource (which can be at any node) is calculated by the scalar product of the vectors mentioned above.

This nodal priority must be transformed into the 2-tuple format, considering the linguistic labels already mentioned. Therefore, it will be necessary to compare each nodal priority value with the average value of each label, the minimum difference of these comparisons will indicate the corresponding label.

B. Calculation of the Priorities or Preferences of the Processes to Access the Shared Resources Available (Calculated in the Centralized Resource Manager) and Determining the Order in Which the Resources Will Be Allocated, and Which Process Each Resource Will Be Assigned

Table XIII and Table XIV are used to calculate the final priorities, in which the priorities or nodal preferences calculated in the previous stage are placed; in this table each row contains information of the nodal priorities of the different processes to access a certain resource.

TABLE XIII. Nodal Priorities of the Processes to Access Each Resource in 2-Tuple (P_{11}, P_{13}, P_{23})

Resources	Nodal Process Priorities				
	p_{11}	...	p_{13}	...	p_{23}
r_{11}	NPT(H;0.0483)	...	NPT(H;0.0483)	...	-
r_{12}	NPT(M;-0.0050)	...	NPT(M;0.0350)	...	NPT(H;-0.070)
r_{13}	-	...	NPT(H;0.0733)	...	-
r_{21}	NPT(L;0.0217)	...	NPT(M;-0.060)	...	-
r_{22}	NPT(M;-0.0150)	...	NPT(M;-0.050)	...	-
r_{23}	NPT(VH;-0.0483)	...	-	...	-
r_{24}	NPT(L;0.0717)	...	-	...	NPT(L;-0.0008)
r_{31}	-	...	NPT(H;-0.0370)	...	NPT(M;-0.0400)
r_{33}	-	...	NPT(H;-0.070)	...	NPT(M;0.0200)
r_{33}	-	...	NPT(H;-0.0517)	...	NPT(H;-0.0317)

TABLE XIV. Nodal Priorities of the Processes to Access Each Resource in 2-Tuple (P_{34}, P_{37})

Resources	Nodal Process Priorities				
	...	p_{34}	...	p_{37}	...
r_{11}	...	-	...	NPT(H;0.0533)	...
r_{12}	...	NPT(H;0.0058)	...	NPT(VH;-0.0733)	...
r_{13}	...	NPT(H;0.0658)	...	-	...
r_{21}	...	-	...	NPT(H;0.0283)	...
r_{22}	...	NPT(VH;-0.0708)	...	-	...
r_{23}	...	NPT(H;0.0083)	...	-	...
r_{24}	...	NPT(M;0.0700)	...	-	...
r_{31}	...	NPT(H;0.0258)	...	-	...
r_{33}	...	NPT(H;0.0083)	...	NPT(H;-0.0192)	...
r_{33}	...	NPT(H;-0.0492)	...	NPT(H;-0.0541)	...

Next, the final weight vector to be used in the final aggregation process to determine the order or priority of access to resources must be calculated. In addition, the recently obtained weights will have to be normalized by dividing each one by the sum of all of them.

The nodal priorities indicated in Table XIII and Table XIV taken row by row, i.e., for each resource, will be multiplied by the final standardized weight vector mentioned above. See in Table XV and Table XVI.

TABLE XV. Final Global Priority Tuple (P_{11}, P_{13}, P_{23})

Resources	Nodal Process Priorities				
	p_{11}	...	p_{13}	...	p_{23}
r_{11}	NPT(EL;0.069)	...	NPT(EL;0.069)	...	-
r_{12}	NPT(EL;0.048)	...	NPT(EL;0.052)	...	NPT(EL;0.058)
r_{13}	-	...	NPT(EL;0.072)	...	-
r_{21}	NPT(EL;0.034)	...	NPT(EL;0.043)	...	-
r_{22}	NPT(EL;0.047)	...	NPT(EL;0.044)	...	-
r_{23}	NPT(EL;0.076)	...	-	...	-
r_{24}	NPT(EL;0.039)	...	-	...	NPT(EL;0.032)
r_{31}	-	...	NPT(EL;0.061)	...	NPT(EL;0.045)
r_{33}	-	...	NPT(EL;0.058)	...	NPT(EL;0.050)
r_{33}	-	...	NPT(EL;0.06)	...	NPT(EL;0.061)

The next step is to normalize Table XV and Table XVI between the extreme values. To do this, subtract the numerical value of the 2-tuple from the minimum value of both tables and divide it by the range, which is the difference between the maximum and minimum values of the two. As can see in Table XVII and Table XVIII.

The largest of these products made for the different processes in relation to the same resource will indicate which of the processes will have access to the resource.

TABLE XVI. Final Global Priority Tuple (P_{34}, P_{37})

Resources	...	Nodal Process Priorities p_{34}	...	p_{37}
r_{11}	...	-	...	NPT(EL;0.07)
r_{12}	...	NPT(EL;0.065)	...	NPT(EL;0.074)
r_{13}	...	NPT(EL;0.071)	...	-
r_{21}	...	-	...	NPT(EL;0.067)
r_{22}	...	NPT(EL;0.074)	...	-
r_{23}	...	NPT(EL;0.065)	...	-
r_{24}	...	NPT(EL;0.055)	...	-
r_{31}	...	NPT(EL;0.067)	...	-
r_{33}	...	NPT(EL;0.065)	...	NPT(EL;0.063)
r_{33}	...	NPT(EL;0.060)	...	NPT(EL;0.059)

TABLE XVII. Normalized Final Global Priority Tuple (P_{11}, P_{13}, P_{23})

Resources	Nodal Process Priorities p_{11}	...	p_{13}	...	p_{23}
r_{11}	NFGPT(VH;0.06)	...	NFGPT(VH;0.06)	...	-
r_{12}	NFGPT(M;0.060)	...	NFGPT(H;-0.046)	...	NFGPT(VH;-0.052)
r_{13}	-	...	**NFGPT(EH;-0.068)**	...	-
r_{21}	NFGPT(L;0.014)	...	NFGPT(M;-0.023)	...	-
r_{22}	NFGPT(M;0.045)	...	NFGPT(M;-0.008)	...	-
r_{23}	**NFGPT(EH;0.00)**	...	-	...	-
r_{24}	NFGPT(M;-0.076)	...	-	...	NFGPT(L;0.007)
r_{31}	-	...	NFGPT(VH;-0.068)	...	NFGPT(M;0.051)
r_{33}	-	...	NFGPT(H;0.053)	...	NFGPT(H;-0.017)
r_{33}	-	...	NFGPT(H;0.075)	...	**NFGPT(VH;-0.061)**

TABLE XVIII. Normalized Final Global Priority Tuple (P_{34}, P_{37})

Resources	...	Nodal Process Priorities p_{34}	...	p_{37}
r_{11}	...	-	...	**NFGPT(VH;0.068)**
r_{12}	...	NFGPT(VH;-0.004)	...	**NFGPT(EH;-0.038)**
r_{13}	...	NFGPT(EH;-0.080)	...	-
r_{21}	...	-	...	**NFGPT(VH;0.030)**
r_{22}	...	**NFGPT(EH;-0.034)**	...	-
r_{23}	...	NFGPT(VH;-0.00)	...	-
r_{24}	...	**NFGPT(H;0.007)**	...	-
r_{31}	...	**NFGPT(VH;0.026)**	...	-
r_{33}	...	**NFGPT(VH;-0.00)**	...	NFGPT(VH;-0.042)
r_{33}	...	NFGPT(H;0.079)	...	NFGPT(H;0.072)

The summation of all these products in relation to the same resource will indicate the priority that this resource will have to be assigned, in relation to the other resources that will also have to be assigned. This constitutes the Linguistics Distributed System Assignment Function (*LDSAF*). Refer to (3).

$$LDSAF(r_{ij}) = \Sigma\ NFGPT_{ijkl} = r_{ij} \text{ resource assignment priority} \quad (3)$$

By calculating the *LDSAF* for all resources, a 2-tuple vector will be obtained and, by ordering its elements from highest to lowest, the priority order of resource allocation will be obtained, which should be normalized ensuring that the 2-tuples obtained are in the interval [0, 1]. As can be seen in Table XIX.

TABLE XIX. Valuations to Normalize the *LDSAF* (First Iteration)

Label	Value
Maximum Value	5.2364
Minimum value	1.8255
Range	3.4109

In addition, as indicated above, the largest of the $NFGPT_{ijkl}$ for each resource will indicate the process to which the resource will be assigned.

The result of normalizing the 2 tuples constitutes what will be called Normalized Linguistics Distributed System Assignment Function a (*NLDSAF*). Refer to (4)

$$NLDSAF(r_{ij}) = \Sigma\ (NFGPT_{ijkl} / (\text{Maximum } (NFGPT_{ijkl}) - \text{Minimum } (NFGPT_{ijkl}))) = r_{ij} \text{ resource assignment priority normalized between extreme values} \quad (4)$$

This can be seen in Table XX.

TABLE XX. Normalized Linguistics Distributed System Assignment Function Ordered By Highest Priority Resource (First Iteration)

Assignment order of resources	Priority of the resource	Process at which assign the resource
r_{33}	T(EH;0.0000)	p_{23}
r_{12}	T(EH;-0.0326)	p_{37}
r_{31}	T(H;-0.0028)	p_{34}
r_{11}	T(H;-0.0355)	p_{37}
r_{22}	T(M;0.0423)	p_{34}
r_{21}	T(M;0.0337)	p_{37}
r_{13}	T(M;0.0274)	p_{13}
r_{32}	T(M;-0.0334)	p_{34}
r_{23}	T(L;-0.0591)	p_{11}
r_{24}	T(EL;0.0000)	p_{34}

The next step is to repeat the procedure but removing the requests of already made allocations; it must be noted that the assigned resources will be available once they are released by the processes and can therefore be allocated to other processes. Table XIX should be recalculated by omitting the resource allocations already done. As can see in Fig. 3.

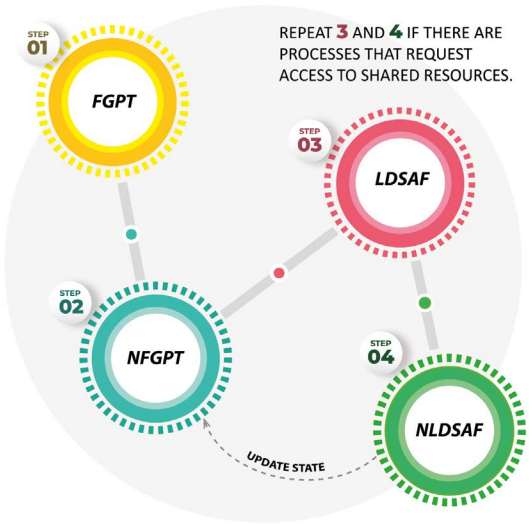


Fig. 3. Process of self-regulation and calculation of *NLDSAF*.

Since the system regulates itself by releasing the resources already assigned to the processes in the previous step, and because there are resource requests from the processes that have not yet been satisfied, the calculations in Table XIX and XX are repeated with their respective values, omitting the processes already completed.

Normalized the 2-tuples in the second iteration can be seen in Table XXI.

TABLE XXI. VALUATIONS TO NORMALIZE THE *LDSAF* (SECOND ITERATION)

Label	Value
Maximum Value	4.8503
Minimum value	1.2514
Range	3.5989

The result of normalizing the 2-tuples for the second iteration can be seen in Table XXII.

TABLE XXII. NORMALIZED LINGUISTICS DISTRIBUTED SYSTEM ASSIGNMENT FUNCTION ORDERED BY HIGHEST PRIORITY RESOURCE (SECOND ITERATION)

Assignment order of resources	Priority of the resource	Process at which assign the resource
r_{33}	T(EH;0.0000)	p_{34}
r_{12}	T(VH;0.0758)	p_{34}
r_{31}	T(H;-0.0392)	p_{13}
r_{11}	T(M;0.0813)	p_{11}
r_{21}	T(M;-0.0076)	p_{25}
r_{22}	T(M;-0.0298)	p_{11}
r_{13}	T(M;-0.0347)	p_{34}
r_{32}	T(M;-0.0676)	p_{37}
r_{23}	T(VL;0.0173)	p_{34}
r_{24}	T(EL;0.0000)	p_{11}

The final tables with the results of all the iterations will be shown below.

The valuations to normalize the *LDSAF* of each iteration can be seen in Table XXIII.

TABLE XXIII. VALUATIONS TO NORMALIZE THE *LDSAF* (OF EACH ITERATION).

Maximum Value	Minimum value	Range	Iteration
5.2364	1.8255	3.4109	1
4.8503	1.2514	3.5989	2
4.1598	0.8147	3.3451	3
4.0308	0.5616	3.4692	4
3.9955	0.4695	3.526	5
3.6488	0.2681	3.3807	6
2.6775	0.0000	2.6775	7
2.4275	0.0000	2.4275	8
2.9932	0.0000	2.9932	9
1.5946	0.0000	1.5946	10
1.0000	0.0000	1.0000	11
0.0169	0.0000	0.0169	12

The result of normalizing the 2 tuples for each iteration can be seen in Table XXIV, ordered by highest priority resource (of each iteration).

V. EXAMPLE OF A DECISION MODEL APPLIED TO ONE OF THE TRADITIONAL ALGORITHMS

A particular case of the proposed decision model is to visualize how some of the methods considered traditional in this work, are a particular case of this method.

As the traditional methods do not consider groups of processes, the calculation will only be done with independent processes and the column that considers whether a process is part of a group of processes should be disabled, see Table XXV.

TABLE XXV. WEIGHTS ASSIGNED TO THE PROCESSES FOR THE CALCULATION OF PRIORITIES

Processes	Group of processes	Independent processes
p_{11}	-	$wf_{11}=1/np$
....	-	....
p_{kl}	-	$wf_{kl}=1/np$
....	-	
p_{np}	-	$wf_{np}=1/np$

The methods considered traditional do not consider most of the criteria contemplated in the proposed model (in addition to not considering the representation by means of linguistic labels or 2-tuples), they are only based on the calculation of the "Process Priority" criterion. The weights assigned to the criteria in Table XXVI, to calculate the global priority, only the "Process Priority" criterion will be considered, disabling the other criteria.

For each requirement of a resource made by a process, the assessment vectors are applied according to the criteria established for the determination of the priority. This is done in the node where the requirement occurs. To obtain the node priority, each rating vector of each requirement must be scaled and multiplied by the weight vector corresponding to the current load category of the node.

TABLE XXVI. WEIGHTS ASSIGNED TO THE CRITERIA FOR CALCULATING PRIORITY

Categories	Process priority	Other Criteria
High	0.1000	-
Medium	0.2000	-
Low	0.1000	-

Although the decision model obtains the information of all the criteria, it should be noted that for the traditional methods, from the weight vector Table XXVI, only the criterion "Process Priority" will affect the calculation of the priority.

The valuations assigned to the criteria for calculating the priority or preference that each node will give to each requirement of each process according to the node load, will be those used in TABLE XII. To calculate the priorities or preferences of the processes, taking into account the state of the node, Table XX will be used, but disabling all the criteria, except "Process Priority". Nodal priorities, final weights and overall process priorities for accessing resources must be calculated.

Table XXVII, Table XXVIII, Table XXIX, Table XXX and Table XXXI are constructed from the nodal priority values, which for this example matches the "Process Priority" criteria.

TABLE XXIV. Concatenated Normalized Linguistics Distributed System Assignment Function (*CNLDSAF*)

Assignment order of resources	Priority of the resource	Process at which assign the resource	Iteration
r_{33}	T(EH;0.0000)	p_{23}	1
r_{12}	T(EH;-0.0326)	p_{37}	1
r_{31}	T(H;-0.0028)	p_{34}	1
r_{11}	T(H;-0.0355)	p_{37}	1
r_{22}	T(M;0.0423)	p_{34}	1
r_{21}	T(M;0.0337)	p_{37}	1
r_{13}	T(M;0.0274)	p_{13}	1
r_{32}	T(M;-0.0334)	p_{34}	1
r_{23}	T(L;-0.0591)	p_{11}	1
r_{24}	T(EL;0.0000)	p_{34}	1
r_{33}	T(EH;0.0000)	p_{34}	2
r_{12}	T(VH;0.0758)	p_{34}	2
r_{31}	T(H;-0.0392)	p_{13}	2
r_{11}	T(M;0.0813)	p_{11}	2
r_{21}	T(M;-0.0076)	p_{25}	2
r_{22}	T(M;-0.0298)	p_{11}	2
r_{13}	T(M;-0.0347)	p_{34}	2
r_{32}	T(M;-0.0676)	p_{37}	2
r_{23}	T(VL;0.0173)	p_{34}	2
r_{24}	T(EL;0.0000)	p_{11}	2
r_{33}	T(EH;0.0000)	p_{34}	3
r_{12}	T(VH;0.0758)	p_{34}	3
r_{31}	T(H;-0.0392)	p_{13}	3
r_{11}	T(M;0.0813)	p_{11}	3
r_{21}	T(M;-0.0076)	p_{25}	3
r_{22}	T(M;-0.0298)	p_{11}	3
r_{13}	T(M;-0.0347)	p_{34}	3
r_{32}	T(M;-0.0676)	p_{37}	3
r_{23}	T(VL;0.0173)	p_{34}	3
r_{24}	T(EL;0.0000)	p_{11}	3
r_{33}	T(EH;0.0000)	p_{37}	4
r_{12}	T(VH;0.0255)	p_{13}	4
r_{31}	T(H;-0.0739)	p_{23}	4
r_{22}	T(M;-0.0178)	p_{13}	4
r_{21}	T(M;-0.0721)	p_{12}	4
r_{11}	T(L;0.0082)	p_{12}	4
r_{13}	T(L;-0.0337)	p_{21}	4
r_{32}	T(L;-0.0762)	p_{23}	4
r_{23}	T(EL;0.0489)	p_{32}	4
r_{24}	T(EL;0.0000)	p_{35}	4
r_{33}	T(EH;0.0000)	p_{12}	5
r_{12}	T(VH;0.0431)	p_{11}	5
r_{31}	T(H;-0.0777)	p_{31}	5
r_{22}	T(M;-0.0467)	p_{12}	5
r_{21}	T(L;0.0674)	p_{22}	5
r_{13}	T(L;-0.0824)	p_{32}	5
r_{11}	T(VL;0.0273)	p_{32}	5
r_{32}	T(VL;-0.0591)	p_{36}	5
r_{23}	T(EL;0.0230)	p_{33}	5
r_{24}	T(EL;0.0000)	p_{36}	5
r_{33}	T(EH;0.0000)	p_{31}	6
r_{12}	T(VH;-0.0133)	p_{12}	6
r_{31}	T(M;0.0638)	p_{12}	6
r_{22}	T(L;0.0275)	p_{21}	6
r_{21}	T(L;-0.0106)	p_{11}	6
r_{13}	T(VL;0.0815)	p_{36}	6
r_{11}	T(VL;0.0443)	p_{36}	6
r_{32}	T(VL;-0.0604)	p_{35}	6
r_{23}	T(EL;0.0032)	p_{24}	6
r_{24}	T(EL;0.0000)	p_{24}	6
r_{33}	T(EH;0.0000)	p_{21}	7
r_{12}	T(VH;-0.0348)	p_{21}	7
r_{31}	T(M;0.0040)	p_{22}	7
r_{13}	T(VL;0.0742)	p_{35}	7
r_{21}	T(VL;0.0682)	p_{33}	7
r_{11}	T(VL;0.0631)	p_{33}	7
r_{22}	T(VL;0.0580)	p_{35}	7
r_{32}	T(VL;-0.0553)	p_{33}	7
r_{33}	T(EH;0.0000)	p_{22}	8
r_{12}	T(EH;-0.0210)	p_{33}	8
r_{31}	T(M;-0.0299)	p_{36}	8
r_{21}	T(VL;0.0191)	p_{36}	8
r_{13}	T(VL;0.0110)	p_{33}	8
r_{22}	T(VL;0.0078)	p_{33}	8
r_{11}	T(VL;-0.0035)	p_{24}	8
r_{12}	T(EH;0.0000)	p_{36}	9
r_{33}	T(VH;-0.0117)	p_{33}	9
r_{31}	T(L;-0.0534)	p_{35}	9
r_{22}	T(EL;0.0000)	p_{36}	9
r_{12}	T(EH;0.0000)	p_{24}	10
r_{33}	T(VH;0.0734)	p_{35}	10
r_{12}	T(EH;0.0000)	p_{32}	11
r_{33}	T(H;-0.0196)	p_{36}	11
r_{12}	T(EH;0.0000)	p_{35}	12

TABLE XXVII. NODAL PRIORITIES OF THE PROCESSES TO ACCESS EACH RESOURCE IN 2-TUPLE (p_{11}, p_{12}, p_{13})

	p_{11}	p_{12}	p_{13}
r_{11}	NPT(EL;0.0800)	NPT(VL;-0.0667)	NPT(EL;0.0600)
r_{12}	NPT(EL;0.0300)	NPT(EL;0.0800)	NPT(VL;-0.0767)
r_{13}	-	-	NPT(VL;-0.0767)
r_{21}	NPT(VL;-0.0767)	NPT(EL;0.0800)	NPT(EL;0.0500)
r_{22}	NPT(EL;0.0800)	NPT(EL;0.0800)	NPT(EL;0.0500)
r_{23}	NPT(VL;-0.0717)	-	-
r_{24}	NPT(EL;0.0600)	-	-
r_{31}	-	NPT(EL;0.0300)	NPT(EL;0.0800)
r_{32}	-	-	NPT(EL;0.0400)
r_{33}	-	NPT(EL;0.0300)	NPT(VL;-0.0767)

TABLE XXVIII. NODAL PRIORITIES OF THE PROCESSES TO ACCESS EACH RESOURCE IN 2-TUPLE (p_{21}, p_{22}, p_{23})

	p_{21}	p_{22}	p_{23}
r_{11}	-	-	-
r_{12}	NPT(VL;-0.0267)	-	NPT(VL;-0.0667)
r_{13}	NPT(VL;-0.0267)	-	-
r_{21}	-	NPT(VL;-0.0467)	-
r_{22}	NPT(VL;-0.0667)	NPT(VL;0.0133)	-
r_{23}	NPT(VL;-0.0667)	-	-
r_{24}	-	-	NPT(EL;0.0600)
r_{31}	NPT(VL;0.0133)	NPT(EL;0.0800)	NPT(VL;-0.0267)
r_{32}	-	-	NPT(EL;0.0800)
r_{33}	NPT(EL;0.0800)	NPT(VL;-0.0067)	NPT(VL;0.0133)

TABLE XXIX. NODAL PRIORITIES OF THE PROCESSES TO ACCESS EACH RESOURCE IN 2-TUPLE (p_{24}, p_{25}, p_{31})

	p_{24}	p_{25}	p_{31}
r_{11}	NPT(VL;-0.0667)	-	-
r_{12}	NPT(VL;0.0133)	-	-
r_{13}	-	-	NPT(EL;0.0700)
r_{21}	-	NPT(EL;0.0800)	-
r_{22}	-	-	-
r_{23}	NPT(VL;-0.0067)	-	-
r_{24}	NPT(EL;0.0600)	-	-
r_{31}	-	-	NPT(EL;0.0700)
r_{32}	-	-	-
r_{33}	-	-	NPT(VL;-0.0767)

TABLE XXX. NODAL PRIORITIES OF THE PROCESSES TO ACCESS EACH RESOURCE IN 2-TUPLE (p_{32}, p_{33}, p_{34})

	p_{32}	p_{33}	p_{34}
r_{11}	NPT(VL;-0.0767)	NPT(EL;0.0600)	-
r_{12}	NPT(EL;0.0800)	NPT(EL;0.0300)	NPT(EL;0.0700)
r_{13}	NPT(VL;-0.0767)	NPT(EL;0.0300)	NPT(EL;0.0800)
r_{21}	-	NPT(EL;0.0800)	-
r_{22}	-	NPT(EL;0.0700)	NPT(VL;-0.0767)
r_{23}	NPT(EL;0.0600)	NPT(EL;0.0600)	NPT(VL;-0.0767)
r_{24}	-	-	NPT(EL;0.0700)
r_{31}	-	-	NPT(EL;0.0700)
r_{32}	-	NPT(EL;0.0400)	NPT(EL;0.0600)
r_{33}	-	NPT(EL;0.0600)	NPT(VL;-0.0767)

TABLE XXXI. NODAL PRIORITIES OF THE PROCESSES TO ACCESS EACH RESOURCE IN 2-TUPLE (p_{35}, p_{36}, p_{37})

	p_{35}	p_{36}	p_{37}
r_{11}	-	NPT(EL;0.0800)	NPT(VL;-0.0767)
r_{12}	NPT(VL;-0.0767)	NPT(EL;0.0800)	NPT(EL;0.0500)
r_{13}	NPT(VL;-0.0767)	NPT(EL;0.0800)	-
r_{21}	-	NPT(EL;0.0700)	NPT(EL;0.0600)
r_{22}	NPT(EL;0.0800)	NPT(EL;0.0800)	-
r_{23}	-	-	-
r_{24}	NPT(EL;0.0600)	NPT(EL;0.0800)	-
r_{31}	NPT(EL;0.0800)	NPT(EL;0.0800)	-
r_{32}	NPT(EL;0.0400)	NPT(EL;0.0800)	NPT(EL;0.0800)
r_{33}	NPT(EL;0.0800)	NPT(EL;0.0800)	NPT(EL;0.0800)

As mentioned, traditional methods do not consider groups of processes. The calculation will only consider that the processes are independent. In the example there are 15 processes and the calculation of the weights is wf_{ij} equal to $1/np$ for independent processes, where np is the number of processes in the system (15), the calculation for the weights of each process (wp_{ij}) is equal to 1/15. For the calculation of the standardized weights (nwp_{ij}) each wp_{ij} value is divided by the sum of all wp_{ij}, this can be seen in the tables above. The final weight vector to be used in the final aggregation process should be calculated to determine the order or priority of access to resources. In addition, the recently obtained weights should be normalized by dividing each one of them by the sum of all of them. For this particular situation, all processes will have the same weight value, since they are only considered as independent processes. The normalized weight vector will have the same value for all processes, this value will be 1/15, which results in 0.0066.

The nodal priorities indicated in Table XXVII, Table XXVIII, Table XXIX, Table XXX and Table XXXI taken row by row, that is, for each resource, will be multiplied by the normalized weight vector (nwp_{ij}). This can be seen in Table XXXII, Table XXXIII, XXXIV, Table XXXV and Table XXXVI.

TABLE XXXII. NODAL PRIORITIES OF THE PROCESSES TO ACCESS EACH RESOURCE IN 2-TUPLE (p_{11}, p_{12}, p_{13})

	p_{11}	p_{12}	p_{13}
r_{11}	NPT(EL;0.005)	NPT(EL;0.007)	NPT(EL;0.004)
r_{12}	NPT(EL;0.002)	NPT(EL;0.005)	NPT(EL;0.006)
r_{13}	-	-	NPT(EL;0.006)
r_{21}	NPT(EL;0.006)	NPT(EL;0.005)	NPT(EL;0.003)
r_{22}	NPT(EL;0.005)	NPT(EL;0.005)	NPT(EL;0.003)
r_{23}	NPT(EL;0.006)	-	-
r_{24}	NPT(EL;0.004)	-	-
r_{31}	-	NPT(EL;0.002)	NPT(EL;0.005)
r_{32}	-	-	NPT(EL;0.003)
r_{33}	-	NPT(EL;0.002)	NPT(EL;0.006)

TABLE XXXIII. NODAL PRIORITIES OF THE PROCESSES TO ACCESS EACH RESOURCE IN 2-TUPLE (p_{21}, p_{22}, p_{23})

	p_{21}	p_{22}	p_{23}
r_{11}	-	-	-
r_{12}	NPT(EL;0.009)	-	NPT(EL;0.007)
r_{13}	NPT(EL;0.009)	-	-
r_{21}	-	NPT(EL;0.008)	-
r_{22}	NPT(EL;0.007)	NPT(EL;0.012)	-
r_{23}	NPT(EL;0.007)	-	-
r_{24}	-	-	NPT(EL;0.004)
r_{31}	NPT(EL;0.012)	NPT(EL;0.005)	NPT(EL;0.009)
r_{32}	-	-	NPT(EL;0.005)
r_{33}	NPT(EL;0.005)	NPT(EL;0.011)	NPT(EL;0.012)

TABLE XXXIV. Nodal Priorities of the Processes to Access Each Resource in 2-Tuple (p_{24}, p_{25}, p_{31})

	p_{24}	p_{25}	p_{31}
r_{11}	NPT(EL;0.007)	-	-
r_{12}	NPT(EL;0.012)	-	-
r_{13}	-	-	NPT(EL;0.005)
r_{21}	-	NPT(EL;0.005)	-
r_{22}	-	-	-
r_{23}	NPT(EL;0.011)	-	-
r_{24}	NPT(EL;0.004)	-	-
r_{31}	-	-	NPT(EL;0.005)
r_{32}	-	-	-
r_{33}	-	-	NPT(EL;0.006)

TABLE XXXV. Nodal Priorities of the Processes to Access Each Resource in 2-Tuple (p_{32}, p_{33}, p_{34})

	p_{32}	p_{33}	p_{34}
r_{11}	NPT(EL;0.006)	NPT(EL;0.004)	-
r_{12}	NPT(EL;0.005)	NPT(EL;0.002)	NPT(EL;0.005)
r_{13}	NPT(EL;0.006)	NPT(EL;0.002)	NPT(EL;0.005)
r_{21}	-	NPT(EL;0.005)	-
r_{22}	-	NPT(EL;0.005)	NPT(EL;0.006)
r_{23}	NPT(EL;0.004)	NPT(EL;0.004)	NPT(EL;0.006)
r_{24}	-	-	NPT(EL;0.005)
r_{31}	-	-	NPT(EL;0.005)
r_{32}	-	NPT(EL;0.003)	NPT(EL;0.004)
r_{33}	-	NPT(EL;0.004)	NPT(EL;0.006)

TABLE XXXVI. Nodal Priorities of the Processes to Access Each Resource in 2-Tuple (p_{35}, p_{36}, p_{37})

	p_{35}	p_{36}	p_{37}
r_{11}	-	NPT(EL;0.005)	NPT(EL;0.006)
r_{12}	NPT(EL;0.006)	NPT(EL;0.005)	NPT(EL;0.003)
r_{13}	NPT(EL;0.006)	NPT(EL;0.005)	-
r_{21}	-	NPT(EL;0.005)	NPT(EL;0.004)
r_{22}	NPT(EL;0.005)	NPT(EL;0.005)	-
r_{23}	-	-	-
r_{24}	NPT(EL;0.004)	NPT(EL;0.005)	-
r_{31}	NPT(EL;0.005)	NPT(EL;0.005)	-
r_{32}	NPT(EL;0.003)	NPT(EL;0.005)	NPT(EL;0.005)
r_{33}	NPT(EL;0.005)	NPT(EL;0.005)	NPT(EL;0.005)

The next step is to normalize Table XXXII, Table XXXIII, Table XXXIV, Table XXXV and Table XXXVI between the extreme values. To do this, you must subtract the numerical value of the 2-tuple by the minimum value of all of them and divide it by the range. The range is the difference between the maximum value and minimum value of the tables already mentioned. This can be seen in Table XXXVII.

TABLE XXXVII. Normalization Assessments

Label	Value
Maximum Value	0.012
Minimum value	0.002
Range	0.010

The result of this standardization can be seen in Table XXXVIII, Table XXXIX, Table XL, Table XLI and Table XLII. The largest of these products made for the different processes in relation to the same resource will indicate which of the processes will have access to the resource.

TABLE XXXVIII. Nodal Priorities of the Processes to Access Each Resource in 2-Tuple Normalized (p_{11}, p_{12}, p_{13})

	p_{11}	p_{12}	p_{13}
r_{11}	NFGPT(L;0.000)	NFGPT(M;-0.033)	NFGPT(VL;0.033)
r_{12}	NFGPT(EL;0.000)	NFGPT(L;0.000)	NFGPT(L;0.067)
r_{13}	-	-	NFGPT(L;0.067)
r_{21}	NFGPT(L;0.067)	NFGPT(L;0.000)	NFGPT(VL;-0.033)
r_{22}	NFGPT(L;0.000)	NFGPT(L;0.000)	NFGPT(VL;-0.033)
r_{23}	NFGPT(M;-0.067)	-	-
r_{24}	NFGPT(VL;0.033)	-	-
r_{31}	-	NFGPT(EL;0.000)	NFGPT(L;0.000)
r_{32}	-	-	NFGPT(EL;0.067)
r_{33}	-	NFGPT(EL;0.000)	NFGPT(L;0.067)

TABLE XXXIX. Nodal Priorities of the Processes to Access Each Resource in 2-Tuple Normalized (p_{21}, p_{22}, p_{23})

	p_{21}	p_{22}	p_{23}
r_{11}	-	-	-
r_{12}	NFGPT(H;0.067)	-	NFGPT(M;-0.033)
r_{13}	NFGPT(H;0.067)	-	-
r_{21}	-	NFGPT(H;-0.067)	-
r_{22}	NFGPT(M;-0.033)	NFGPT(EH;0.000)	-
r_{23}	NFGPT(M;-0.033)	-	-
r_{24}	-	-	NFGPT(VL;0.033)
r_{31}	NFGPT(EH;0.000)	NFGPT(L;0.000)	NFGPT(H;0.067)
r_{32}	-	-	NFGPT(L;0.000)
r_{33}	NFGPT(L;0.000)	NFGPT(VH;0.033)	NFGPT(EH;0.000)

TABLE XL. Nodal Priorities of the Processes to Access Each Resource in 2-Tuple Normalized (p_{24}, p_{25}, p_{31})

	p_{24}	p_{25}	p_{31}
r_{11}	NFGPT(M;-0.033)	-	-
r_{12}	NFGPT(EH;0.000)	-	-
r_{13}	-	-	NFGPT(L;-0.067)
r_{21}	-	NFGPT(L;0.000)	-
r_{22}	-	-	-
r_{23}	NFGPT(VH;0.033)	-	-
r_{24}	NFGPT(VL;0.033)	-	-
r_{31}	-	-	NFGPT(L;-0.067)
r_{32}	-	-	-
r_{33}	-	-	NFGPT(L;0.067)

TABLE XLI. Nodal Priorities of the Processes to Access Each Resource in 2-Tuple Normalized (p_{32}, p_{33}, p_{34})

	p_{32}	p_{33}	p_{34}
r_{11}	NFGPT(L;0.067)	NFGPT(VL;0.033)	-
r_{12}	NFGPT(L;0.000)	NFGPT(EL;0.000)	NFGPT(L;-0.067)
r_{13}	NFGPT(L;0.067)	NFGPT(EL;0.000)	NFGPT(L;0.000)
r_{21}	-	NFGPT(L;0.000)	-
r_{22}	-	NFGPT(L;-0.067)	NFGPT(L;0.067)
r_{23}	NFGPT(VL;0.033)	NFGPT(VL;0.033)	NFGPT(L;0.067)
r_{24}	-	-	NFGPT(L;-0.067)
r_{31}	-	-	NFGPT(L;-0.067)
r_{32}	-	NFGPT(EL;0.067)	NFGPT(VL;0.033)
r_{33}	-	NFGPT(VL;0.033)	NFGPT(L;0.067)

TABLE XLII. NODAL PRIORITIES OF THE PROCESSES TO ACCESS EACH RESOURCE IN 2-TUPLE NORMALIZED (p_{35}, p_{36}, p_{37})

	p_{35}	p_{36}	p_{37}
r_{11}	-	NFGPT(L;0.000)	NFGPT(L;0.067)
r_{12}	NFGPT(L;0.067)	NFGPT(L;0.000)	NFGPT(VL;-0.033)
r_{13}	NFGPT(L;0.067)	NFGPT(L;0.000)	-
r_{21}	-	NFGPT(L;-0.067)	NFGPT(VL;0.033)
r_{22}	NFGPT(L;0.0000)	NFGPT(L;0.000)	-
r_{23}	-	-	-
r_{24}	NFGPT(VL;0.033)	NFGPT(L;0.000)	-
r_{31}	NFGPT(L;0.000)	NFGPT(L;0.000)	-
r_{32}	NFGPT(EL;0.067)	NFGPT(L;0.000)	NFGPT(L;0.000)
r_{33}	NFGPT(L;0.000)	NFGPT(L;0.000)	NFGPT(L;0.000)

The sum of all these products in relation to the same resource will indicate the priority that should be assigned to this resource, in relation to the other resources that should also be assigned. This constitutes the Linguistic Distributed System Assignment Function (*LDSAF*). Refer to (5).

$$LDSAF(r_{ij}) = \Sigma\ NFGPT_{ijkl}\ with\ r_{ij}\ \text{resource allocation priority} \quad (5)$$

When calculating the *LDSAF* for all resources, a 2-tuple vector will be obtained. Sorting their elements from highest to lowest, you will get the priority order of resource allocation. This should be normalized by ensuring that the 2-tuples obtained are in the range [0, 1]. This can be seen in Table XLIII.

In addition, as indicated above, the largest of the $NFGPT_{ijkl}$ of each resource will indicate the process to which the resource will be assigned.

TABLE XLIII. ASSESSMENTS TO NORMALIZE THE *LDSAF*

Label	Value
Maximum Value	4.5999
Minimum value	1.4003
Range	3.1996

The result of the standardization of the 2-tuples constitutes what will be called the Normalized Linguistic Distributed System Assignment Function (*LDSAF*). Refer to (6).

$$LDSAF(r_{ij}) = \Sigma\ (NFGPT_{ijkl} / (\text{Maximum}\ (NFGPT_{ijkl}) - \text{Minimum}\ (NFGPT_{ijkl})) = r_{ij}\ \text{priority of resource allocation normalized between extreme values} \quad (6)$$

This is shown in Table XLIV.

TABLE XLIV. ASSESSMENTS TO NORMALIZE THE *LDSAF*

Order of resource assignment	Priority of assignment	Process selected
r_{33}	T(EH;0.0000)	p_{23}
r_{12}	T(EH;-0.0625)	p_{24}
r_{31}	T(H;0.0209)	p_{21}
r_{22}	T(H;0.0209)	p_{22}
r_{13}	T(M;-0.0418)	p_{21}
r_{11}	T(M;-0.0626)	p_{12}
r_{21}	T(L;0.0417)	p_{22}
r_{23}	T(L;0.0311)	p_{24}
r_{24}	T(EL;0.0000)	p_{36}
r_{32}	T(EL;0.0000)	p_{23}

The next step is to repeat the procedure but eliminating the requests for assignments already made. It should be noted that the allocated resources will be available once the processes release them and therefore, they can be allocated to other processes. The system is self-regulating by releasing the resources already assigned to the processes in the previous step.

The resource requests from the processes that have not yet been satisfied, that is, the calculations in Table XLIII and Table XLIV are repeated with their respective values, omitting the processes already completed.

The result of the concatenation of all allocation rounds for this example can be seen in Table XLV. The diagram in Fig. 4 shows a graph that allows the flow and relationship between the different rounds of resource allocation to processes to be represented.

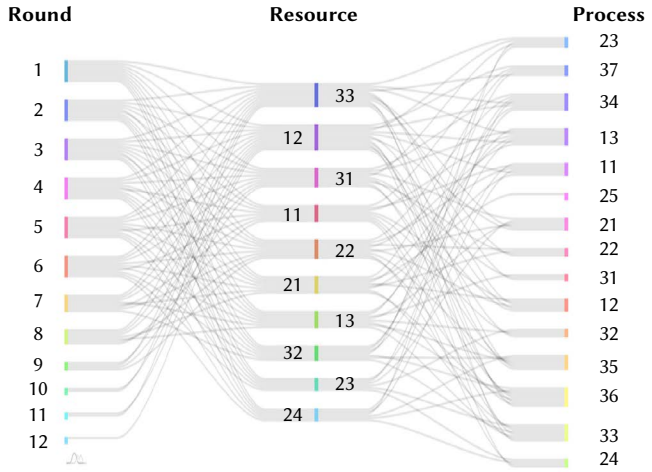


Fig. 4 Flow and relationships between different rounds of resource allocation.

VI. COMPARISON OF RESULTS OBTAINED WITH TRADITIONAL METHODS

Table XLV of this scenario is compared to Table XXIV of scenario E1 and look for each first round allocation of the latter in Table XLV.

In Table XLVI you can see the order of assignments for the first round of the E1 scenario, while Table XLVII represents the order in which the same assignments appear but for the traditional method, in which round they appear and in which position with respect to each round.

TABLE XLVI. VALUES CORRESPONDING TO THE FIRST ROUND OF ITERATION OF SCENARIO E1 (*ONLDSAF*)

Pos.	Resource	2-tuple	Process	Round
1	r_{33}	T(EH;0.0000)	p_{23}	1
2	r_{12}	T(EH;-0.0326)	p_{37}	1
3	r_{31}	T(H;-0.0028)	p_{34}	1
4	r_{11}	T(H;-0.0355)	p_{37}	1
5	r_{22}	T(M;0.0423)	p_{34}	1
6	r_{21}	T(M;0.0337)	p_{37}	1
7	r_{13}	T(M;0.0274)	p_{13}	1
8	r_{32}	T(M;-0.0334)	p_{34}	1
9	r_{23}	T(L;-0.0591)	p_{11}	1
10	r_{24}	T(EL;0.0000)	p_{34}	1

TABLE XLV. CONCATENATION OF ALL ASSIGNMENT ROUNDS (*CNLDSAF*) FOR TRADITIONAL METHODS

Resource	2-tuple	Process	Round
r_{33}	T(EH;0.0000)	p_{23}	1
r_{12}	T(EH;-0.0625)	p_{24}	1
r_{31}	T(H;0.0209)	p_{21}	1
r_{22}	T(H;0.0209)	p_{22}	1
r_{13}	T(M;-0.0418)	p_{21}	1
r_{11}	T(M;-0.0626)	p_{12}	1
r_{21}	T(L;0.0417)	p_{22}	1
r_{23}	T(L;0.0311)	p_{24}	1
r_{24}	T(EL;0.0000)	p_{36}	1
r_{32}	T(EL;0.0000)	p_{23}	1
r_{33}	T(EH;0.0000)	p_{22}	2
r_{12}	T(EH;-0.079)	p_{21}	2
r_{31}	T(H;-0.0613)	p_{23}	2
r_{22}	T(H;-0.0614)	p_{21}	2
r_{11}	T(M;-0.0001)	p_{24}	2
r_{13}	T(M;-0.0791)	p_{13}	2
r_{21}	T(L;0.0351)	p_{11}	2
r_{23}	T(VL;0.0832)	p_{21}	2
r_{24}	T(EL;0.0000)	p_{34}	2
r_{32}	T(EL;0.0000)	p_{36}	2
r_{33}	T(EH;0.0000)	p_{13}	3
r_{12}	T(EH;-0.0333)	p_{23}	3
r_{22}	T(H;0.0335)	p_{34}	3
r_{31}	T(M;0.0669)	p_{13}	3
r_{11}	T(M;0.0666)	p_{32}	3
r_{13}	T(M;-0.0001)	p_{32}	3
r_{21}	T(M;-0.0665)	p_{12}	3
r_{23}	T(VL;0.0832)	p_{11}	3
r_{24}	T(EL;0.0333)	p_{11}	3
r_{32}	T(EL;0.0000)	p_{37}	3
r_{33}	T(EH;0.0000)	p_{31}	4
r_{12}	T(EH;-0.069)	p_{13}	4
r_{22}	T(H;0.0231)	p_{11}	4
r_{31}	T(H;-0.0803)	p_{22}	4
r_{11}	T(M;0.0517)	p_{37}	4
r_{13}	T(M;-0.0172)	p_{35}	4
r_{21}	T(M;-0.0516)	p_{25}	4
r_{23}	T(VL;0.0402)	p_{34}	4
r_{24}	T(VL;-0.0632)	p_{23}	4
r_{32}	T(EL;0.0000)	p_{34}	4
r_{33}	T(EH;0.0000)	p_{34}	5
r_{12}	T(EH;-0.0769)	p_{35}	5
r_{22}	T(H;0.0257)	p_{12}	5
r_{31}	T(M;0.077)	p_{35}	5
r_{11}	T(M;0.0001)	p_{11}	5
r_{13}	T(M;-0.0769)	p_{34}	5
r_{21}	T(M;-0.0769)	p_{33}	5
r_{23}	T(VL;-0.0513)	p_{32}	5
r_{24}	T(VL;-0.0513)	p_{24}	5
r_{32}	T(EL;0.0000)	p_{13}	5
r_{33}	T(EH;0.0000)	p_{21}	6
r_{12}	T(VH;0.0714)	p_{12}	6
r_{22}	T(H;0.0000)	p_{35}	6
r_{31}	T(M;0.0238)	p_{36}	6
r_{11}	T(M;-0.0714)	p_{36}	6
r_{13}	T(L;0.0000)	p_{36}	6
r_{21}	T(L;0.0000)	p_{36}	6
r_{23}	T(EL;0.0476)	p_{33}	6
r_{24}	T(EL;0.0476)	p_{35}	6
r_{32}	T(EL;0.0000)	p_{33}	6
r_{33}	T(EH;0.0000)	p_{35}	7
r_{12}	T(VH;0.0556)	p_{32}	7
r_{22}	T(H;-0.0556)	p_{36}	7
r_{31}	T(M;-0.0556)	p_{31}	7
r_{11}	T(L;0.0000)	p_{13}	7
r_{21}	T(L;-0.0556)	p_{37}	7
r_{13}	T(VL;0.0556)	p_{31}	7
r_{32}	T(EL;0.0556)	p_{35}	7
r_{33}	T(EH;0.0000)	p_{36}	8
r_{12}	T(VH;0.0128)	p_{36}	8
r_{22}	T(M;-0.0385)	p_{33}	8
r_{31}	T(L;-0.0256)	p_{34}	8
r_{11}	T(VL;0.0641)	p_{33}	8
r_{21}	T(VL;-0.0128)	p_{13}	8
r_{33}	T(EH;0.0000)	p_{37}	9
r_{12}	T(VH;-0.0833)	p_{34}	9
r_{22}	T(L;-0.0833)	p_{13}	9
r_{33}	T(EH;0.0000)	p_{33}	10
r_{12}	T(H;0.0000)	p_{37}	10
r_{12}	T(EH;0.0000)	p_{11}	11
r_{13}	T(M;0.0000)	p_{33}	11
r_{31}	T(M;0.0000)	p_{12}	11
r_{33}	T(M;0.0000)	p_{12}	11
r_{12}	T(EH;0.0000)	p_{33}	12

TABLE XLVII. VALUES CORRESPONDING TO THE SAME PROCESS RESOURCE ASSIGNMENTS FROM THE *ONLDSAF* TABLE OF THE FIRST E1 ITERATION FOUND IN THE *CNLDSAF* TABLE OF THE TRADITIONAL METHODS

Pos.	Resource	2-tuple	Process	Round
1	r_{33}	T(EH;0.0000)	p_{23}	1
2	r_{12}	T(H;0.0000)	p_{37}	10
4	r_{31}	T(L;-0.0256)	p_{34}	8
5	r_{11}	T(M;0.0517)	p_{37}	4
3	r_{22}	T(H;0.0335)	p_{34}	3
6	r_{21}	T(L;-0.0556)	p_{37}	7
6	r_{13}	T(M;-0.0791)	p_{13}	2
10	r_{32}	T(EL;0.0000)	p_{34}	4
8	r_{23}	T(VL;0.0832)	p_{11}	3
9	r_{24}	T(EL;0.0000)	p_{34}	2

The first element of Table XLVII, assignment of r_{33} to p_{23}, is the only one that occurs in the same iteration (first), all other assignments in the example of traditional methods occur in different rounds and in different positions. It has been seen that in this comparison, the results of assignments in the traditional methods are not the same as those in the proposed model. This is because traditional methods consider only one type of criterion (process priority), and do not consider the number of processes, %CPU, %Mem, %MV, etc., that is, the load of each node and the overall state of the system.

In this sense, it can be said that besides being the traditional methods, a particular case of the proposed method. This new model allows a more approximate evaluation to the real state of the system, which would allow to obtain better results in the assignments.

The global model, for the example of the traditional methods, does not consider the collection of information on the overall state of the system, nor the predisposition (nodal priority), nor the load of the node, only the process priority is considered.

It should be noted that the results obtained are adjusted to each particular scenario. That is, when the conditions of the scenario change, the results obtained in the application of the Decision Model may be different.

The Fig. 5 shows the different values of the nodal loads. It is observed that the process p_{36} requests the resource r_{13}, whose nodal load has a value of 6.4. By the intensity of the color, you can see that this node is highly loaded.

In the traditional methods, the order of assignment is made only considering the initial priority of the processes. Following this premise, the assignment of resource r_{13} to process p_{36} is made in the first position.

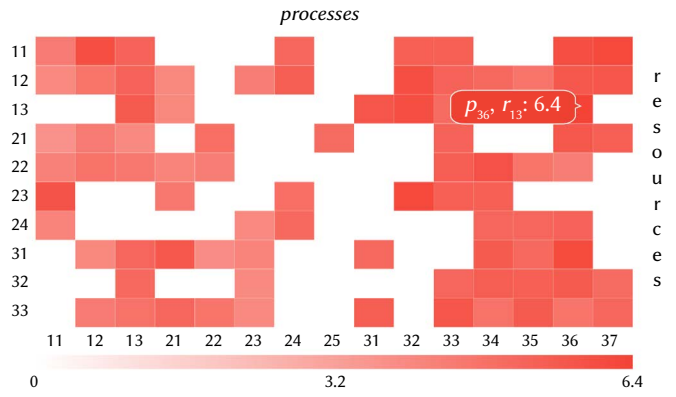


Fig. 5. Heat map showing sample values of nodal loads. The heat map has been created with AMCharts v4 JavaScript library.

As explained above, the proposed method evaluates a set of criteria (including initial priority and nodal load) to determine the order of allocation. Considering that, the same assignment of the previous example, for this method, is in position 56. This is because the node where the assignment is made is heavily loaded. This can be seen in the Table XLVIII.

TABLE XLVIII. COMPARISON OF THE TRADITIONAL METHOD WITH THE PROPOSED METHOD

Nodal Preference	Traditional Method			Proposed Method		
	Process	Resource	Pos.	Process	Resource	Pos.
6.4	p_{36}	r_{13}	1	p_{36}	r_{13}	56
6.2	p_{32}	r_{23}	2	p_{32}	r_{23}	48
6.2	p_{37}	r_{11}	3	p_{37}	r_{11}	35
6.1	p_{36}	r_{31}	4	p_{36}	r_{31}	54
6.0	p_{32}	r_{12}	5	p_{32}	r_{12}	62
6.0	p_{32}	r_{13}	6	p_{32}	r_{13}	26
6.0	p_{36}	r_{11}	7	p_{36}	r_{11}	55
6.0	p_{12}	r_{11}	8	p_{12}	r_{11}	6
5.9	p_{34}	r_{22}	9	p_{34}	r_{22}	23
5.9	p_{11}	r_{23}	10	p_{11}	r_{23}	28

VII. DISCUSSIONS AND COMMENTS

It highlights the dynamism, the magnitude of the nodes, processes and the number of requirements that can be applicable to large systems, through a global solution, or to systems with fewer nodes and requirements. The load of traffic, processes and requirements varies, so that systems that were operating stop operating because they end and others appear, or because nodes with resources and processes that are needed are activated. The nodes can be active, but they are incorporated to the algorithm when some process requests some resource, or some resource is requested by another process of another node. A node can be active but not part of the assignment evaluations.

In each node, an interface is defined between the applications and the operating system, which through a Runtime (software at runtime complementary to the operating system) included in that interface, manages the processes and shared resources, and defines the corresponding scenario, as can see in Fig. 6.

In addition, the Runtimes interact with each other to exchange information and in one of the nodes there is a global coordinating Runtime that evaluates and executes the decision model and the corresponding aggregation operator.

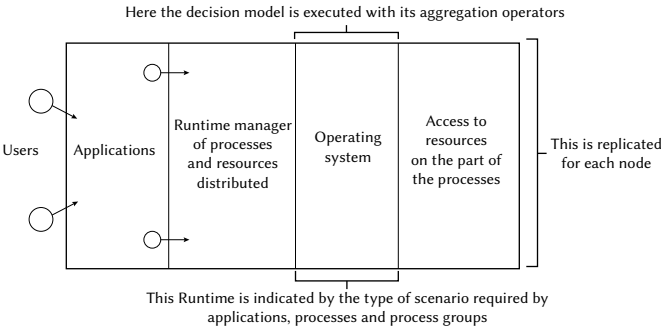


Fig. 6. Runtime global coordinator.

The proposed model manages to establish a consensus that allows groups of processes to access all their resources sequentially and that these cannot be removed until the same group of processes that maintains them, releases them. The order of allocation will be determined by the overall average priority of all the allocations of

each group. The distributed system constantly regulates and updates the local status of each node. The decisions of access to the resources modify these states so it must be readjusted repeatedly, guaranteeing the mutual exclusion, and reordering new priorities. The method should be repeated whenever there are groups of processes that require shared resources.

An important feature to note about the Neat OWA operators used is that the values to be added do not need to be ordered for processing. This implies that the formulation of a neat operator can be defined using the arguments directly instead of the sorted elements.

In the proposed aggregation operator, the weights are calculated based on the context values from which the values to be aggregated arise.

The characteristics of the Decision Model allow us to evaluate the possible alternatives and consequences and thus be able to clearly define the objectives. The best optimization has been achieved by selecting the best possible alternative in each particular case. As the main objective of the proposed model, the environment of distributed execution of processes was considered, the access to shared resources was established according to different consensus requirements. This allows the generation of the sequence of resource allocation to the processes that request them by using the most appropriate aggregation method for each possible scenario, respecting the mutual exclusion in the access to such resources. It has been explained that the decision model uses a Runtime that manages the shared processes and resources and defines the corresponding scenario. The traditional models have been compared by means of an example of application with respect to the proposed model and the considerations of the aggregation operators developed have been commented on.

The aggregation method used, and the data structure mentioned in this work are not fully covered by traditional methods, for example, do not contemplate the predisposition (nodal priority), node load, the nodal state (nodes, processes, groups, resources.) or the overall system state, for the calculation of priorities in resource allocations to processes.

VIII. Conclusions

The proposed model makes it possible for the distributed system to self-regulate repeatedly according to the local state of the n nodes, resulting in an update of their local states, as a consequence of the evolution of their respective processes and the decisions of access to resources: the distributed system in whose groups of processes access to critical resources is executed, produces access decisions to resources that modify the state of the system and readjusts it repetitively, also guaranteeing the mutual exclusion in access to the shared resources, indicating the priority of granting access to each resource and the process to which it is assigned. This process is repeated if there are processes that request access to shared resources.

In this work, fuzzy logic has been used as a tool to innovatively solve the management of resources and processes in distributed systems. The use of the 2-tuple linguistic model allows to improve accuracy and facilitate word processing by treating the linguistic domain as continuous but maintaining the linguistic base (syntax and semantics), through symbolic translation.

What makes the proposed method innovative is that it allows system self-regulation, respects the initial priority of the processes, maintains the status of the nodes updated through the self-regulation, the mutual exclusion is guaranteed, the symbolic translation is incorporated for nodes that use different types of tags, the collaborative nodal priority is established that collaborates in the self-regulation of the system and also includes traditional methods as particular cases of the proposed method.

A prototype simulator has been developed to evaluate the performance of the new decision models and aggregation operators proposed against the main traditional models.

In this research, a software has been developed that simulates the execution of a central runtime of a node located in a distributed system, it is a web application that has been developed with the php language.

When evaluating the results obtained with the simulator, it was possible to verify that the solution produced contemplates an adequate workload balancing, according to the theoretical support used for the development of the simulator. It was also possible to demonstrate that the proposed theoretical solution is more adequate than traditional algorithms that allocate resources to processes only according to the priority of the processes. The values for the figures 4 and 5 has been obtained from the simulator.

For future work, it is planned to develop variants of the proposed method considering other aggregation operators (especially the OWA family) and the possibility of being used by a resource manager shared (instead of centralized as in the proposed method).

It is also planned to continue the development of a simulator with other scenarios.

Acknowledgment

This work has been supported by the Project: "Decision models for resource and process management in distributed systems considering process migration, data imputation and fuzzy logic in new aggregation operators.", code 20F005 of Northeastern National University (Argentine), and the Project: "Development of a simulator for the evaluation of classical and new algorithms for the management of shared resources in distributed systems contemplating mutual exclusion.", code PI 126/20 of the National University of the Southern Chaco (Argentine).

References

[1] A. S. Tanenbaum, *Sistemas Operativos Distribuidos*, México: Prentice - Hall Hispanoamericana S.A., 1996.

[2] A. S. Tanenbaum, *Sistemas Operativos Modernos*. 3ra. Edición: México, Pearson Educación S. A., 2009.

[3] D. Agrawal, A. El Abbadi, "An Efficient and Fault-Tolerant Solution of Distributed Mutual Exclusion," *ACM Transactions on Computer Systems*. Vol. 9, USA, 1991 pp. 1-20.

[4] G. Ricart, A. K. Agrawala, "An Optimal Algorithm for Mutual Exclusion in Computer Networks". *Communication of the ACM*. Vol. 24, USA, 1981, pp. 9-17.

[5] G. Cao, M. Singhal, "A Delay-Optimal Quorum-Based Mutual Exclusion Algorithm for Distributed Systems". *IEEE Transactions on Parallel and Distributed Systems*. Vol. 12, no. 12, USA, 2001, pp. 1256-1268.

[6] S. Lodha, A. Kshemkalyani, "A Fair Distributed Mutual Exclusion Algorithm". *IEEE Transactions on Parallel and Distributed Systems*. Vol. 11, no. 6, USA, 2000, pp. 537-549.

[7] W. Stallings, *Sistemas Operativos*. 5ta. Edición. Madrid, España, Pearson Educación S.A., 2005.

[8] G. Andrews, *Foundation of Multithreaded, Parallel, and Distributed Programming*. Reading, MA, USA, Addison Wesley, 2000.

[9] R. Guerraoui, L. Rodrigues, *Introduction to Reliable Distributed Programming*. Springer-Verlag, Berlin, Germany, 2006.

[10] N. Lynch, *Distributed Algorithms*, San Mateo, CA, USA, Morgan Kauffman, 1996.

[11] G. Tel, *Introduction to Distributed Algorithms*. 2nd ed., Cambridge, UK, Cambridge University Press, 2000.

[12] H. Attiya, J. Welch, *Distributed Computing Fundamentals, Simulations, and Advanced Topics*, 2nd ed., New York, USA, John Wiley, 2004.

[13] P. Saxena, J. Rai, "A Survey of Permission-based Distributed Mutual

Exclusion Algorithms". *Computer Standards and Interfaces*, vol. (25)2, pp 159-181, 2003.

[14] M. Velazquez, "A Survey of Distributed Mutual Exclusion Algorithms". *Technical Report* CS-93-116, University of Colorado at Boulder, 1993.

[15] S.-D. Lin, Q. Lian, M. Chen, Z. Zhang, "A Practical Distributed Mutual Exclusion Protocol in Dynamic Peer-to-Peer Systems". *Proceeding of the Third International Workshop on Peer-to-Peer Systems*, vol. 3279 of Lecture Notes in Computer Sciences, (La Jolla, CA). Springer-Verlag, Berlin, 2004.

[16] L. Sha, R. Rajkumar, J. P. Lehoczky, "Priority inheritance protocols: An approach to real-time synchronization". *Computers, IEEE Transactions on*, vol. 39(9), pp 1175– 1185, 1990.

[17] F. Agostini, D. L. La Red Martínez, J. C. Acosta. "Modeling of the consensus in the allocation of resources in distributed systems". *International Journal of Advanced Computer Science and Applications* (IJACSA). The Science and Information (SAI) Organization, England, U.K. Vol. 9, no. 12., 2018.

[18] L. A. Zadeh, "The Concept of a Linguistic Variable and its Application to Approximate Reasoning-I", *Information Sciences*, Volume 8, Issue 3, pp 199-249, 1975, DOI: 10.1016/0020-0255(75)90036-5.

[19] C. González García, E. R. Núñez-Valdez, V. García-Díaz, B. C. Pelayo G Bustelo, J. M. Cueva Lovelle, "A Review of Artificial Intelligence in the Internet of Things", *International Journal of Interactive Multimedia and Artificial Intelligence - IJIMAI Journal*, 2019, DOI: 10.9781/ijimai.2018.03.004

[20] L. A. Zadeh, "Fuzzy Logic = Computing with Words", *IEEE Transactions On Fuzzy Systems*, VOL. 4, NO. 2, 103-111, 1996.

[21] Ta-Chun Wen, Kuei-Hu Chang, Hsin-Hung Lai, "Integrating the 2-tuple linguistic representation and soft set to solve supplier selection problems with incomplete information", *Engineering Applications of Artificial Intelligence*, Vol. 87, January 2020.

[22] Jerry M. Mendel, "Computing with Words: Zadeh, Turing, Popper and Occam", *IEEE Computational Intelligence Magazine*, pp 10-17, November 2007.

[23] F. Herrera, L. Martinez. "An Approach For Combining Linguistic And Numerical Information Based On The 2-Tuple Fuzzy Linguistic Representation Model In Decision-Making". *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 8, No. 5, pp 539-562, 2000.

[24] F. Herrera, L. Martínez, "A 2-Tuple Fuzzy Linguistic Representation Model for Computing with Words". *IEEE Transactions On Fuzzy Systems*, VOL. 8, NO. 6, pp. 746-752, December 2000

[25] F. Herrera, E. Herrera-Viedma, S. Alonso, F. Chiclana: "Computing with Words in Decision Making: Foundations, Trends and Prospects", *Fuzzy Optimization and Decision Making*, 8, 337-364, 2009 (ISSN: 1568-4539). doi:10.1007/s10700-009-9065-2

[26] S. Zapata, D. Fuentealba, G. Valenzuela. "Aplicación del modelo de representación de información lingüística 2-tuplas con información multigranular". *Revista Trilogía: Ciencia, Tecnología y Sociedad*. Vol. 27, N° 37, pp 110-127, July 2015. Facultad de Ingeniería UTEM.

[27] Jiménez, G.E. and Zulueta, Y., "A 2-tuple linguistic multi-period decision making approach for dynamic green supplier selection". *DYNA*, 84(202), pp. 199-206, September 2017.

[28] M. Ying. "A Formal Model of Computing With Words". *IEEE Transactions On Fuzzy Systems*, VOL. 10, N° 5, October 2002.

[29] J. Liu, L. Yi and Z. Pei, "A new linguistic term transformation method in linguistic decision making". *Journal of Intelligent & Fuzzy Systems* 35, 2403–2412, IOS Press, 2018, DOI:10.3233/JIFS-17987.

[30] David L. la Red Martínez, "Aggregation Operator for Assignment of Resources in Distributed Systems", *(IJACSA) International Journal of Advanced Computer Science and Applications*, Vol. 8, No. 10, 2017.

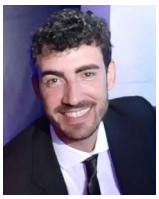
[31] R. Yager. "On Ordered Weighted Averaging Aggregation Operators in Multi-Criteria Decision Making". *IEEE Transactions On Systems, Man and Cybernetics* 18: 183-190, 1988.

[32] R. Yager and G. Pasi. "Modelling Majority Opinion in Multi-Agent Decision Making". *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 2002.

[33] F. Herrera and L. Martínez, "A model based on linguistic 2-tuples for dealing with multigranular hierarchical linguistic contexts in multi-expert decision-making," *IEEE Transactions on Systems, Man and Cybernetics. Part B (Cybernetics)*, vol. 31, no. 2, 2001, doi: 10.1109/3477.915345.

Jorge Tomás Fornerón Martínez

Jorge Tomás Fornerón Martínez received the diploma degree in Information Systems Analysis from the Autonomous University of Asunción (Paraguay) in 2015 and the MSc degree in Informatic and Computation from the National University of Pilar (Paraguay) in 2020. He is professor of General Systems Theory and is Thesis Director in National University of Pilar (Paraguay), since 2015. Currently, he holds the position of Dean at Faculty of Applied Sciences in the same University, and its lines of research are focused on distributed operating systems.



Federico Agostini

Federico Agostini received the diploma degree in Information Systems from the National University of the Northeast (Argentine) in 2013, and the MSc degree in Telecommunications Systems and Networks from the National University of the Northeast (Argentine), in 2019. He is professor of Data Communications and Operating Systems at the same University, since 2013. Currently, he is working at the Northeast Botanical Institute (National University of the Northeast - National Council for Scientific and Technical Research), and its lines of research are focused on bioinformatics.

David L. la Red Martínez

David L. la Red Martínez received the diploma degree in Information Systems from the National University of the Northeast (Argentine) in 1979, the MSc degree in Informatic and Computation from the National University of the Northeast (Argentine), in 2001, the Specialist degree in University Teaching from the National University of the Northeast (Argentine), in 2003 and the Doctoral degree in Computer Systems Engineering from the University of Malaga (Spain), in 2011. He made a postdoctoral research stay in cyber-physical security systems at Florida Atlantic University (USA), in 2019. He is professor of Databases, Data Communications and Operating Systems at the National University of the Northeast (Argentine), since 1983. Currently, its lines of research are focused on distributed systems and data mining.

A Hybrid Secure Cloud Platform Maintenance Based on Improved Attribute-Based Encryption Strategies

Abhishek Kumar[1]*, Swarn Avinash Kumar[2], Vishal Dutt[3], Ashutosh Kumar Dubey[1], Sushil Narang[1]

[1] Chitkara University School of Engineering and Technology, Chitkara University, Himachal Pradesh (India)
[2] Indian Institute of Information Technology, Allahabad, UP (India)
[3] Department of Computer Science, Aryabhatta College, Ajmer (India)

Received 3 April 2021 | Accepted 27 August 2021 | Published 12 November 2021

Abstract

In the modern era, Cloud Platforms are the most needed port to maintain documents remotely with proper security norms. The concept of cloud environments is similar to the network channel. Still, the Cloud is considered the refined form of network, in which the data can easily be stored into the server without any range restrictions. The data maintained into the remote server needs a high-security feature, and the processing power of data should be high to retrieve the data back from the respective server. In the past, there were several security schemes available to protect the remote cloud server reasonably. However, the attack possibilities over the cloud platform remain; only all the researchers continuously work on this platform without any delay. This paper introduces a hybrid data security scheme called the Improved Attribute-Based Encryption Scheme (IABES). This IABES combines two powerful data security algorithms: Advanced Encryption Standard (AES) and Attribute-Based Encryption (ABE) algorithm. These two algorithms are combined to provide massive support to the proposed approach of data maintenance over the remote cloud server with high-end security norms. This hybrid data security algorithm assures the data cannot be attacked over the server by the attacker or intruder in any case because of its robustness. The essential generation process generates a credential for the users. It cannot be identified or visible to anyone as well as the generated certificates cannot be extracted even if the corresponding user forgets the credentials. The only way to get back the certification is resetting the credential. The obtained results prove the accuracy level of the proposed cypher security schemes compared with the regular cloud security management scheme, and the proposed algorithm essential generation process is unique. No one can guess or acquire it. Even the person may be the service provider or server administrator. For all, the proposed system assures data maintenance over the cloud platform with a high level of security and robustness in Quality of Service.

Keywords

Advanced Encryption Standard (AES), Attribute-Based Encryption (ABE), Cloud Security, Improved Attribute-Based Encryption Scheme (IABES).

DOI: 10.9781/ijimai.2021.11.004

I. Introduction

CLOUD Computing environments are supporting users to manage their data globally over the remote server with the high end of security. The cloud server processes data remotely and provides the resulting features to the client port without any hurdles.

Many research papers illustrated that cloud computing environments are highly secure and robust in their performance and cipher policies [1],[2],[3],[4]. However, the problems in the cloud environment usually sustain until now. The issues are growing every day, and the researchers are identifying many new mechanisms day by day to tackle these security issues [5], [6].

The cloud computing environment requires a new methodology to avoid security issues and provide a high-level security measure to the proposed approach in an exemplary manner without any interventions. This paper introduces a new hybrid methodology, which integrates the two best algorithms and operates the proposed cloud server system accordingly, called Improved Attribute-Based Encryption Scheme (IABES).

This proposed algorithm combines two powerful algorithms such as Advanced Encryption Standard (AES) and the Attribute-Based Encryption (ABE) Mechanism. The concept of AES follows the Rijndael process, which is formed as a cipher-block with a 256-bit encryption technique. Each block is divided into 128-bit capacity with associated essential space. This is one of the powerful crypto algorithms usually followed over many real-time applications such as banking, mobile applications, etc. The next one is called ABE, in which the algorithm is operating based on the attributes and process the data accordingly based on cipher keys [7], [8], [9].

This is also a robust security principle, allowing users to maintain the data into the server end without any hurdles. But instead of keeping the public crypto keys, in this ABE approach, a new essential generation standard is followed based on user input attributes and based on that input attributes. The keys are generated, and the input data or document is encrypted [10], [11], [12]. So, that the encryption

* Corresponding author.
E-mail address: abhishek.kumar@chitkara.edu.in

standards are highly unique with such systems [13], [14], [15]. There are several approaches which has been covered the scenario of different ciphertext policies and other aspects for dealing the problems in different types of security system [16], [17], [18], [19].

However, the individuality of above mentioned two algorithms are working fine, but in the case of higher-end security threats, both of these algorithms struck up into a specific range [5], [8], [10]. So, that a new algorithm is designed based on the efficiency of the two separate algorithms, such as ABE and AES, and named the hybrid algorithm as IABES. It adapts the benefits of mentioned two algorithms and provides the ultimate security features over the proposed cloud server management system. The submitted paper is intended to make the new algorithm concentrate on security threats concerning different attack possibilities such as the Query-Regeneration attack, Query Modification attack, Searchable-Query attack and the Query-Removal attack. These different kinds of security threats are handled adequately over the proposed system with an advanced cipher handling algorithm. These attacks are coming under the SQL Injection attack category, which will be illustrated in detail below.

A. Query-Regeneration Attack

The attackers or intruders usually attack the server from the client end only and generate a query to regenerate multiple data over the standard table presented into the server end. For example, the table contains ten numbers of records with different unique identities; this kind of Query-Regeneration attack creates duplication over the proposed system server end, which will automatically degrade the performance of the entire server management system. The attackers usually try these kinds of attacks to copy the whole server data and place it again into the weak node presented into the network.

B. Query Modification Attack

The attackers try to modify the data available into the server using Query Modification logic. The data presented into the server must be integrity enabled and robust against multiple scenarios of attacks. But in the regular cloud server maintenance system, the usual attack is called a query modification attack. Consider the design of government organization; if the quotation is raised for some commercial contract, if the attackers modify the quoted amount, the complete reference gets spoiled. These kinds of attacks are presented based on modification attack over the server. It is considered one of the most dangerous query attacks in the information technology industry.

1. Searchable-Query Attack

The attackers not only try to attack the data or document presented over the remote server instead attempted to view the records submitted over the server without having any access control norms and proper credentials. For some of the weak cloud servers, the attacker can easily surf and get the records without the knowledge of the respective data owner. This kind of attack is usually raised to identify an individual's personal or official details and target the corresponding individual based on private information. This kind of attack is also crucial to concentrate more on it over the proposed data handling approach over the proposed system.

2. Query-Removal Attack

This kind of Query Removal attack is composed to remove the data presented into the server, which causes some severe reflections over the cloud server management scheme. Because of the removal, the entire trust over the server will lack, and this removal problem raises many legal issues in the industry. For example, if the organization maintains the employee salary records into the server means, the documents need to be robust in all ways. Suppose any intruder removes the basic pay of all employees in the server or deletes all employee records over the server means. In that case, entire operations are collapsed on the company and employees facing massive trouble. This has also happened in most server mediums; that is why all are periodically back up the server with some proper intervals.

All these issues are handled using our proposed approach of Cloud-based data maintenance concerning Improved ABE Scheme IABES. This proposed algorithm has taken care of all these mentioned injection attacks, provides the problem accessible server to the users in a suitable manner, and provides the high-level security threats elimination mechanism to real-world cloud servers.

The significant motivation behind this study is that, in the modern era, the cloud computing platform is being used at a very high level. In today's technology era, the importance of data is immense. Given the increased importance of data, its security needs to be taken very seriously. Even though people are using cloud storage in abundance, there are still many apprehensions regarding the safety of the data, which proves to be an important and significant reason for reducing the use of cloud systems. And due to this, the use of cloud storage remains limited and compressed.

The main objective of this research is to provide a high degree of security to the data stored in the cloud storage. So that the data can be protected from attackers or intruders and its use can be promoted. The processing power of the data must be high to retrieve the data from the respective server. Hence, managing data access time is also an essential part of cloud storage systems to maintain the meaningfulness of data availability and data security. To provide tight security to the data without compromising on the processing power and provide seamless access to the data is the study's main objective.

The rest of this paper have been arranged in the following manner: Section II illustrates the proposed system methodologies in detail with proper algorithm flow. Section III demonstrates the result and discussion portion of the paper, and the final section, Section IV, illustrates the concept of conclusion and future scope.

II. Proposed System

It is difficult to find out the intruders and trace them over the digital world in this modern era. The security mechanisms available nowadays provide acceptable security norms to the clients to preserve their data safely. But the consistency and stability of such security mechanisms are still raising an issue to manage the data integrity over different levels. The proposed system is intended to provide an efficient data security and integrity maintenance scheme, which will be suitable for all kind of textual data maintenance over the cloud server in an intelligent manner.

A. Components of IABES

- User: It is an entity that is going to consume the services of cloud storage.
- Authentication: This is an essential component in architecture. Through this, the identity of the participant or user is checked by the system. If a user wants to join this system and go for its services, he has to first go through the proposed authentication process. He has to prove his identity that he is an Authorized and Authentic user.
- Text Uploading: As soon as the system confirms that the user is authentic, the user gets permission to upload the data.
- Secure Data: When the data is uploaded, then the process of securing the information is started. In this process, encryption and decryption is an important addition. Some algorithms for encryption and decryption have been proposed in this study.
- Cloud server: The cloud server is included as an essential

component in this architecture. The data uploaded by the user will be stored in encrypted form on the cloud server itself so that the user can access the data anywhere and anytime via an internet connection. Just as the uploading of data has to go through authentication to access the data, the user must also follow this authentication process.

The significant contributions of the proposed algorithm IABES are as follows:

- In this proposed approach, the attackers or intruders who share their unique secret keys to others, whatever may be the purpose, needs to be traceable from our proposed logic. The present system needs to generate random access to extract some portion of the user identities and develop a new secret key so that the generated private key cannot be guessed or identified by the attackers in any case.
- The proposed approach of secure hashing allows the user to generate dynamic user credentials (refer to Fig. 1.) concerning the user's identity and the random key generation process. With these associations, a new dynamic credential is generated, and that will be forwarded to the user mail with decrypted mode. The respective user can only get to know the credential ultimately until and unless without the user knowledge. It won't be shared with anyone, and this credential cannot be breakable by anyone because we know that the secure hashing technique is a unidirectional encryption scheme using this SHA based data hashing and storing those credential values to the server so that the server administrator cannot retrieve the credentials from the server. Since each characteristic of the clients is analyzed based on the tracing values stored on the server end when the feature contained in the key satisfies the access control norms of the server, that would be able to be decoded effectively, and the particular user only can access all the features of the proposed approach. Contrasting and the related detectable ABE Scheme, it is of functional significance to present the idea of the secret key into the noticeable proposed system so that our proposed method is nearer to the genuine circumstance.
- Under the suspicion of the proposed approach, the developed hybrid algorithm of IABES is demonstrated to enhance protection from the plaintext attack in the standard model, and the trial results show that the proposed scheme of IABES is viable in the cloud condition. The proposed approach, Improved ABE Scheme (IABES), is intended to provide security and proper access control norms over the cloud server with the help of the following procedures: User Attribute Segregation Secret Key Generation Process, Encryption Process and Decryption Process. All these processes are described below. The architectural view of the proposed system is shown in Fig. 2.

B. User Attribute Segregation

This attribute segregation process as illustrated in Fig. 3., gathers the user attribute such as name, mobile number and email-id from the respective user and process the collected data with the segregation principle. For example, the User X identity is grasped and segregates the required features from that collected attribute employing data split logic over the proposed approach. The collected attributes are used to generate the dynamic credential of the user. This logic is used only to extract the attribute from the user identities, which is sufficient for the credential; instead, it generates a robust cloud network credential for further access. The process is shown in Algorithm 1.

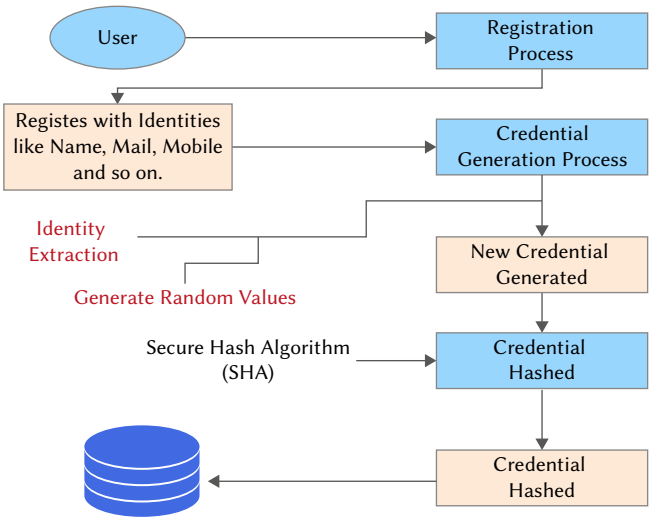
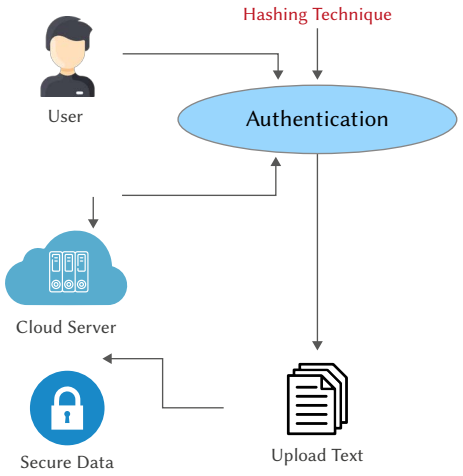


Fig. 1. Secured Credential Generation Process using SHA.



Improved Attribute Based Encryption Scheme (IABES)

Fig. 2. IABES Architectural View.

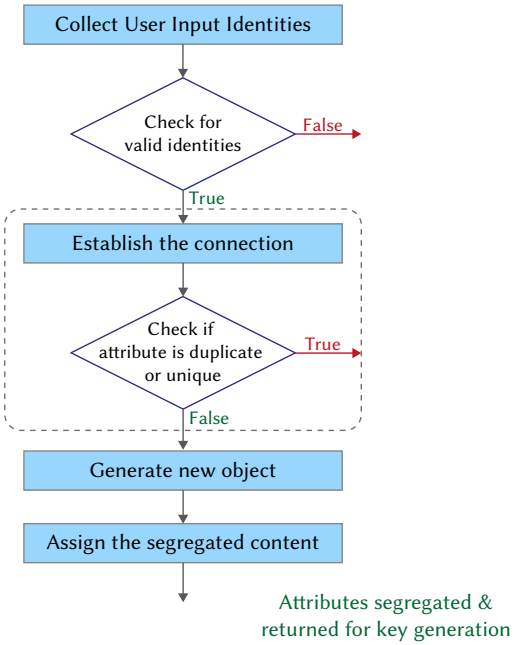


Fig. 3. Work flow of user segregation Process.

Algorithm 1: User Attribute Segregation

Input: User Attributes (Name, Mobile Number, Mail-ID and Contact)

Output: Segregated Portion of User Credential.

Step-1: Collect user input identities from the client end web portal.

Input = id_c, id_{wp}

Step-2: Check for valid identities. In this step, the accumulated information from Step-1 is validated and allowed the user once the given identities are correct.

$Validate(Input(id_c, id_{wp}))$

Step-3: Establish the connection between Client end and the Server end. Check whether the given attribute is duplicate or unique.

$Conn(Client,Server))$

$Obj_i == Input\ ? \rightarrow return(True):Obj_{i+1} = Input\ ?$

$\rightarrow return(True):Obj_{i+n} = Input\ ?$

$\rightarrow return(True):return(False)$

Step-4: Once the return statement returns false means, the identity is unique.

Step-5: Generate new object for String Segregation Obj_{Str}

Step-6: Assign the segregated content of the user identity to created string object.

$\text{Obj}_{\text{str}} = \sum_{i=0}^{n} I(1,2,\ldots,n)$

Step-7: User attributes segregated and returned to further process of key generation.

$Sua \xrightarrow{Generates} P(Key_{Generation})$

Here 'S' denotes Segregated attributes, and 'P' denotes the process.

C. Secret Key Generation and Processing

The secret key generation process is dependent on the Attribute Segregation process over the proposed approach of IABES. The processed attributes from the user attribute segregation scheme are collected over this approach as an input and generate the random key based on the RandomClass function and merge the created random key with the already segregated user attribute. So, that the generated secret key is ultimately vital to compare to any other traditional approaches. No one can judge this kind of secret keys, or it cannot be guessable to others. As well as the created secret key is not only enough for authentication, because of providing high-level security norm, the dynamic one-time password will be generated and send to the respective users' mail-id after verifying the credentials given by the user. The system allows the user to proceed further once the given high secured one-time password is correct; otherwise, it blocks the user to proceed further. It is shown in Algorithm 2.

$n = pq\ \varphi(n) = (p-1)(q-1)$

$e, 1 < e < \varphi(n)$ gcd $(e, \varphi(n) = 1$

$d = e^{-1}\ mod\varphi(n)$

Where,

P, Q ⟶ Prime Numbers

n ⟶ Composite Numbers

D. Encryption Process

The proposed system follows the Improved ABE Scheme as illustrated in Fig. 4., which integrates two powerful cipher algorithms: AES and ABE.

The AES scheme is a traditional scheme that encrypts the given text document or text data into cipher form based on the Rijndael encryption scheme with 256-bit operational frequency and an essential algorithm over the innovative real-world application at present. The proposed system algorithm integration is called ABE, the encryption algorithm but the difference between AES and ABE is based on key

Algorithm 2: Secret Key Generation and Processing

Input: Segregated User Attribute $Segg_{attr}$

Output: Secret Key Sk.

Step-1: Collect the segregated user attribute from segregation process as:

$SegregationProcess \xrightarrow{yields} SgAttr$

Step-2: Initiate the function for generating random key, using Random_Class Rc.

Create Function and assign new object New_{Obj} to Random_Class Rc.

Provide int_{Min} and int_{max} values as a parameter to the Random_Class Rc as:

Function $Integer\_Key_{Generation}$ (int_{Min}, int_{max}) $Rc \leftarrow NewObj$;

Step-3: The Rc methodology takes min value and max value parameter as an input and generates a new random key between these given parameter values.

$int_{Min}, int_{max} \xrightarrow{Generates} Random_{Key}$

Step-4: Return the generated random value to the required function.

$Return \rightarrow Random_{Key} \rightarrow RequiredFunction$

Step-5: Random value retained and stored that into a new variable called Rn as: $Rn = RequiredFunction(int_{Min}, int_{max})$

Step-6: Concatenate the generated random value Rn and Obj_{Str} to generate a new credential to user as:

$String_{Cred} = Rn + Obj_{Str}$

Send this $Cred$ to the respective user mail-id;

Step-7: Authenticate process required the given credential and one time password to access the system further.

Step-8: Checks the input credential with the existing server credential over encrypted form.

Step-9: Checks the one-time password with server session password.

Step-10: Allows the user to proceed further, if credential and the one-time password matched with the server credential and the server session.

Step-11: Access Control provided properly based on the generated secret key

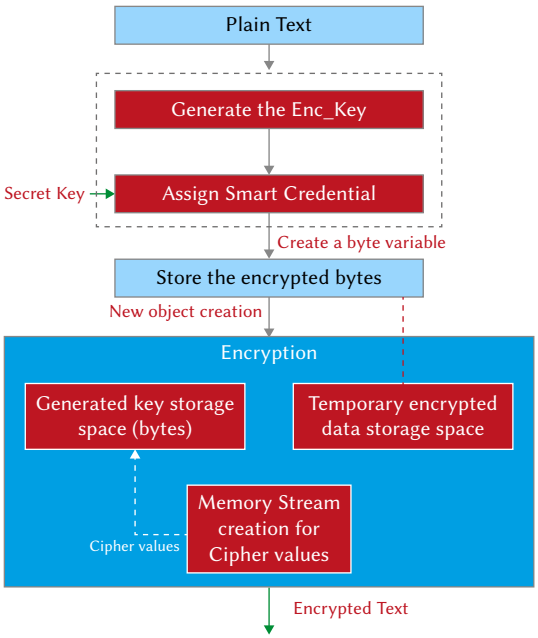


Fig. 4. Detailed work flow of encryption.

generation. In the ABE, the key generation process is based on user attributes, and the ABE generates the symmetric key for processing. All these features are integrated and make the hybrid algorithm process the entire system more securely. The Algorithm 3 explains the encryption process in detail. The following algorithm explains the encryption process in detail.

$C = E(K_E, P)$

$P = D(K_D, E(K_E, P))$

Where:

$C \rightarrow$ Cipher Text,

$E \rightarrow$ Encryption

$K_E \rightarrow$ Encryption Key

$K_D \rightarrow$ Decrypton Key

Algorithm 3: Encryption Algorithm

Input: Text Data or Document from user end as $Plain_{Text}$

Output: Encrypted Cipher Data (Cipher text) as: $Cipher_{Text}$

Step-1: Collect the Plain text or data from user end.

Step-2: Generate the string variable called *Enc_Key* and assign the generated smart credential from Algorithm2 as:

$Enc_{key} \leftarrow String_{Cred}$

Step-3: Create a byte variable to store the encrypted bytes as:

$Byte_{var[100]} \leftarrow Encoding\_Unicode\_GetBytes.GetBytes(Enc\_Text)$

Step-4: Create a new object to perform encryption based on advanced encryption procedure with respect to Rijndael process as:

$AES\_Encryptor_{aes\ enc} \leftarrow AES_{Encryptor}.Create()$

$RFC\_2898\_Derive\_Bytes_{Derived\_Bytes}$

$\leftarrow new\ (RFC\_2898\_Derive\_Bytes(Enc_{key}, Byte_{var[100]}))$

$0_x49, 0_x50, 0_x51, \cdots 0_xn$

Step-5: Generate the key storage space with respect to Derived_Bytes over Step-4 as:

$$StorageSpace_{Key} \leftarrow Derived_{Bytes}$$

Step-6: Generate the temporary encrypted data storage space with respect to Derived_Bytes over Step-3 and key storage space generated over Step-5 as:

$$Encryptor_{Key} \leftarrow Derived_{Bytes}.getBytes(32)$$

$$Encryptor_{Data} \leftarrow Derived_{Bytes}.getBytes(16) + Enc_{key}$$

Step-7: Create $Memory_{Stream}$ for storing the cipher values one by one to the generated encrypted data storage space over Step-6.

$$Memorystream_{obj} \leftarrow Memory_{stream}$$

Step-8: Store the encrypted data to the storage space in byte format.

$$Crypto_{stream}\left\{\begin{matrix} Memorystream_{obj}, Enc\_Create\_Encryptor(), \\ Crypto\_Stream\_Mode[Write] \end{matrix}\right\};$$

Step-9: Return the encrypted text.

$$return \rightarrow Cipher_{Text}$$

E. Decryption Process

The proposed system decryption process illustrated in Fig. 5., is just a reverse of the encryption process, in which it is associated with the AES procedure.

Still, the variation over here is the key used to decrypt the data is unique compared to the traditional approach, which is extracted from the user attributes, and the dynamically generated key is mailed to the receiver. The receiver needs to provide the correct dynamic secret access to the system to decrypt the data. In this case, the given key is the valid means. The data is decrypted and allows the user to download the same otherwise, and the system blocks the user to proceed further. The algorithm 4 clearly illustrates the process of decryption straightforwardly.

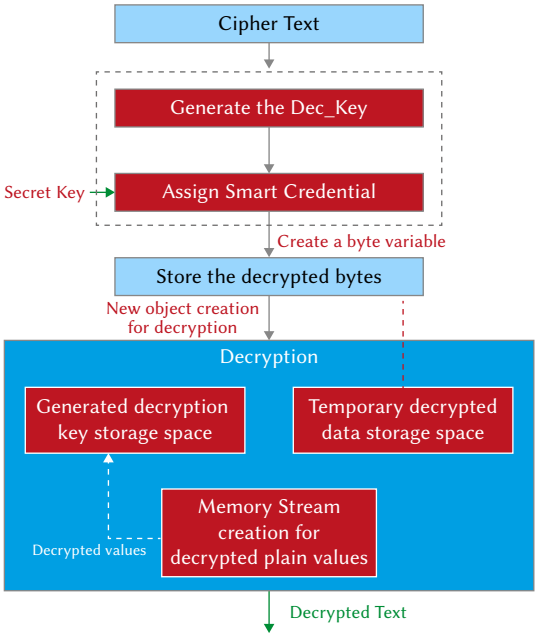


Fig. 5. Detailed work flow of decryption.

P = D(C)

Where:

C → Plain Text,

D → Decryption Function

P → Plain Text

Algorithm 4: Decryption Algorithm

Input: Encrypted Cipher Data $Cipher_{Text}$.

Output: Decrypted Text Data or Document Dec_{Text}

Step-1: Collect the Cipher text from the server end.

Step-2: Generate the string variable called Dec_{key} and assign the generated smart credential from Algorithm2.

$Dec_{key} \leftarrow String_{Cred}$

Step-3: Create a byte variable to store the decrypted bytes as:

$Byte_{var[100]} \leftarrow Encoding\_Unicode\_GetBytes.GetBytes(Dec\_Text)$

Step-4: Create a new object to perform decryption based on advanced encryption procedure with respect to Rijndael process.

$AES\_Decryptor_{aes\_dec} \leftarrow AES\_Dncryptor.Create()$

$RFC\_2898\_Derive\_Bytes_{Derived\_Bytes}$

$\leftarrow new\ (RFC\_2898\_Derive\_Bytes(Dec_{key}, Byte_{var[100]}))$

$0_x49, 0_x50, 0_x51, \cdots 0_xn$

Step-5: Generate the key storage space with respect to Derived_Bytes over Step-3.

$$StorageSpace_{Key} \leftarrow DerivedBytes$$

Step-6: Generate the temporary decrypted data storage space with respect to Derived_Bytes over Step-3 and decryption key storage space generated over Step-5.

$$Decryptor_{Key} \leftarrow Derived\_Bytes.getBytes(32)$$

$$Encryptor_{Data} \leftarrow Derived\_Bytes.getBytes(16) + Dec_{key}$$

Step-7: Create a memory stream for storing the decrypted plain values one by one to the generated decrypted data storage space over Step-6.

$$Memorystream_{obj} \leftarrow Memory_{stream}$$

Step-8: Store the decrypted data to the storage space in byte format.

$$Crypto_{stream}\left\{\begin{matrix} Memorystream_{obj}, Dec\_Create\_Encryptor(), \\ Crypto\_Stream\_Mode[Write] \end{matrix}\right\}$$

Step-9: Return the decrypted text.

$return \rightarrow Dec_{Text}$

III. Results and Discussion

In this summary, the experimental analysis of the proposed algorithm Improved ABE Scheme is to be discussed transparently with a practical graphical outcome. The entire process estimation proves the performance ratio of the proposed system with IABES is high compared to the classical cloud service structure. The proposed system performance and accuracy measures are estimated in terms of cost and time required to process the entire system over a real-time working environment. The whole programming and analysis are composed by using Microsoft supported tool platform, and the resulting units are properly accumulated pleasingly. The graphical estimations prove the resulting summary of the proposed approach and the proposed encryption and decryption accuracy levels in detail. Table I illustrates the performance measures of the proposed method, and that has been compared with many existing algorithms.

Table I shows the proposed IABES approach for evaluating the performance of key generation process. On the other hand, the ARMAX [20], took 29.61 milliseconds as response time with 80% robustness and 92% of accuracy. Whereas, the proposed approach took very less response time of 10.26 milliseconds, and achieved 92% of robustness with 98% of accuracy.

TABLE I. Performance Measures of Proposed Approach

Algorithm	Response Duration (ms)	Robustness (%)	Accuracy (%)
[20]	29.61	80%	92%
IABES	10.26	92%	98%

Fig. 6. illustrates the evaluation of the Secret Key generation process and its time requirement of the proposed system, which is explained in terms of several taken user attributes and key generation duration.

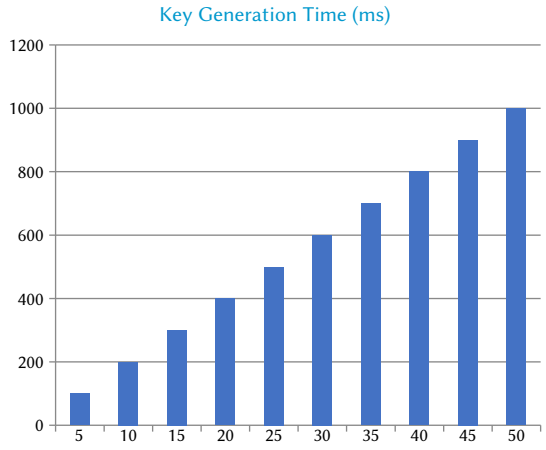


Fig. 6. Proposed algorithm key generation time evaluation.

Fig. 7. illustrates the proposed algorithm encryption time evaluation concerning the evaluation of processing time in milliseconds versus the number of user attributes taken for processing. Fig. 8. illustrates the proposed algorithm decryption time evaluation concerning the evaluation of processing time in milliseconds versus several users' attributes taken for processing. Fig. 9. illustrates the impact of the number of attributes used by users on the time cost. To analyze the effect of the number of attributes used by users on time cost, we set the total number of attributes to 21 and change the number of attributes used by users, the number of nodes in the access tree, and the access tree's depth shift together. This can affect KeyGen(R) and test as previously analyzed. The change in the number of attributes used by users will also affect the time consumption of the proposed algorithm. The impact is observed that the time cost of the proposed algorithm has a positive linear correlation with the number of attributes used by users. The proposed approach needs to bind attributes to the ciphertext, and computing the corresponding cost and time for each attribute results in a longer encryption time [21]. Fig. 10. illustrates that the impact of the number of keywords in the cipher-text on the time cost [21].

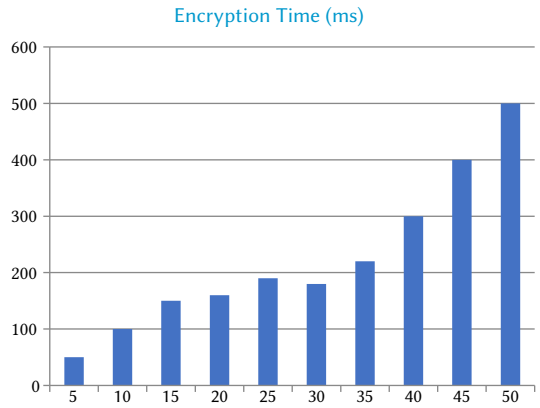


Fig. 7. Proposed algorithm encryption time evaluation.

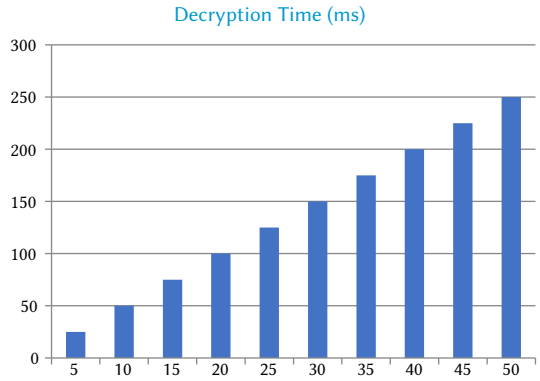


Fig. 8. Proposed algorithm decryption time evaluation.

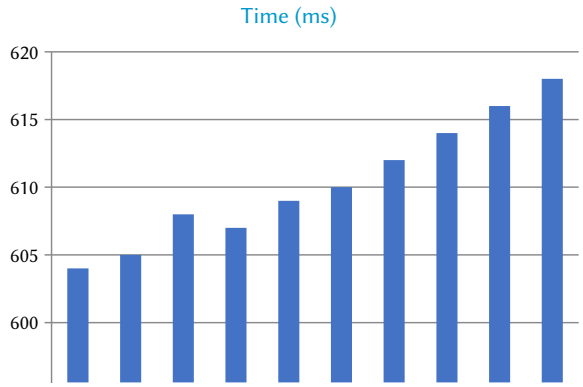


Fig. 9. Impact of the number of attributes used by users on the time cost.

IV. Conclusion and Future Scope

This paper demonstrates the performance and security features of the proposed algorithm Improved ABE Scheme (IABES). It shows the accuracy levels as high over the result and discussion section. This paper provides the secure hashing principle to prove the access control security in a detailed manner over the proposed system summary section. The hashing algorithm provides deep security to the users during authentication into the system with complete

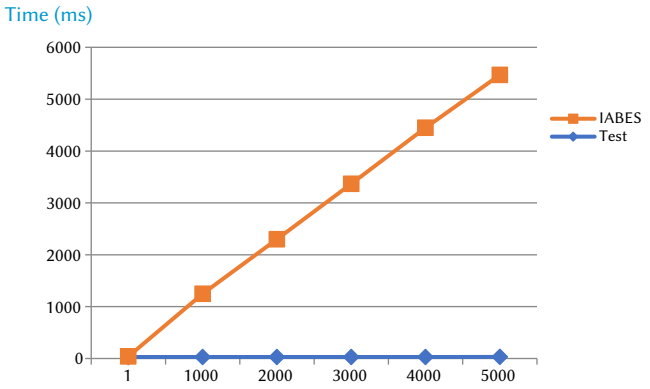


Fig. 10. Impact of the number of keywords in the cipher text on the time cost.

access control. The key generation process is handled through ABE logic, in which it accumulates the secret processing key from user attributes instead of using a general symmetric key principle. So, the proposed algorithm's security and processing nature are high as well as the proposed algorithm time efficiency is proved over the result and discussion section. The time accuracy and consumption scenario will diversely prove the cost efficiency of the proposed approach. The entire work is more suitable to provide cloud storage security to the data maintenance scheme with proper access control norms. In future, the work is further extended by adding some deep learning or machine learning algorithms to train the machine based on security threats. That kind of artificial intelligence approaches improves the efficiency of the overall system in terms of robustness and accuracy.

References

[1] N. A. Wigati, A. Wibisono, and A. N. Hidayanto, "Challenges of Infrastructure in Cloud Computing for Education Field: A Systematic Literature Review," Insight, vol. 43, no. 23, pp. 351-358, 2021.

[2] G. Ramachandra, M. Iftikhar, and F. A Khan, "A comprehensive survey on security in cloud computing," Procedia Computer Science, vol. 110, pp. 465-72, 2017.

[3] M. Samvatsar and P. Kanungo, "An Analytical Review and Analysis for The Data Control and Security in Cloud Computing," International Journal of Advanced Technology and Engineering Exploration, vol. 7, no. 73, pp. 241-246, 2020.

[4] S. M. Sasubilli, A. K. Dubey, and A. Kumar, "A Computational and Analytical Approach for Cloud Computing Security with User Data Management," In International Conference on Advances in Computing and Communication Engineering, IEEE, pp. 1-5, 2020.

[5] J. Li, W. Yao, J. Han, Y. Zhang, and J. Shen, "User Collusion Avoidance CP-ABE with Efficient Attribute Revocation for Cloud Storage," IEEE Systems Journal, vol. 12, no. 2, pp. 1767-1777, 2018.

[6] J. Li, N. Chen and Y. Zhang, "Extended File Hierarchy Access Control Scheme with Attribute Based Encryption in Cloud Computing," IEEE Transactions on Emerging Topics in Computing, vol. 9, no. 2, pp. 983-993, 2019.

[7] J. Li, Q. Yu, and Y. Zhang, "Hierarchical Attribute-Based Encryption with Continuous Leakage-Resilience," Information Sciences, vol. 484, pp. 113-134, 2019.

[8] J. Li, Q. Yu, Y. Zhang, and J. Shen, "Key-policy attribute-based encryption against continual auxiliary input leakage," Information Sciences, vol. 470, pp. 175-188, 2019.

[9] A. K. Dubey, A. K. Dubey, M. Namdev, and S.S. Shrivastava, "Cloud-user Security Based on RSA And MD5 Algorithm for Resource Attestation and Sharing in Java Environment," In Sixth International Conference on Software Engineering, IEEE, pp. 1-8, 2012.

[10] J. Li, Y. Wang, Y. Zhang, and J. Han, "Full Verifiability for Outsourced Decryption in Attribute Based Encryption," IEEE Transactions on Services Computing, vol. 13, no. 3, pp. 478-487, 2017.

[11] X. Liu, J. Ma, J. Xiong, Q. Li, and T. Zhang, "Ciphertext Policy Weighted Attribute-Based Encryption Scheme," Journal of Xi'an Jiaotong University, vol. 47, no. 8, pp. 4448, 2013.

[12] Y. T. Wang, K. F. Chen, and J. H. Chen, "Attribute-based traitor tracing," Journal of Information Science and Engineering, vol. 27, no. 1, pp. 181195, 2011.

[13] W. J. Chung and T.H. Cho, "A Security Scheme Based on Blockchain and A Hybrid Cryptosystem to Reduce Packet Loss in IoV," International Journal of Advanced Technology and Engineering Exploration, vol. 8, no. 81, pp. 945-956, 2021.

[14] X. Liu, J. Ma, J. Xiong, Q. Li, and J. Ma, "Ciphertext-Policy Weighted Attribute-Based Encryption for One-Grained Access Control," In International Conference on Intelligent Networking and Collaborative Systems, IEEE, pp. 51-57, 2013.

[15] Z. Liu, Z. Cao, and D. S. Wong, "White-Box Traceable Ciphertext-Policy Attribute-Based Encryption Supporting any Monotone Access structures," IEEE Transactions on Information Forensics and Security, vol. 8, no. 1, pp. 76-88, 2013.

[16] J. Li, Q. Huang, X. Chen, S. S. M. Chow, D. S. Wong, and D. Xie, "Multi-Authority Ciphertext-Policy Attribute-Based Encryption with Accountability,'' In Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security, IEEE, pp. 386-390, 2011.

[17] Y. Jiang, W. Susilo, Y. Mu, and F. Guo, "Ciphertext-Policy Attribute-Based Encryption Against Key-Delegation Abuse in Fog Computing," Future Generation Computer Systems, vol. 78, pp. 720-729, 2018.

[18] Q. Li, H. Zhu, Z. Ying, and T. Zhang, "Traceable Ciphertext-Policy Attribute-Based Encryption with Variable Outsourced Decryption in Ehealth Cloud," Wireless Communications and Mobile Computing, vol. 2018, pp. 112, 2018.

[19] J. Zhou, Z. Cao, X. Dong, and X. Lin, "TR-MABE: White-Box Traceable and Revocable Multi-Authority Attribute-Based Encryption and Its Applications to Multi-Level Privacy-Preserving E-Healthcare Cloud Computing Systems,'' In proceedings of IEEE Conference on Computer Communication, IEEE, pp. 2398-2406, 2015.

[20] F. Piltan, S. TayebiHaghighi, and N. B. Sulaiman, "Comparative Study Between ARX and ARMAX System Identification," International Journal of Intelligent Systems and Applications, vol. 9, no. 2, pp. 25-34, 2017.

[21] Y. Yu, J. Shi, H. Li, Y. Li, X. Du, and M. Guizani, "Key-Policy Attribute-Based Encryption with Keyword Search in Virtualized Environments," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 6, pp. 1242-1251, 2020.

Abhishek Kumar

Dr. Abhishek Kumar is Doctorate in computer science from University of Madras and done M.tech in Computer Sci. & Engineering from Government engineering college Ajmer, Rajasthan Technical University, Kota India. He has total Academic teaching experience of more than 7 years with more than 80 publications in reputed, peer reviewed National and International Journals, books & Conferences. He has guided more than 20 M.Tech Projects and Thesis and guiding 2 PhD Scholar. His research area includes- Artificial intelligence, Image processing, Computer Vision, Data Mining, Machine Learning. He has been Session chair and keynote Speaker of many International conferences, webinars in India and Abroad. He has been the reviewer for IEEE and Inderscience Journal. He has authored/Co-Authored 6 books published internationally and edited 16 books (Published & ongoing with Elsevier, Wiley, IGI GLOBAL Springer, Apple Academic Press, De-Grueter and CRC etc. He has been member of various National and International professional societies in the field of engineering & research like Senior Member of IEEE, IAENG (International Association of Engineers), Associate Member of IRED (Institute of Research Engineers and Doctors), He has got Sir CV Raman National award for 2018 in young researcher and faculty Category from IJRP Group. He is Editor of Special issue in the Journal Computer materials and continua [SCI and SCOPUS.IF- 4.98] and Intelligent Automation and Soft Computing [SCI, SCOPUS, IF-1.276] Cognitive Neuro dynamics, Springer [SCI, SCOPUS, IF-3.925].

Swarn Avinash Kumar

Mr. Swarn Avinash Kumar is a Research Engineer at the self-driving division of Lyft. He has previously worked at AI divisions of Google and Amazon as well. He has a total professional research experience of 6 years with more than 10 publications in reputed, peer-reviewed national and international journals, books & conferences. His research area includes: Artificial intelligence, computer vision, robotics, data mining, machine learning. He has filed multiple patents in the field of AI. He has been the reviewer for IEEE and IET conferences. He has co-authored 2 books (ongoing with Institution of Engineering and Technology & Eureka publications) and edited 2 books(Published & ongoing with Institution of Engineering and Technology & Eureka publications).

Vishal Dutt

Dr. Vishal Dutt is Doctorate in computer science from University of Madras, Chennai and has done MCA (Gold Medalist) from MDS University, Ajmer, Rajasthan, India. He has been working as the Assistant Professor of Computer Science at Aryabhatta College, Ajmer and also visiting faculty in Maharshi Dayanand Saraswati University (State Govt. University) Ajmer. He has total Academic teaching experience of more than 4 years. He has more than 35 publications in reputed, peer reviewed National and International, Scopus Journals & Conferences and Book Chapters. He has edited 2 books with Wiley, Eureka publications. He has been keynote Speaker and resource person of many workshops and webinars in India. He has been the reviewer for Elsevier, Springer, and IEEE Access. He has been Program Committee Member and Reviewer in the International Conference on Computational Intelligence and Emerging Power System ICCIPS 2021. He has recently presented 2 articles in Sixth International Conference on Advances in Computing & Communication Engineering Las Vegas USA ICACCE 2020 (22-24 June) IEEE EXPLORE Digital Library [SCOPUS] and 2 articles in the Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021 (4-6 Feb. 2021) IEEE EXPLORE Digital Library [SCOPUS]. His research area includes- Data Science, Data Mining, Machine Learning and Deep Learning. He also has Data Analytics Experience in Rapid Miner, Tableau, and WEKA. He has been working for more than 4 years in the field of data analytics, Java & Assembly Programming, Desktop Designing and Android Development.

Ashutosh Kumar Dubey

Dr. Ashutosh Kumar Dubey is currently in the department of Computer Science and Engineering, Chitkara University School of Engineering and Technology, Chitkara University, Himachal Pradesh, India. He received his PhD degree in Computer Science and Engineering from JK Lakshmipat University, Jaipur, Rajasthan, India. He is the Senior Member of IEEE and ACM. He has more than 14 years of teaching experience. He has authored a book name Database Management Concepts. He has been associated with many international and national conferences as the Technical Program Committee member. He is also associated as the Editor/Editorial Board Member/ Reviewer of many peer-reviewed journals. His research areas are Data Mining, Health Informatics, Optimization, Machine Learning, Cloud Computing, Artificial Intelligence and Object-Oriented Programming.

Sushil Narang

Dr. Sushil Narang is an Associate Professor in the Department of Computer Science and Engineering at Chitkara University, Rajpura, Punjab (India) since 2019. From 2006-2019, He was head of IT department at SAS Institute of IT & Research, Mohali, Punjab (India). From 1996-2006, He was Assistant Professor at Department of Computer Science & Applications, MLN College, Yamunanagar, Haryana (India). He Completed his Ph.D. at Panjab University, Chandigarh (India). His Research on “Feature Extraction and Neural Network Classifiers for Optical Character Recognition for Good quality handwritten Gurmukhi and Devnagari Characters” focused on various image processing, machine as well as deep learning algorithms. His research interests lie in the area of programming languages, ranging from theory to design to implementation, Image Processing, Data Analytics and Machine Learning. He has collaborated actively with researchers in several other disciplines of computer science, particularly Machine Learning on real world use cases. He is a certified Deep Learning Engineer from Edureka. He possesses expertise in Object-Oriented Analysis; Design and Development using Java and Python programming using OpenCV in Image Processing and Neural Network construction. He has strong knowledge of C++ and Java with experience in component architecture of product interface. With Solid training and management skills, He has demonstrated proficiency in leading and mentoring individuals to maximize levels of productivity, while forming cohesive team environments.

Real World Anomalous Scene Detection and Classification Using Multilayer Deep Neural Networks

Atif Jan[1]*, Gul Muhammad Khan[2]

[1] Department of Electrical Engineering, University of Engineering & Technology Peshawar (Pakistan)
2 National Center of Artificial Intelligence, University of Engineering & Technology Peshawar (Pakistan)

Received 30 May 2021 | Accepted 1 September 2021 | Published 31 October 2021

Abstract

Surveillance videos record malicious events in a locality utilizing various machine learning algorithms for detection. Deep-learning algorithms being the most prominent AI algorithms are data-hungry as well as computationally expensive. These algorithms perform better when trained over a diverse and huge set of examples. These modern AI methods have a dire need of utilizing human intelligence to pamper the problem in such a way as to reduce the ultimate effort in terms of computational cost. In this research work, a novel methodology termed Bag of Focus (BoF) based training methodology has been proposed. BoF is based on the concept of selecting motion-intensive blocks in a long video, for training different deep neural networks (DNN's). The methodology reduced the computational overhead by 90% (ten times) in comparison to when full-length videos are entertained. It has been observed that training networks using BoF are equally effective in terms of performance for the same network trained over the full-length dataset. In this research work, firstly, a fine-grained annotated dataset including instance and activity information has been developed for real-world volume crimes. Secondly, a BoF-based methodology has been introduced for effective training of the state-of-the-art 3D, and 2D Convolutional Neural Networks (CNNs). Lastly, a comparison between the state-of-the-art networks have been presented for malicious event recognition in videos. It has been observed that 2D CNN even with lesser parameters achieved a promising classification accuracy of 98.7% and Area under the curve (AUC) of 99.7%.

Keywords

Volume Crime Classification, Volume Crime Detection, Malicious Activity Detection, Deep Learning.

DOI: 10.9781/ijimai.2021.10.010

I. Introduction

NOWADAYS surveillance cameras are installed at decisive locations across the city. The surveillance videos can record various malicious activities in its locality. To reduce the impact of crimes recorded by CCTV cameras, timely detection of the activity is needed for prompt actions by the concerned authorities. The network of CCTV cameras is monitored through a central control room operated round-the clock by human observers. However, firstly a large expert task force is required to monitor these hundreds of video streams. Secondly, the probability of detecting anomalous activities decreases with an increase in the number of video streams and the time of attention. According to [1] an operator may efficiently monitor a video stream for about 12 minutes continuously, after which he may miss up to 45% of screen activity. After 22 minutes this miss-rate may even elevate to 95%. Thus, to improve the efficiency of surveillance systems, various technological solutions have been adopted to continuously analyze the CCTV recordings. The solutions are based on various machine learning algorithms. Deep-learning being the most effective earning technique is data-hungry and required huge computation. Currently, deep neural networks are trained over a large set of videos for understanding motion information. Training on a large set of full-length videos is a computationally expensive problem. So, a human intelligence-based solution is needed to design data for algorithms in such a way as to avoid redundant information in the full-length video, to reduce the computational overhead for training DNN's. Considering the importance of human intelligence-based methodology for training DNN's, we propose BoF methodology. Comprising of a dataset for the identification of volume crimes in public places for training learning systems has been introduced.

Various approaches have been entertained to develop a system for automatic detection of abnormal behaviors in CCTV recording. The initial studies in abnormal event detection were focused on object tracking [2], [3], [4], where a moving object is considered abnormal if its trajectory doesn't follow the fitted model during the training period. Trajectory analysis can perform well in the case of an individual moving object in a scene but is less effective for complex and crowded scenes. Such efforts are less effective in tracking the motion of abnormal shapes. Handcrafted feature extraction techniques are also exploited for anomaly detection [5], [6]. The fundamental problem with the mentioned approach is that the selection of efficacious features, which was resolved through deep features by Gong et al [7]. They have used unsupervised deep learning-based features for addressing anomaly detection. Usually, deviation from the normal is considered as an anomaly in unsupervised learning; however, this may not be the case due to the existence of a _ne line between normal and

* Corresponding author.
E-mail address: atifjan@uetpeshawar.edu.pk

abnormal behavior, which results in a large number of false alarms. The strongest approach used so far is supervised deep learning algorithms. In supervised deep learning techniques, various labeled datasets are used for the detection of a particular group of activities. The latest approach used by Sultani et al. is Multiple Instance Learning (MIL) for real-world anomaly detection. They introduced an assorted dataset of 13 real-world malicious activities. They managed to achieve a classification accuracy of 28%. The developed dataset contains a very diverse set of classes. Nevertheless, class labels are assigned to whole videos while only a part of these videos contained the occurrence of the actual event. This causes MIL to perform poorly leading to low classification accuracy. We in this research, acquired a subset of the above mentioned dataset to address volume crimes consisting of four frequently occurring abnormalities (assault, fight, shooting, vandalism). Some of these examples are illustrated in Fig. 1.

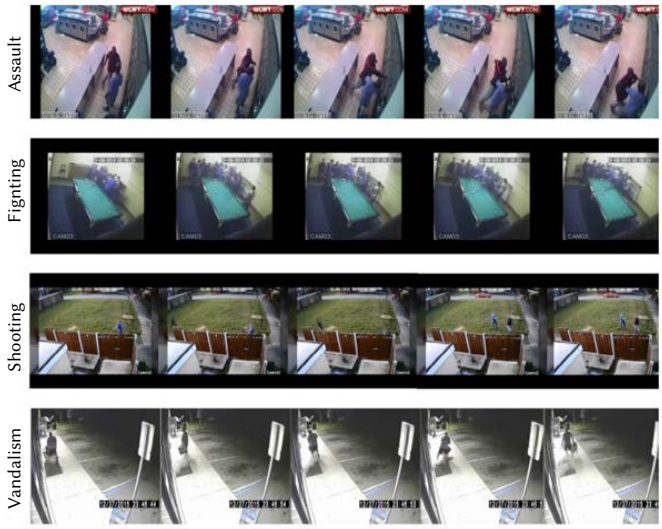


Fig. 1. Sequence of frames taken from video stream can related to different malicious events e.g. Assault, Fighting, Shooting, and Vandalism.

State-of-the-art models implemented for understanding motion information in videos adopt the methodology of training network on full-length videos. The conventional methodology trains the networks on redundant data having almost similar motion information and thus creates a computational overhead during the training process. The work presented here provides a mechanism where a properly labeled dataset is developed in which each video instance is labeled based on the type of activity present in it to make it suitable for supervised learning methodology. Furthermore, a BoF-based methodology for training networks has been introduced. The proposed methodology reduces the computational overhead for training a deep-learning network for understanding motion information in videos, without affecting the overall accuracy of the network. In addition to properly arranging the training datasets, the most effective concept of intermediate frame fusion and early frame fusion for Spatio-temporal analysis of videos has been adopted and deployed using state-of-the-art 2D, and 3D CNN. This research work has three tier contributions including:

1. Development of dataset for anomalous activity detection.
2. Development of BoF based methodology for training networks.
3. Validation of the proposed methodology for training deep-learning networks over state-of-the-art 3D-CNN, and 2D-CNN based networks developed for understanding motion information in videos.

The remainder of this paper is organized as follows: Section II reviews the related work in anomalous activity detection. Section III presents the architectures of the proposed models including problem formulation and its implementation details. Section IV explains the experimental setup including dataset while section V provides the results and analysis. Section VI concludes the paper.

II. Related Work

In this section, we will be discussing various approaches entertained in past for malicious activity detection using Spatio-temporal CNNs. Generally, malicious activities are anomalous parts of the video sequence; however, it could be observed in the literature that maliciousness of the activity in the given video sequence has been contextually specified according to the target application. According to [4], the definition of malicious activities changes with the scenario. It is a complicated task to differentiate the malicious event from the rest of the video recording at the given instance. Some of the malicious events identified in CCTV footages have been shown in Fig. 1. The following subsections present the overview of various approaches used for the detection of context-sensitive anomalies and established anomalies.

A. Context Sensitive Anomalies

Object tracking as abnormal motion in restricted areas, running, and loitering are often considered as an anomaly in a sensitive environment. Such type of movements in a video sequence are detected by various trajectory analysis techniques [8], [4]. Calderara et al. considered inter node transition pattern in a graph as trajectory for abnormal motion detection [2] while Morris et al. found the interesting node using Gaussian Mixture Model and then Hidden Markov Model for the same purpose [9]. Moreover, techniques based on low level local features have been used for detection of abnormal motion [10], [11], [6]. Ermis et al. constructed a probabilistic model for abnormal motion detection by generating behavior cluster derived from behavior profile [12]. Reddy et al. exploited ground truth segmentation in combination with the motion and size feature modeled by kernel density estimation [11]. This approach claims its effectiveness in detecting abnormal object in crowded scenes. Xiao et al. used hybrid combination of sparse semi non-generative matrix factorization (SSMF) and histogram of non-negative coefficient (HNC) for anomaly detection in surveillance videos [6]. In their work only normal data is used for parameter tuning and deviation from normal motion is considered as an anomaly. Li et al. [13] proposed a Spatio-temporal model for anomaly detection in complex and crowded scene. Dynamic texture model in combination is used for considering both dynamics and appearance information. In proposed model spatial saliency score is computed using a center-surround discriminant. Whereas, temporal saliency score is produced using a model of normal behavior learned from data. Although, all of these techniques performed well for abnormal motion detection such as running in a scene and walking in the wrong direction, however, they are specifically designed for tracking objects in image sequence.

B. Established Anomalies

Anomaly detection remained in focus for the last decade to detect the abnormal human behavior in surveillance videos. A group of researchers working in the area of ensuring safety of pedestrian walkways focused on the detection of non-pedestrian entities in public walkways [13], [19], [22]. Li et al. [13] proposed the Real Time Volume Crime Detection and Classification using Deep Learning. [14] proposed the use of Spatio-temporal information Center-surround discriminant saliency detector and normal behavior model for extracting spatial and temporal saliency score respectively, for the categorization of pedestrian abnormalities. Tahboub et al. [14] used local binary pattern in combination with random forest for detecting pedestrian anomalies. Ravan et al. [3] used generative adversarial

TABLE I. Existing Datasets for Malicious Activity Detection

Name of Dataset	Total Videos	Environmental Conditions	Identified Anomalous Activities
UCSD PED1	170	Outdoor	Non-pedestrian entities, and walking across walkways
UCSD PED2	28	Outdoor	Non-pedestrian entities, and walking across walkways
Subway Entrance	1	Indoor	Wrong direction, no payment, and loitering
Subway Exit	1	Indoor	Wrong direction, no payment, and loitering
Avenue	37	Outdoor	Running, and Throwing object
UMN	5	Outdoor, Indoor	Running
Hockey Fight	1000	Playground	Fighting
The Movies	200	Outdoor, Indoor	Fighting
UCF Crimes	1900	Outdoor, Indoor	Abuse, Accident, Arrest, Burglary Explosion, and Fighting

network for learning normal pattern of public walkways and deviation from the learned pattern is considered as an abnormality. Ameer et al. [15] proposed combination of connected component analysis, histogram of oriented gradient and Gaussian mixture model for non-pedestrian object detection. Khan et al. [7] used Gaussian discriminant model in combination with k-means clustering for classification of events recorded in surveillance videos installed in pedestrian walkways. Various applications of the anomaly detection have been previously validated with certain benchmark dataset listed in Table I. These include some of the prominent datasets of UCSD (PED1, PED2), Avenue, Subway Entrance, Subway Exit, and UMN. Most of these datasets have been developed for identification of a particular set of anomalies in specific environments e.g., UCSD is used to identify non-pedestrian entities and walking across walkways, Subway dataset is for identifying walking in wrong direction, non-payments, and loitering, while Avenue, UMN, Hockey Fight, and The Movies datasets have been used for identification of single activity mostly fighting. Moreover, these datasets have been developed with the help of actors and does not provide any real-life situation in any of their videos [16]. We can conclude that the approaches discussed in the previous section mainly targeted contextual anomalies tested over fabricated datasets in which anomalous scenes are acted by the actors. Considering the importance of detecting real-world anomalies recorded in real-time CCTV footage to assist security agencies. Sultani [16] introduced UCF crimes dataset incorporating 13 real-world anomalies and proposed multiple instance learning (MIL) for the classification of abnormal activities. The dataset is developed by downloading videos of CCTV recordings from live leaks and youtube. Each video is labeled according to the type of anomaly recorded in that video.

According to Sultani the developed dataset is weekly annotated. The proposed algorithm achieved a classification accuracy of 28%, thus, a state-of-the-art solution for detecting real-world abnormality in CCTV recording is still a dream.

C. Spatio-temporal Analysis

Spatio-temporal analysis is usually desired for identification of a function between spatial and temporal data to affect the performance of any process. While it defines a relation between GPS coordinates and its time instance for activity recognition in [17] and a relation between location and time of the day for prediction of criminal activity in [18], it links the spatial information of each frame of a video sequence to its temporal distribution in [19]. Extraction of useful information from a video sequence relies not only on the visual information spread spatially in each static frame but also on the complex motion information distributed along the continuous sequence of frames. Previously, hand-crafted features are used for obtaining appearance and motion information from video streams [20], [14]. However, these hand-crafted features contain less discriminant information and recent deep supervised and unsupervised features are employed for different applications [16]. More advanced, Spatio-temporal convolutional neural networks (CNN) [21] are introduced to learn appearance as well as motion information in video stream [19] which previously gained popularity in area of action recognition [22], [23] and hand gesture recognition [24]. Inspired from the popularity of Spatio-temporal CNN's, we in this research propose the use of a similar CNN modified for malicious activity detection.

In this research, we have developed a dataset specifically to identify malicious events in a video stream. For this purpose, videos are taken from UCF crimes dataset and are annotated for instance recognition task. Followed by a BoF based training mechanism to reduce the computational overhead. The proposed methodology is then verified by training state of the art Spatio-temporal CNN's developed for understanding motion information in videos.

III. Proposed Framework

Instance recognition in videos demands analysis of the information spread across spatial and temporal domains of video sequence. We can acquire semantic information of the scene from spatially distributed objects in a single frame, while sequence of such consecutive frames provides the positional changes of objects, hence enabling us to understand the overall activity in the video stream. To perform this task through CNNs, we have designed a framework that takes in samples of the video stream and outputs the activity performed in each sample. The overall architecture has been divided into the modular phases of video processor, feature extraction, and instance recognition. The overall architecture is presented in Fig. 2.

A. Video Pre-Processing

Videos consist of a sequence of stationary image frames. For the interpretation of useful information through convolutional neural network from these videos, all video frames are processed for understanding motion information in full length video. So, training data on full length video creates a huge computational overhead. To train CNN's for understanding motion information in videos, we have designed a framework for training CNN's over a defined set of frames i.e. BoF containing key information required for understanding motion in particular videos. The overall proposed framework is divided into modular phase of BoF extraction, block formation, block selection, and down framing.

BoF Extraction The set of frames containing the activity in full length video recording has been termed as BoF. Initially BoF has been determined in each video. The process of BoF extraction is given in Fig. 3. The figure shows that only a portion of the full-length video labeled as assault contains the actual activity. So, 128 out of 294 frames have been considered as a BoF. The same process has been repeated for all videos in the dataset and thus reduced the training data by removing the unwanted information for understanding. A comparison of total frames in dataset and the number of frames in the set of BoF's has been mentioned in Table II.

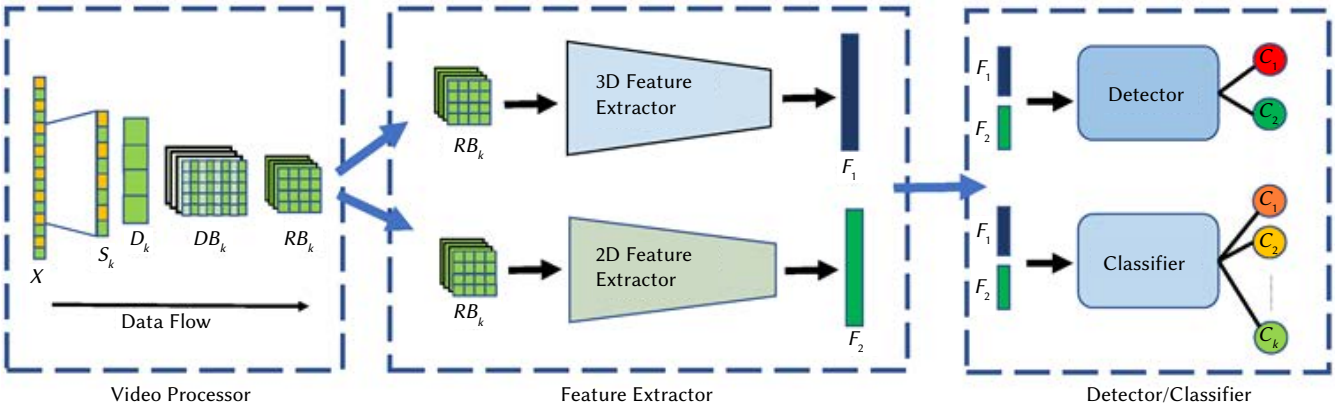


Fig. 2. Overall architectural diagram of the framework developed for malicious instance recognition.

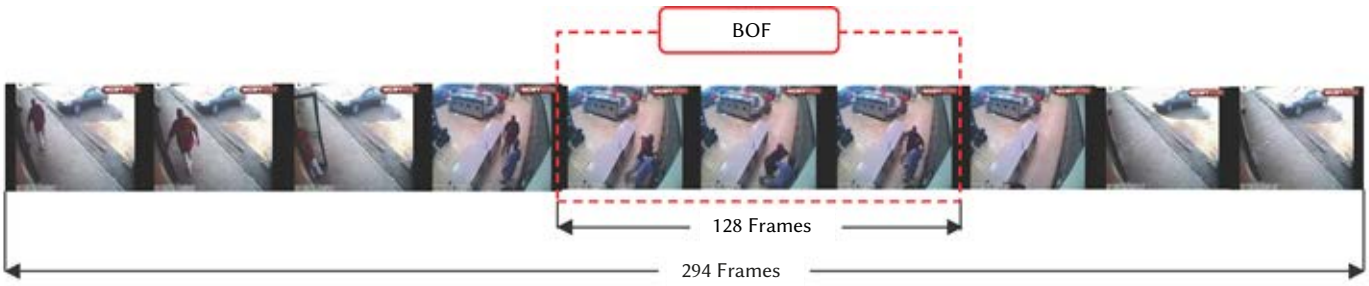


Fig. 3. BOF Extraction.

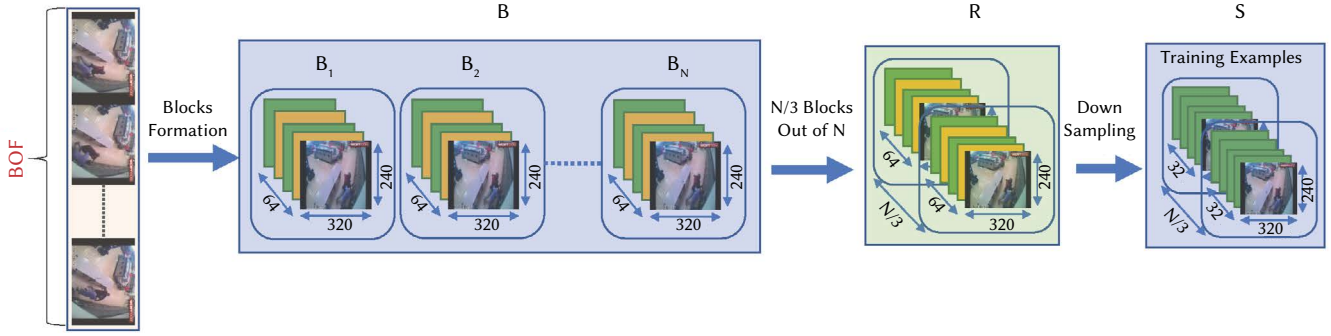


Fig. 4. The Process of block formation, block selection, and down sampling is illustrated in the figure.

TABLE II. Comparison of Frames in the Original Dataset and in the Reduced Dataset

Activity	Total Frames	Frames in BOF	Frames in Training set
Assault	129,284	54,912	9,152
Fighting	258,144	124,608	20,768
Shooting	146,624	56,768	9,472
Vandalism	146,400	84,352	14,059
Total	564,152	265,728	53,451

Block Formation The BoF is then divided into blocks of 64 frames each. The block of 64 frames covers a video length of almost 2sec. the process of block formation is mentioned in the first section of Fig. 4. The whole process of block formation is expressed mathematically in Equation (1). Considering the sequence of discrete frames from BoF expressed as X[n] we can present the m*th* block B*m* as:

$$Bm = X[n](u(n - 64m) - u(n - 64(m + 1))) \quad (1)$$

where u(n) represents the unit step function.

Block Selection and Down Sampling The set of blocks B obtained in previous section consists of total N blocks. 1/3rd of the total blocks is randomly selected for training the CNN's. Let the set of randomly selected blocks is represented by R this step is based on the fact that consecutive blocks will contain almost same motion information. Each block of set R is then down sampled to block *s* by removing alternating frame to avoid redundant information in consecutive frames. The process of down framing has been presented in Fig. 4. The overall process of block formation could be expressed mathematically by Equation (2) whereas $\delta(.)$ represents the unit impulse function. Table II shows that only 10% of the total frames have been selected for training the networks. Thus 90% of the redundant data has been avoided during training process.

$$S[n] = \sum_{k=0}^{31} R[k]\delta[n - 2k] \quad (2)$$

B. Feature Extraction

Deep CNNs have been extensively used for feature extraction in various image domains. They proved to extract much more representative features from images compare to previous hand-crafted approaches that relied mostly on local features in images. To validate

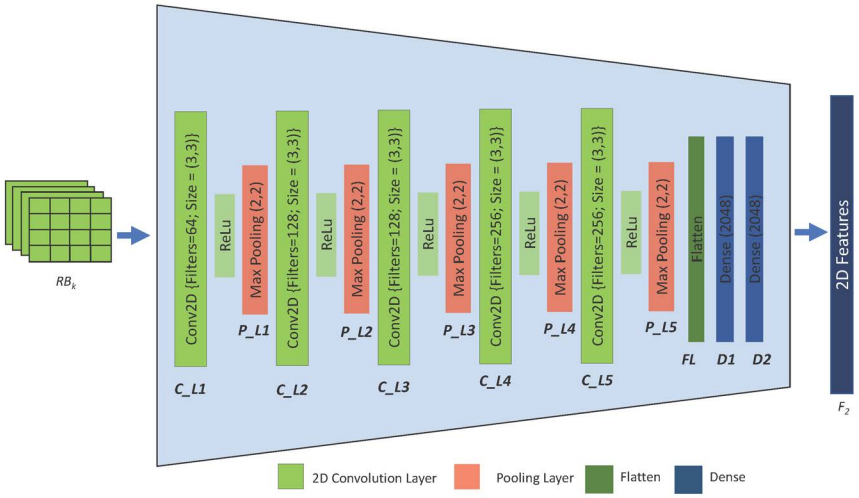


Fig. 5. 2D Convolutional Neural Network.

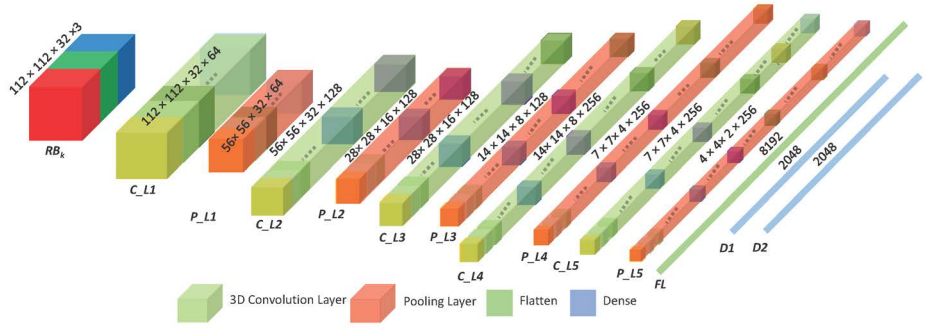


Fig. 6. 3D Convolutional Neural Network.

the concept of BoF based training mechanism, we have explored two types of deep learning architectures for extracting features from the given block of video stream. Both of these networks have been developed in such a way that take the block of video and extract a single dimensional feature.

3D CNN based Feature Extractor We believe motion information of the objects to be equally important for instance recognition in addition to the spatial distribution of objects in a given frame. For this purpose, we developed a CNN network comprising of 3D convolution layers that could learn spatial as well as temporal features from the given block of the video sequence. The proposed model is obtained by removing a few convolution layers from the standard C3D network to reduce the network complexity. Our 3D CNN feature extractor uses 5-tiered 3D convolution layers followed by 2 fully connected layers to learn a single-dimensional feature vector. Each 3D convolution layer is followed by a Max-pooling layer with stride 2×2×2 to transform the object and motion information from spatial and temporal dimension to depth. This transformation leaves us with a frame size of 4×4×2×256. Recently, 3D-CNN gained popularity in the area of action recognition [6], [23], [17]. Inspired by the performance of 3D-CNN in the field of action recognition, we developed the model in Fig. 5.

$$Y[x,y,z] = \sum_{i=0}^{L-1}\sum_{j=0}^{M-1}\sum_{k=0}^{N-1} h[i,j,k]S[x-i,y-j,z-k] \quad (3)$$

2D CNN based Feature Extractor Network parameters of proposed system are 24,109,437. In order to reduce the network parameter, the dimensions of input block are reduced from (112×1×12×3×32) to (112×112×32) by converting each frame to grayscale. Conversion to grayscale does not affect the system performance in case of activity detection due to the fact that activity detection procedure is not sensitive to the color tone in video frames. The gray-scale block is then fed to 2D CNN of same number of convolution layers and dense layers. Thus, reduces the network parameters to 13,722,437. Block diagram of the proposed system based on 2D-CNN is given in Fig. 6. In the figure, the last layer represents the feature vector. The feature vector is then fed to two fully connected layers and an output layer. The number of neurons in fully connected layers and dense layer are same as that in 3D-CNN. For instance, S is the block of 32 gray-scale frames of a video, S_k is the kth frame of block S, h is the 2D filter of dimension L _M. The mathematical process for considering temporal as well as special information is shown in Equation (4).

$$Y[x,y] = \sum_{i=0}^{L-1}\sum_{j=0}^{M-1}\sum_{k=0}^{N-1} h[i,j]S_k[x-i,y-j] \quad (4)$$

C. Instance Recognition

The objective of the research is to detect the combination of frames in a video stream with one of the categories of the volume crimes mentioned earlier on. For this purpose, features of the block acquired in section III.B are classified through various classification algorithms including Gaussian Naive Bays, Decision Tree, Support Vector Machine (SVM), k-Nearest Neighbor, and Softmax. All these classifiers are developed in such a way to address both instance detection (binary classification) and instance classification (multi-class classification). Among all these, softmax classifier has been designed with softmax activation optimizing the binary cross entropy loss and categorical cross entropy loss with detection and classification, respectively. Both of these losses are mathematically presented as:

$$L_b = -\frac{1}{N}\sum_i^N t_i\, log(p_i) + (1-t_i)\, log(1-p_i) \quad (5)$$

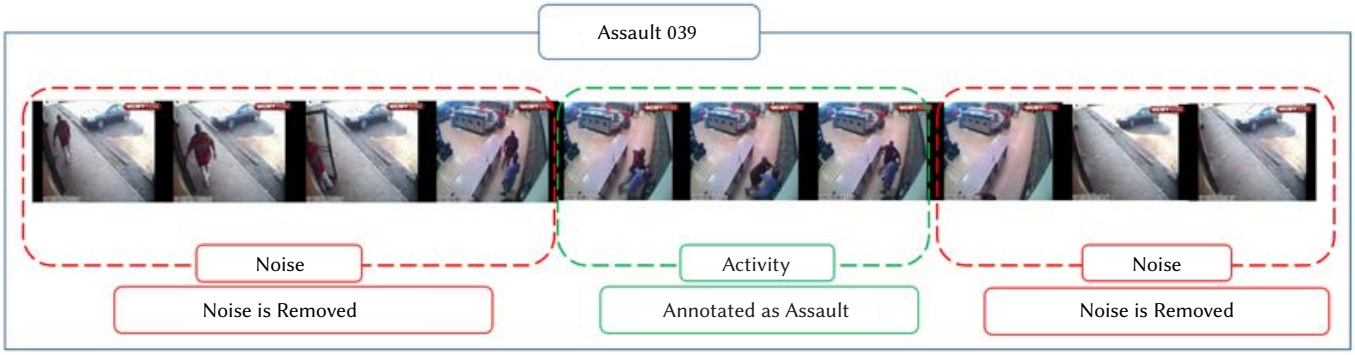


Fig. 7. Annotation of video sample labeled as Assault in original dataset.

$$L_C = -\frac{1}{N}\sum_{i=1}^{N}\sum_{i}^{C} t_{ij}\, log(p_{ij}) \quad (6)$$

where N represent total number of training examples (blocks), t is target value, p is predicted value, and C denotes the number of classes for multi-class-classification.

IV. Experimental Setup

We have developed a unified framework for the tasks of detection and classification. For the case of instance detection, the block is identified as a normal or anomalous event. This is like the concept introduced in [28] which considered everything that doesn't look normal as anomaly. In classification, on the other hand the specific type of activity associated with each of volume crime is identified. Technically, detection performs a binary classification task (0; 1) while in classification we perform a multi-class classification task (0; 1; 2; 3; 4) with the same framework given in Fig. 2. Both tasks have been validated through the dataset developed specially for malicious instance recognition.

A. Dataset

Apart from the methodology for effective training of CNN's, this research work is also focused on the anomaly detection in safe-city environment, this is why the subset of a very recent dataset (UCF crimes) developed for real-world anomaly detection in surveillance videos has been considered. The dataset consists of CCTV footages of real-world anomalies from Liveleaks, and Youtube including 13 real-world anomalies containing 1900 videos spanned over 128 hours. For this research, a subset of four most crucial anomalies (shooting, assault, fighting, and vandalism) are annotated for in-video event recognition. This is carried out by specifically separating normal frames from the ones that contain anomalous activity. The process has been demonstrated in Fig. 7. Previously, it was very difficult to use the video labeled as assault, it was observed that videos labeled as (Assault 039) contain (43:8%) frames belonging to normal activity and the rest belonging to the assault. Each video in our dataset consists of frame-level labels for its class annotated by three skilled annotators through visually inspecting each video stream.

B. Experimentation

We have conducted experiments for detection and classification using 3D-CNN and 2D-CNN features with various classifiers explained in section III.3. Hyper parameters setting for 2D and 3D CNNs are listed in Table III. We have performed our experiments on Intel Core-i5 with 8Gb RAM and Nvidia GTX 1050Ti Graphical Processing Unit. In each experiment Stochastic Gradient Descent optimizer is used for learning weights.

TABLE III. Hyper Parameter Choice for 2D-CNN and 3D-CNN

Parameters	2D-CNN	3D-CNN
No. of Epochs	140	70
Initial Learning rate	0.001	0.001
Momentum	0.9	0.9
Kernel Size	(3,3)	(3,3,3)
Pooling window	(2,2)	(2,2,2)

V. Results and Analysis

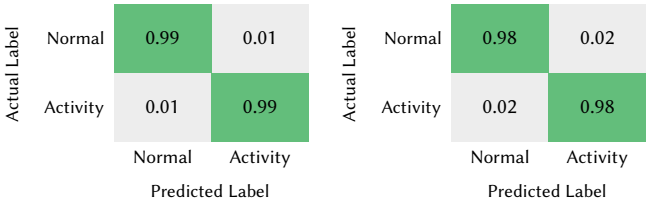
We have conducted numerous experiments for detection and classification using features extracted from 2D and 3D-CNNs. The performance of the mentioned model is evaluated based on the performance metric like AUC, accuracy, and false-positive rate. This section provides a performance comparison of our proposed system. Table IV summarizes the results.

A. Anomalous Event Detection

From the results on event detection using different features, it was observed that 2D CNN outperforms 3D CNN achieving the overall detection accuracy of 99:0%. Although, all the classifiers equally performed well for both type of features, however, Softmax classifier outperforms the rest in detecting malicious video events as shown in Table IV. Similar patterns have been observed in confusion matrices and t-SNE plots of the detection process as shown in Fig. 8 and 11 respectively.

TABLE IV. Comparison of the Features From 2D CNN and 3D CNN With Other Classifiers

Classifier	2D CNN Detection	3D CNN Detection	2D CNN Classification	3D CNN Classification
Gaussian Naïve Bays	98.80	98.67	88.83	88.72
Decision Tree	98.88	98.65	74.41	74.41
Support Vector Machine	97.80	98.57	74.41	74.41
K Nearest Neighbor	98.86	**98.69**	82.15	81.80
Softmax	**99.00**	98.10	**98.80**	**97.80**



(a) 2D-CNN (b) 3D-CNN

Fig. 8. Confusion Matrix for 2D features and 3D features using Softmax classifier.

B. Anomalous Event Classification

For classification of the event among one of assault, fighting, shooting, vandalism, and normal, the features from 2D and 3D CNN have been extracted in similar manner and evaluated with various classifiers. It is observed that the performance of Softmax classifier combined with 2D features is much better in comparison to the rest of classifiers. Overall classification accuracy of 98:89% has been achieved for this specific task. Even though, 3D features also performed well in classification; however, the number of network parameters in 2D CNN are much less as compared to 3D-CNN.

C. Extended Analysis on Anomalous Event Classification

We also presented the results in terms of detail performance metrics including Precision, Recall, and F1-Score for each class of anomalous events. Mathematical expressions for Precision, Recall, and F1-Score are given in equations (7-9), respectively. Upon observation of table VI, we concluded that 2D-CNN performs better for each class in comparison to the 3D-CNN. A similar phenomenon could be observed in the recall score for the shooting as 0:846 and 0:916 for 3D-CNN and 2D-CNN, respectively. Fig. 10 and 9 show the confusion matrices and t-SNE plots for event classification using 2D and 3D features with Softmax classifier. It should be noted that visibly separable clusters could be seen in t-SNE plots which validates the accuracy achieved for the given task of classification.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{7}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{8}$$

$$F1 = \frac{2 \times Precission \times Recall}{Precision + Recall} \tag{9}$$

D. Comparison With the State-of-the-art

We have also compared our approaches with the state-of-the-art techniques using 14 UCF crimes dataset. It is observed from Table V

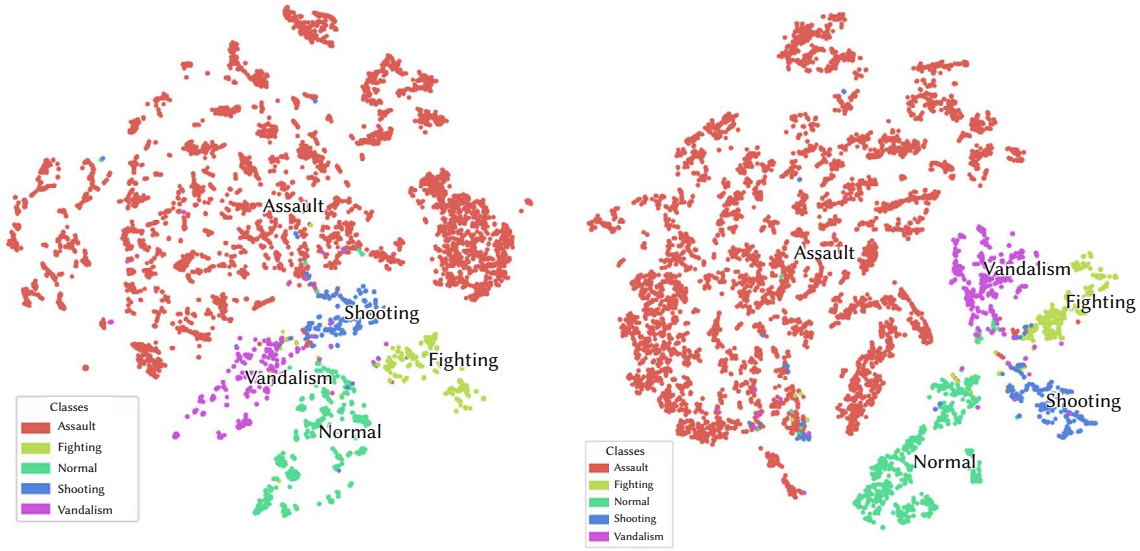
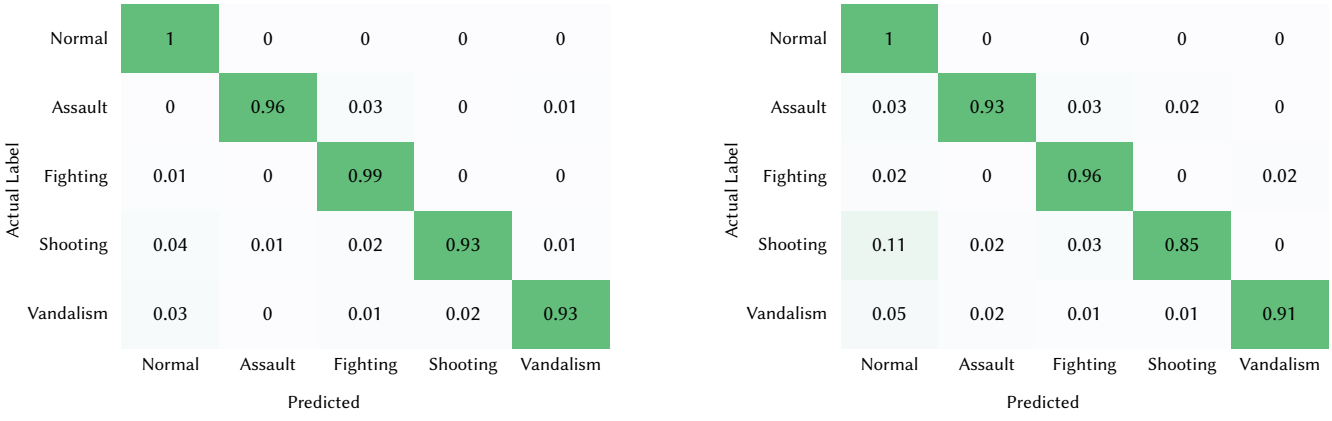


Fig. 9. t-SNE plots for 2D features and 3D features using Softmax classifier.



(a) 2D-CNN (b) 3D-CNN

Fig. 10. Confusion Matrix for Classification Task for 2D features and 3D features using Softmax classifier.

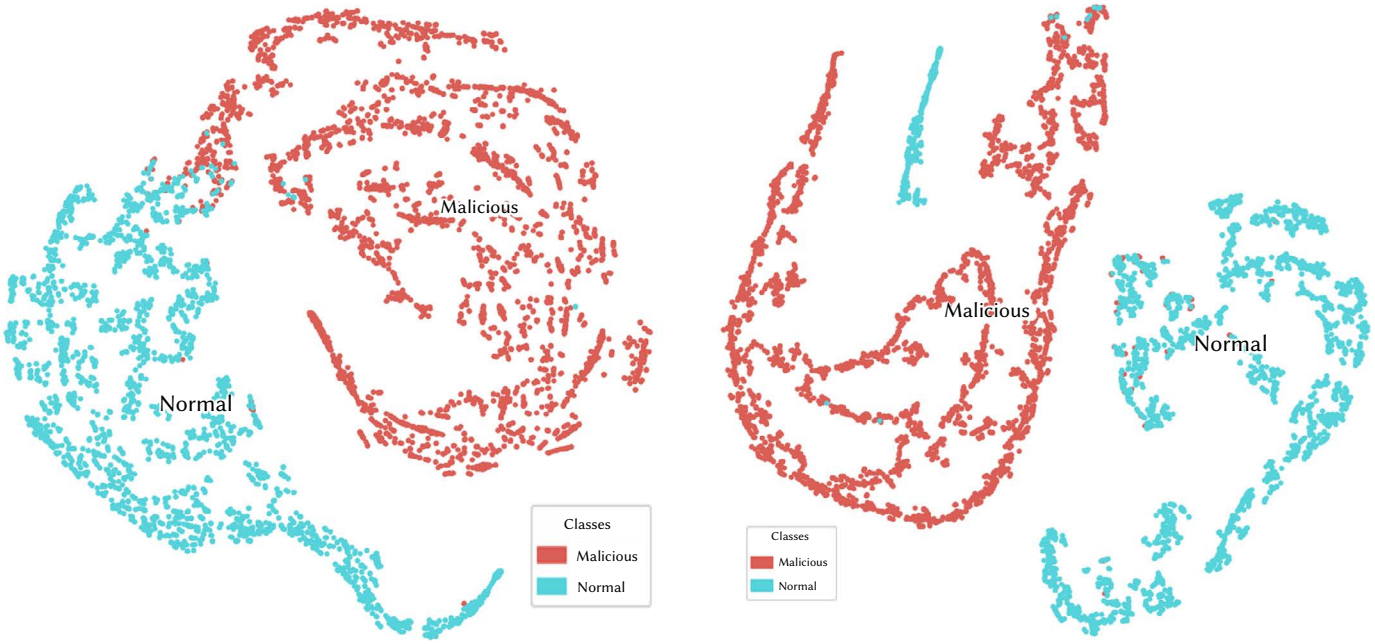


Fig. 11. t-SNE plots for Classification Task for 2D features and 3D features using Softmax classifier.

TABLE V. Comparison of Our Approach to the State-of-the-art Techniques Using the Dataset Proposed by [21]

Authors	Technique	Accuracy	AUC	False Positive
Sultani *et al.* [16]	MIL	28.4	75.41	1.9
Khan *et al* [7]	GDA	-	64.30	-
Ours	2D CNN	**98.8**	**99.7**	1.2
Ours	3D CNN	97.4	99.5	**0.7**

TABLE VI. Extended Analysis on Anomalous Event Classification

Technique	Activity	Precision	Recall	F1 Score
ResNet3D [29]	Normal	0.997	0.990	0.993
	Assault	0.993	0.944	0.968
	Fighting	0.955	0.998	0.976
	Shooting	0.941	0.971	0.955
	Vandalism	0.994	0.96	0.976
(2+1) D [26]	Normal	0.996	0.995	0.995
	Assault	0.996	0.944	0.970
	Fighting	0.984	0.994	0.989
	Shooting	0.911	0.988	0.948
	Vandalism	0.994	0.965	0.979
P3D [30]	Normal	0.993	0.996	0.995
	Assault	0.967	0.965	0.966
	Fighting	0.994	0.986	0.990
	Shooting	0.953	0.956	0.955
	Vandalism	0.990	0.986	0.988
3D-CNN	Normal	0.985	0.996	0.991
	Assault	0.933	0.929	0.931
	Fighting	0.968	0.963	0.965
	Shooting	0.962	0.846	0.900
	Vandalism	0.966	0.912	0.938
2D-CNN	Normal	0.993	0.997	0.995
	Assault	0.986	0.965	0.975
	Fighting	0.985	0.990	0.988
	Shooting	0.955	0.916	0.935
	Vandalism	0.952	0.943	0.948

that our approach is performing far better on 5 classes as compared to the rest of the techniques using MIL and GDA techniques. Moreover, our approach achieves the least false positive rate of 0:7% for our 3D CNN. It is observed that the proposed 2D model outperforms the performance of 3D model in case of the activity detection and manage to achieve false positive rate of 0.1. Furthermore, the model is suitable for real time application due to its low false positive rate and high frame processing rate that is 1000 frames/sec. Also, the proposed methodology for training CNN's is further verified by training three (ResNet3D [29], (2+1)D [26], and P3D [30]) well known algorithms for video understanding. Validation results of the trained model on the data samples removed from the dataset following the proposed methodology discussed in section III.A are mentioned in Table VI. It has been observed that training the state-of-state-of-the-art algorithm by following the methodology mentioned in section III.A performed in its order of popularity. Hence, it is verified that even training on 10% of video frames are enough for understanding motion information in video, and thus reduces the computation overhead during training process by 90% irrespective of the network used.

VI. Conclusion and Future Work

Lack of implementable software solutions for the identification of real-world malicious activities from video streams in a safe city environment requires a blend of computer vision and machine learning algorithms. In this regard, a training mechanism has been introduced to reduce the computation required for training the learning algorithms. The proposed training methodology achieved a promising accuracy even reducing the computational overhead by 90%. Also, an optimal solution for the analysis of temporal frames extracted from CCTV recordings is proposed. Our proposed models managed to achieve high accuracy for not only the identification of malicious events but also classification of real-world volume crimes including assault, fighting, shooting, and vandalism in a video sequence. Furthermore, our models are also suitable for real-time applications due to their high frame processing rate and low false alarm rate, with high classification accuracy of 98.7% and AUC of 99.7% on four classes.

The system can further be modified for other classes of crimes including but not limited to burglary, riots, attempted murder, arson, explosion, robbery, theft, and arrest etc. In order to get a unified framework for the detection of multiple malicious activities recorded by a CCTV camera, we need to train the same system with the data for the above-mentioned events.

Acknowledgment

Indeed, it was a tough journey to compile all this work. At many instances, I felt that I would not be able to do this. But thanks to my team at National Center of Artificial Intelligence who helped, supported, and motivated me on every step.

References

[1] T. Ainsworth, "Buyer beware," *Security Oz*, vol. 19, pp. 18–26, 2002.

[2] S. Calderara, U. Heinemann, A. Prati, R. Cucchiara, and N. Tishby, "Detecting anomalies in people's trajectories using spectral graph analysis," *Computer Vision and Image Understanding.*, vol. 115, no. 8, pp. 1099–1111, 2011.

[3] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, "Abnormal event detection in videos using generative adversarial nets," in 2017 *IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 1577–1581.

[4] S. Wu, B. E. Moore, and M. Shah, "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes," in 2010 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2054–2060.

[5] T. Wang and H. Snoussi, "Detection of abnormal visual events via global optical flow orientation histogram," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 6, pp. 988–998, 2014.

[6] T. Xiao, C. Zhang, and H. Zha, "Learning to detect anomalies in surveillance video," *IEEE Signal Processing Letters*, vol. 22, no. 9, pp. 1477–1481, 2015.

[7] M. U. K. Khan, H.-S. Park, and C.-M. Kyung, "Rejecting motion outliers for efficient crowd anomaly detection," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 2, pp. 541–556, 2018.

[8] F. Jiang, J. Yuan, S. A. Tsaftaris, and A. K. Katsaggelos, "Anomalous video event detection using spatiotemporal context," *Computer Vision and Image Understanding*, vol. 115, no. 3, pp. 323–333, 2011.

[9] B. T. Morris and M. M. Trivedi, "Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 11, pp. 2287–2301, 2011.

[10] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *International journal of computer vision*, vol. 74, no. 1, pp. 17–31, 2007.

[11] V. Reddy, C. Sanderson, and B. C. Lovell, "Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture," in *CVPR 2011 WORKSHOPS*, 2011, pp. 55–61.

[12] E. B. Ermis, V. Saligrama, P.-M. Jodoin, and J. Konrad, "Motion segmentation and abnormal behavior detection via behavior clustering," in *2008 15th IEEE International Conference on Image Processing*, 2008, pp. 769–772.

[13] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 1, pp. 18–32, 2013.

[14] K. Tahboub, A. R. Reibman, and E. J. Delp, "Accuracy prediction for pedestrian detection," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 4192–4196.

[15] S. Amraee, A. Vafaei, K. Jamshidi, and P. Adibi, "Anomaly detection and localization in crowded scenes using connected component analysis," *Multimedia Tools and Applications*, vol. 77, no. 12, pp. 14767–14782, 2018.

[16] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6479–6488.

[17] J. Wang, J. Jiao, L. Bao, S. He, Y. Liu, and W. Liu, "Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4006–4015.

[18] L. Duan, T. Hu, E. Cheng, J. Zhu, and C. Gao, "Deep convolutional neural networks for spatiotemporal crime prediction," in *Proceedings of the International Conference on Information and Knowledge Engineering (IKE)*, 2017, pp. 61–67.

[19] S. Zhou, W. Shen, D. Zeng, M. Fang, Y. Wei, and Z. Zhang, "Spatial--temporal convolutional neural networks for anomaly detection and localization in crowded scenes," *Signal Processing: Image Communication*, vol. 47, pp. 358–368, 2016.

[20] V. Kaltsa, A. Briassouli, I. Kompatsiaris, L. J. Hadjileontiadis, and M. G. Strintzis, "Swarm intelligence for detecting interesting events in crowded environments," *IEEE transactions on image processing*, vol. 24, no. 7, pp. 2153–2166, 2015.

[21] M. Khari, A. K. Garg, R. Gonzalez-Crespo, and E. Verdú, "Gesture Recognition of RGB and RGB-D Static Images Using Convolutional Neural Networks," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 7, p. 22, 2019.

[22] T. Lima, B. Fernandes, and P. Barros, "Human action recognition with 3D convolutional neural network," in *2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, 2017, pp. 1–6.

[23] J. D.Pujari, R. Yakkundimath, and A. S. Byadgi, "SVM and ANN Based Classification of Plant Diseases Using Feature Reduction Technique," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 3, no. 7, p. 6, 2016.

[24] N. L. Hakim *et al.*, "Dynamic Hand Gesture Recognition Using 3DCNN and LSTM with FSM Context-Aware Model," *Sensors*, vol. 19, no. 24, p. 5429, 2019.

[25] F. Cronje, "Human action recognition with 3D convolutional neural

networks," University of Cape Town, 2015.
[26] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun and M. Paluri, "A Closer Look at Spatiotemporal Convolutions for Action Recognition," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450-6459
[27] Z. Liu, C. Zhang, and Y. Tian, "3D-based deep convolutional neural network for action recognition with depth sequences," *Image and Vision Computing*, vol. 55, pp. 93–100, 2016.
[28] D. Gong *et al.*, "Memorizing Normality to Detect Anomaly: Memory-augmented Deep Autoencoder for Unsupervised Anomaly Detection," *arXiv Prepr. arXiv1904.02639*, 2019.
[29] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-Temporal features with 3D residual networks for action recognition," *Proc. - 2017 IEEE International Conference on Computer Vision Workshops. ICCVW 2017*, vol. 2018-January, pp. 3154–3160, 2017.
[30] J. Chen, J. Hsiao, and C. M. Ho, "Residual Frames with Efficient Pseudo-3D CNN for Human Action Recognition," *arXiv preprint* pp. 5–9, 2020.

Atif Jan

Atif Jan obtained his B.Sc. degree in Electrical Engineering from University of Engineering and Technology (UET), Peshawar in 2011 and his master's degree in Electrical Engineering from UET, Peshawar in 2015. He is also pursuing his Ph.D. Currently; he is working as a Lecturer at Department of Electrical Engineering, UET Peshawar. His research interests include image processing, computer vision, machine learning and deep learning.

Dr. Gul Muhammad Khan

Dr. Gul Muhammad Khan is blessed with the capability to do things and face challenges. He graduated from UET Peshawar with distinction. He completed his PhD in intelligent System design from University of York, UK. During his PhD, he devised a brain inspired "learning to learn "system that has the capability of learning for itself at run-time. He is the pioneer of Cartesian Genetic Programming evolved Developmental network (CGPDN) providing an indirect method of decoding for neural programs. He is amongst the top neuro-developmental scientists around the world. He has also introduced a new concept for neuro-evolution and presented new algorithms for Automatic generation of feedforward (CGPANN)/ feedback (CGPRNN), markovian and non-markovian non-linear systems. He has introduced the concept of plastic networks and presented algorithms for both feed forward and feedback system, termed as Plastic CGPANN and Plastic RCGPANN. He joined UET as Assistant Professor on Tenure Track System (TTS) in Aug 2008 after completion of his PhD. In March 2012, he established a Research facility, Centre for Intelligent Systems and Networks Research (CISNR) at UET Peshawar and was appointed as Director of the Centre, still serving. In a short period of three years, he obtained research funding of more than 50 million, completed five research projects, with three in progress, supervising over 25 PhD/MPhil students and Researchers. He won the Research Productivity award for the year 2017. He is the sole author of a book published by springer titled: Evolution of Artificial Neural Development.

RIADA: A Machine-Learning Based Infrastructure for Recognising the Emotions of *Spotify* Songs

P. Álvarez, J. García de Quirós, S. Baldassarri *

Computer Science and Systems Engineering Department. María de Luna, 1, Ada Byron Building, Zaragoza, University of Zaragoza (Spain)

Received 28 January 2021 | Accepted 21 February 2022 | Published 21 April 2022

Abstract

The music emotions can help to improve the personalization of services and contents offered by music streaming providers. Many research works based on the use of machine learning techniques have addressed the problem of recognising the music emotions during the last years. Nevertheless, the results obtained are only applied on small-size music repositories and do not consider what the users feel when they listen to the songs. These issues prevent the existing proposals to be integrated into the personalization mechanisms of the online music providers. In this paper, we present the RIADA infrastructure which is composed by a set of systems able to annotate emotionally the catalog of songs offered by Spotify based on the users' perception. RIADA works with the Spotify playlist miner and data services to build emotion recognition models that can solve the open challenges previously mentioned. Machine learning algorithms, music information retrieval techniques, architectures for parallelization of applications and cloud computing have been combined to develop a complex result of engineering able to integrate the music emotions into the *Spotify*-based applications.

Keywords

Affective Annotation, Cloud Computing, Emotion Recognition, Machine Learning, Music, Spotify.

DOI: 10.9781/ijimai.2022.04.002

I. Introduction

Currently, the music streaming services are facing the chal-lenge of offering personalised media contents to their users [1]. The huge size of their music catalogs has promoted the develop-ment of innovative tools that help users to find among so many choices the songs that best suit their tastes. Most of these tools analyse the users' profiles and listening habits applying artificial intelligence techniques (such as collaborative filtering or content-based filtering), and then make personalised music recommenda-tions to the users [2]. These automatic tools are compatible with other types of content access services, for example, with services that publish the playlists created by other users or with social networks in which the users can share their listening experience. In all these solutions there are some factors that play a relevant role in the process of selecting the music, such as the musical genre and the popularity of the songs, the listening context and the activity that the user is doing, or certain cultural criteria, for instance. Nevertheless, other interesting factors have not had too much prominence among the tools offered by the streaming ser-vices, for example, the music emotions.

The relationship between music and emotions has been widely studied during the last years and the interest of including the users' emotions as a factor for the content personalization has promoted the research area commonly refereed to as Music Emotion Recognition (MER) [3]. The goal of this area is to annotate automatically the songs from an emotional point of view. These annotations usually represent the perceived or the felt emotions by the users when listening the songs, that is, the perception of emotions or the induction of emotions [4]. These two emotional dimensions are clearly different: the former is related to the emotions expressed by the music through the songs' structure and sound properties, whereas the second depends on the listener's experience and is influenced by her/his mood and context, among other factors. During the last years machine learning and deep learning techniques are being widely used to determine automatically both types of emotions in order to improve the music retrieval and recommendation systems [5].

The MER systems that work with perceived emotions are mainly based on the songs' audio. These audio files are processed by specialised tools in order to extract the acoustic characteristics of the songs, called audio features. Then, some of these features are manually selected and used to build a recognition model that acts as a classifier. The recognition function determines the emotions that the listener perceives when listening to an input song from its audio features. The resulting emotions are finally translated to affective tags that enhance the songs' attributes. Although most of these recognition approaches obtain acceptable accuracy results, some works focus on including new features that can improve the classifiers, for example, features related to the songs' lyrics [6], [7]. On the other hand, the deep learning based approaches automate the extraction of features by providing more expressive representations of the music low-level and high-level characteristics [5]. Learning algorithms (mainly, different classes of neural networks [8]–[10]) are applied on music spectrograms for determining gradually the features of interest, and then for finding the relationship between these features and the output emotional categories. These solutions require less domain knowledge than machine learning approaches, but have a higher computational cost.

* Corresponding author.

E-mail addresses: alvaper@unizar.es (P. Álvarez), jgarciaqg@unizar.es (J. García de Quirós), sandra@unizar.es (S. Baldassarri).

Regardless of the learning method applied, the previous solutions present some drawbacks. Firstly, the lack of public large-size datasets that contain high-quality annotations about the songs' emotions. The reference datasets in the field of MER research are usually small (most of them have between 250 and 2, 000 songs) and have not solved the challenge of the subjective perception (the annotations are usually based on the users' feedback, which is influenced by different emotional and contextual factors that cause the quality of these annotations less than desirable) [11]. Secondly, their emotion recognition methods are usually applied on their own datasets or some of the reference datasets (mainly, the MediaEval Database for Emotional Analysis in Music [12] or the MIREX mood dataset [13]), but not on the music catalogues of the streaming services. The application to these catalogues would require to develop systems that integrate the recognition solutions with the technological infrastructure of the streaming providers. Thirdly, there is no consensus on which type of learning method is the best option, and it is even difficult to compare the existing approaches between them. Each proposal applies different feature extraction algorithms, selects different features to build the models, creates the models from different datasets and/or validates the results with different metrics and methodologies [3]. And, finally, most of the approaches determine the emotions perceived by the listeners, instead of considering the emotions that they feel. The problem of determining the emotional response of each user is complex. Nevertheless, wearable technology is demonstrating to be a good opportunity to make progress on the recognition of the listeners' feelings [14].

In this paper, we propose an infrastructure of services, called RIADA, for annotating emotionally the catalog of songs available in *Spotify*. The infrastructure interacts with the *Spotify* service platform and can be used to include the emotional dimension in the music recommendation services offered by the streaming provider. As part of the solution, we have built an automatic music emotion recognition system that classifies and annotates the songs according to the emotions perceived by the listeners. These emotions have been deduced from the playlists that the registered users publish in *Spotify*. The recognition system is based on machine learning techniques and the audio feature services available in the provider's service platform. A parallel version of the system has been programmed to be deployed and executed on cloud environments in order to be applied on large-size music datasets. The main contributions of the proposal with respect to the existing solutions are:

- it consists in a complex result of engineering able to solve a real-life problem related to the emotion recognition,
- the *Spotify* playlists have been used for deducing the emotions that the users perceive when listening to certain types of songs and for creating the dataset of annotated songs involved in the building of the recognition models,
- the emotion recognition is based on a set of multi-label classification models that work from the information published by the *Spotify* data services,
- finally, the system prototype has been successfully tested in a real cloud-based operating environment and, therefore, it has achieved a TRL-6 maturity level in the scale *Technology Readiness Level* [15].

The rest of the paper is structured as follows. Section II presents a review of the music emotion recognition systems based on machine learning techniques. It also reviews the music systems that have been programmed by integrating the *Spotify* services paying attention to those that consider the users' emotions. Section III describes the software architecture of the RIADA infrastructure. The process of building and validating the emotion recognition models is presented in detail in Sections IV and V. Section VI details the parallel and cloud-based implementation of the recognition system and shows its application to large-size music repositories. And, finally, Section VII discusses the main conclusions obtained and the future work.

II. Related Work

In this section, a review of the Music Emotion Recognition (MER) systems based on machine learning techniques and the*Spotify*-based systems that combine music and emotions are presented.

A. MER Systems Based on Audio Features and Machine Learning

There are many research works that propose automatic systems for the recognition of music emotions based on the combination of audio features and machine learning methods [3], [11]. These proposals differ from each other in terms of the method used for extracting the music features, the form of mapping those music features to emotions and, finally, the machine learning algorithms applied in the building of the recognition systems. In the following paragraphs these three issues are detailed from the perspective of the existing solutions in the field of MER research.

The first step of a typical MER system is the extraction of music features. In this review we are specially interested in those features extracted directly from the songs' audio files, called *audio features*. Several studies have analysed the relationship between certain audio features and the emotions that they produce in the listeners [16], [17]. Unfortunately, there is no consensus about which audio features are most appropriate to recognise the music emotions. Therefore, the process of feature selection is a difficult task that is usually based on researchers' experience and knowledge. This problem gets worse since there is a wide variety of processing audio tools that can be used for the feature extraction, such as *MIR toolbox* [18], *Marsyas* [19], *PsySound* [20], *OpenSmile* [21] or *JAudio* [22]. These tools apply different processing methods and, therefore, they compute different features. For this reason many works combine these toolkits for obtaining a large variety and number of features. Intuitively, we may think that it is a good option for increasing the accuracy of emotion recognition models, but some experiments have demonstrated that too many features lead to performance degradation [23].

On the other hand, it is necessary to determine and represent the emotions ascribe to the songs (the perceived or induced emotions, as was discussed in the introduction). This relationship between emotions and songs is affected by a strong subjectivity, because it depends on the listeners' character, musical preferences, genre or cultural factors, for instance. Therefore, the process of annotating manually the music emotions requires to involve many and diverse participants and, as consequence, it is time-consuming and prone to faults and impressions. With respect the representation of emotions, two different models are usually used in the field of the MER research: categorical and dimensional models [24]. The former conceptualise the emotions as a set of distinct categories (such as the *Hevner* model [25] or the *MIREX mood clusters* [13]); whereas the seconds map the emotions onto a two-dimensional space characterised by those emotions' feeling and intensity (such as the *Russell*'s affective model [26], the *Tellegen-Watson-Clark* model (TWC) [27] or the *Thayer* model [28]). Most of the MER systems use the *Russell*'s model, probably the most popular dimensional model in the development of emotion-based systems. Some proposal even work with simplifications or variations of this affective model. According to the presented in the above paragraphs, the creation of datasets that can be used for building MER models is a complex and difficult process. Most of these datasets are small in size and usually contain the songs' audio features and annotations that describe the emotions perceived by the users when listening to those songs. There is some reference datasets in the field of MER research, such as the *MIREX mood* dataset, which is the largest

one, containing about 2, 000 songs [13], the *DEAM* dataset (Database for Emotional Analysis of Music) composed by 1, 800 songs [12], or the *Allmusic* dataset composed by 900 songs. A more detailed description of the released and freely available datasets can be found in [11]. The advantage of using these datasets is that their songs are already emotionally annotated. In particular, the annotations of the *MIREX* dataset are based on their mood clusters (a categorical approach), and the annotations of *DEAM* and *Allmusic* on the *Russell*'s model (a dimensional approach). In any case, the challenge of having large-size datasets that contain the appropriate audio features and the emotional annotations with low levels of subjectivity is still open.

Finally, the different methods based on machine learning that are applied in the creation of computational models able to annotate automatically the songs' emotions are revised. We are interested in those methods that use the combination of audio features and emotion annotations. Most of these MER solutions are based on classification algorithms. Their goal is to obtain one or more emotion labels from the input song's features (single-label and multi-label classification, respectively). Recently, the multi-label classification has gained popularity because it takes in account the inaccuracy of human annotations and classifies each song into a number of different emotion categories. Different machine learning algorithms have been used for creating these classifiers, such as *Support-Vector Machines* (SVM) [29]–[32], *Random Forest* (RF) [33]–[35], *K-Nearest Neighbor* (KNN), *Decision Trees* (DT) [36], *Naïve Bayes* (NB) [33], [37], [38], *Linear Discriminant* (LD) [39] or *Gradient Boosting Machines* (GBM) [40], etc. Among these, SVM is the most used and a good option option for recognising the music emotions from the songs' audio features [33], [34], [39], [40]. This supervised method usually achieves good accuracy results with low computational power. Nevertheless, during the last years SVM has been usually combined with other classification methods in order to improve the classification results [41], [42]. In [3] a detailed review of the emotion classifiers proposed between 2003 and 2017 is presented and discussed (Table 4, pages 384-386). Regardless of the classification method used for the MER, in many cases is necessary to reduce the dimension of the feature space before building the recognition models. The choice of the appropriate features is many times more important than the machine learning method selected. *Principal Component Analysis* (PCA) [30], [33], [40], [43] and the ReliefF algorithm [32], [39] are two techniques commonly used for the feature reduction in the field of the MER research. These techniques help to create a more meaningful representation of the feature space by selecting the features of interest from the recognition point of view, and to improve the final results obtained by the emotional classifiers.

As conclusions, firstly, it is difficult to compare the results of the reviewed proposals because they work with different feature extraction tools, heterogeneous emotion annotated datasets and different classification strategies and methods. The same conclusion was reached by [44], as part of its interesting state of the art about the MER systems based on audio features. And, secondly, future MER solutions should address some drawbacks of interest, such as to avoid the necessity of having the audio of the songs for extracting their features, to have available large-scale reference datasets, or to improve the accuracy of learning-based recognition by applying a multi-method approach.

B. Music Intelligent Systems Based on Spotify

In recent years, a wide variety of intelligent systems based on the *Spotify* services have been proposed. We are especially interested in those that extract knowledge from the songs' audio features and that help users to discover songs and to create their playlists. Within this review, our focus is set on how these proposals integrate the emotional dimension into their solutions.

Spotify offers a data service for accessing the audio feature of the songs available in its music catalogue. These features have been used to predict the future success of a song [45]–[47] or to determine the influence of music on the walking practice in urban space [48], for instance. These solutions analyse the audio features that are determinant for explaining the popularity of a song or the different way of walking, respectively, and then use these features to create machine-learning models (mainly, regression models) that solve the problem. On the other hand, the *Spotify* audio features have been also used for making music recommendations [49]–[52]. These recommendation systems combine the user preferences with the features of songs that she/he usually listens to. The preferences are determined by utilizing the users' past interactions [52] or by processing the messages published by those users in social networks, such as *Twitter* [51] or *Facebook* [50]. Then, different content and collaborative filtering techniques are applied to determine the similarity between songs based on their audio features and the similarity between users based on their preferences in order to make the recommendations. The same approach is even used by *Spotify* as part of its recommendation algorithms [53]. As conclusion, despite the recent interest in using the songs' audio features to develop *Spotify*-based intelligence systems, these solutions ignore the music emotions.

Other works related to the exploitation of playlists created on *Spotify* consider the emotions. These works apply different procedures for determining the emotions of playlists, as will be presented in the following paragraphs.

In some cases, these emotions are deduced by applying natural language processing over the titles of the songs contained in the playlist [54] or over the songs' lyrics [55]. In [54] the songs' titles are concatenated to build a sentence, and then linguistic analysis techniques are used to infer the emotions that will be possibly produced in the listeners. The author concludes that the results are not as expected and only the affection of love may be detected. On the other hand, in [55], a music emotion recognition method based on the sentimental analysis of the words contained in a song's lyric is proposed. The method consists in the building of a recognition model that combines machine learning and natural language processing techniques. This model is trained using the dataset *MoodLyrics4Q* and manually applied over a reduced dataset of songs in order to validate the approach.

In other cases, the emotions of a *Spotify* playlist are recognised by processing the audio features of the songs included in it [56]. A Support Vector Machine model classifies each song of the playlist as happy, sad or angry, and then a voting strategy is used to determine the emotion of that playlist. The classifier recognises the emotions from some of the audio features offered by *Spotify* for describing their songs. Despite the similarities with our work, this proposal is a work in progress that presents some relevant weaknesses: the dataset used for building the model was manually created and consists of a small number of songs (579 songs) reducing the reliability of the classifier, only 3 different emotions are recognised, the features used in the recognition were intuitively selected and are a restricted set, and finally the results are not formally validated (a playlist is only labelled as example).

Instead of analysing the existing playlists, other works provide tools for searching *Spotify* songs applying emotional criteria and supporting the creation of new playlists. In [57], the users classify emotionally the songs based on their personal experience listening to music. Each song is manually annotated using a colour scale that represents the different vibes produced in the listener. It makes difficult the application of this solution to large-size repositories of songs. Then, an user can introduce an input colour and find songs that could produce the desired effect. As an alternative, in [58], a prototype for searching

Spotify songs according to the user's mood is presented. The emotions of the songs are not explicitly recognised, but the authors assume that certain *Spotify* audio features can be mapped directly to moods (the validation of this assumption is not discussed). The mood-based search of songs is programmed applying similarity techniques over the features of interest and integrated into a prototype of application.

III. Description of the Proposal

In this section a high-level description of the RIADA infrastructure is presented. It is composed of a set of systems that collaborate for annotating emotionally the *Spotify* songs using net-accessible data resources. The semantics of these annotations and the affective model used for representing them are two relevant issues that are discussed in advance. After that, the architecture of the proposed system is presented.

A. Music and Emotions

The goal is to build a large-size database of emotionally annotated songs. These annotations represent the emotions that a user perceives when she/he listens to a song. In this subsection, the music data source and the affective model selected for implementing the songs' annotation are briefly explained. *Spotify* is the most popular online music streaming provider with more than 35 million of songs and 100 million of subscribers. Besides, it has recently published a platform of Web services and online tools for accessing the songs' metadata, searching the registered users' playlists, browsing the listeners' habits or making simple music recommendations [59]. These data services are available for encouraging the development of novel *Spotify*-based applications. As today, the emotions that the user perceives or feels when listening to the songs have not been included in the data offered by the music provider. Nevertheless, other data available on its platform could be combined in order to integrate the emotional dimension in its products, and to solve the open challenge of annotating a large-size catalog of songs.

On the other hand, the *Russell's* affective model has been selected for representing the emotions [26]. In this model, the affective states are represented over a two-dimensional space defined by *valence* (X-axis) and *arousal* (Y-axis) dimensions. The valence represents the intrinsic pleasure/displeasure (positive/negative) of an event, object or situation, and the arousal the feeling's intensity. The combination of these two dimensions (valence/arousal) determines four different quadrants: the *happy* (positive/positive), the *angry* (negative/positive), the *sad* (negative/negative) and the *relaxed* (positive/negative) quadrant. Then, each emotion is mapped to a point in the two-dimensional space and, therefore, is also located into one of the mentioned quadrants. Alternately, the emotions can be also represented as a probability vector of four values, one per each of the Russell's quadrants. These values are the probability that the emotion represented belongs to the corresponding quadrant. For example, the "I want to hold your hand" song by "The Beatles" has the following emotional annotation [0.174, 0.765, 0.155, 0.006] which represents that is a *happy* song with a 0.765 probability (the sad, angry and relaxed probabilities are 0.174, 0.155 and 0.006, respectively).

Therefore, the proposal consists of annotating the songs considering the four quadrants of the *Russell's* affective model. The probability of that the emotions ascribe to a song belong to each of those quadrants is mainly estimated from the song's audio features. Those features can be obtained from the *Spotify* data services and, therefore, are available without the need for having the song's audio file. This last issue is very important from our proposal point of view because it will allow us to apply the solution on a large-scale, although it involves delegating the feature extraction process to Spotify.

B. Architecture of the Proposed System

The *RIADA* infrastructure presented in this section is composed of the set of software systems that are responsible for creating and updating the database of emotionally annotated songs. The infrastructure has been integrated into a *multi-tier architecture* [60], [61] in order to make easier the logical and physical decomposition in different tiers of functionality involved in the global solution.

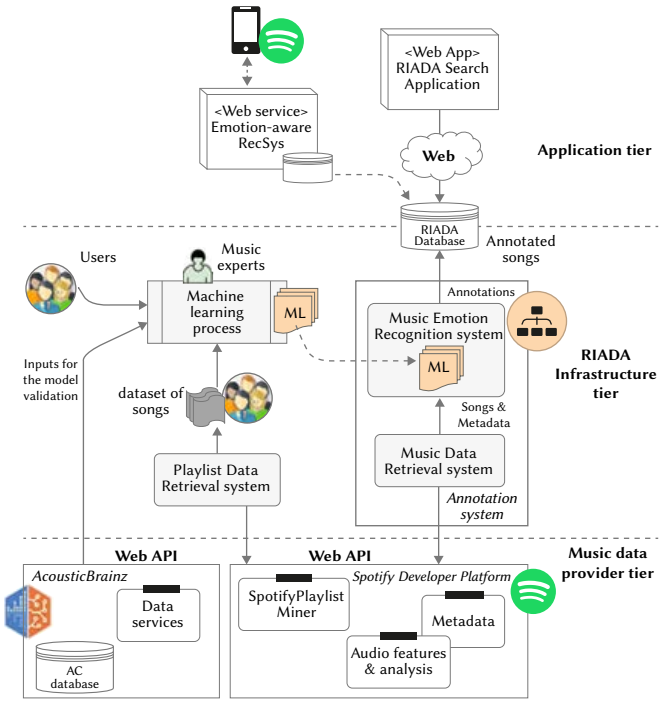


Fig. 1. High-level architecture of the solution.

As shown in Fig. 1, the solution has been divided in three tiers: the *music data provider tier*, the *RIADA infrastructure tier* and, finally, the *application tier*. The first is composed of the online services and the tools offered by the music data providers, in particular, the *Spotify* and *AcousticBrainz* [62] solutions have been integrated into this data tier. The second tier contains the systems involved in the annotations processes and the resulting database of annotated songs. These systems works with the music data providers for building the music emotion recognition models and applying these models over the *Spotify* catalog of songs. The RIADA database is the interface of this second tier from the applications point of view. These RIADA-based applications constitute the last tier of the architecture.

Following the different functional elements of the system are described in more detail. The music data tier is mainly composed of the net-accessible services integrated into the *Spotify Developer Platform*. These offer a set of Web APIs that allow to access the music database of the provider and to retrieve information about *Spotify* songs and the most popular playlists published by registered users. Additionally, the *AcousticBrainz* services have been also included in this data tier, and provide functionality for extracting the songs' acoustic characteristics and for accessing to high-level data computed from those characteristics. Some of these high-level data are related to the mood.

On the other hand, the core component of the RIADA infrastructure is the music emotion annotation system (represented in the right side of the RIADA tier). It consists of a *Music Emotion Recognition* (MER) system which integrates a set of machine-learning models for annotating emotionally *Spotify* songs. These models work with the songs' audio features and predict the emotions that the users perceive

when they listen to each of these songs. Then, these predictions are translated to emotional labels (probability vectors based on the Russell's quadrants) which are stored into the RIADA database. In the recognition process is involved the *Music Data Retrieval* (MDR) system which is responsible for interacting with the *Spotify* data services in order to get the information needed to make the emotional predictions. Note that the annotation system has been programmed implementing parallelism techniques to be applied over large-sized catalogs of songs, as will be presented in Section VI.

A fundamental component of the MER system are the machine learning models used for the emotion recognition. Before building these models, it is necessary to have a dataset of songs emotionally annotated. In the proposal, this dataset is created by the *Playlist Data Retrieval* (PDR) system. Its functionality is based on the *Spotify Playlist miner API* which aggregates the top songs from the most popular playlists created by the *Spotify*'s users. The PDR system processes the names and descriptions of these top songs and from that textual information deduces the emotions that the users can perceive when listening to them. This process is explained in detail in Section IV.

Then, a *Machine learning process* is responsible for building the recognition models from the dataset of annotated songs. This process implements a multi-model hybrid method in which a different model is created for recognising the emotions contained in each of the Russell's quadrants. A detailed description of the process will be presented in Section V. Finally, the models are integrated into the MER system in order to support the massive annotation of *Spotify* songs.

Finally, the emotionally annotated songs are stored into the *RIADA database*. The attributes and annotations of these songs are stable and do not require to be downloaded or computed again. Nevertheless, as *Spotify* is continuously adding new songs to its online catalog, it is necessary to update periodically the contents of the RIADA database. These updates are made by executing the music emotion annotation system previously presented. The system can be configured to work in update mode, and in this case it will process and annotate those songs that are not already included in the database.

IV. Creation of a dataset Based on the *Spotify* Playlists

A dataset of emotionally annotated songs has been created to be used in the building and training of the emotion recognition models. *Spotify* provides certain information about its playlists, but not about the emotions that the users perceive when they listen to those playlists. In this work, a method for deducing those emotions from the playlists that are available through the *Spotify Playlist miner* is presented.

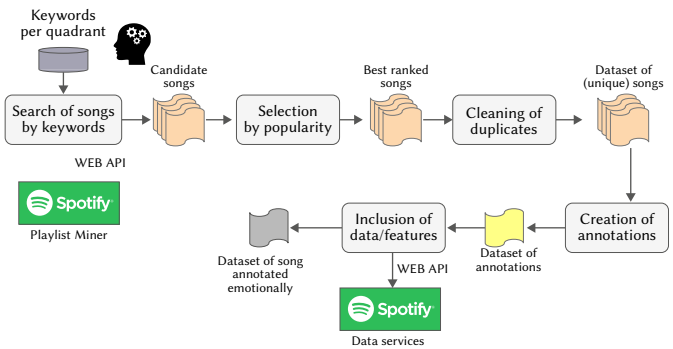


Fig. 2. Description of the data preprocessing process.

Fig. 2 shows the process followed for creating the dataset of annotated songs and the tools involved in it. Before starting the process, a set of keywords have been defined for each Russell's quadrant. These keywords correspond with emotions mapped to each particular quadrant, for example, the keywords *happy*, *joy*, *motivating* or *excited* are some of those included in the *Happy* quadrant.

The process begins executing the task *Search of songs by keywords*. This task invokes to the *Spotify Playlist miner API* which returns aggregations of songs contained in the most popular playlists published by the *Spotify*'s users. These aggregations are created from search criteria based on keywords which are matched with names and descriptions of published playlists. We have assumed that a song contained into a playlist called "Motivating music for running" is likely that conveys positive energy and emotions. Therefore, that song could be annotated as *happy*. Considering this, a set of requests are executed for each quadrant. The search criterion of a request contains a subset of the keywords defined for the quadrant of the interest and some unwanted keywords. These latter are selected from among those included in the other three quadrants. Different non-repeating combinations of keywords have been calculated for each quadrant in order to determine its set of search criteria. For example, *"Happy AND Joy AND Motivating AND NOT Sad AND NOT Relaxed"* or *"Joy AND Motivating AND NOT Angry"* are some of the criteria configured for getting songs that are probably contained in the *happy* quadrant. In the future these criteria could be improved by analysing the combinations of keywords that return the most appropriated playlists. Sentiment analysis techniques based on text could be applied for deducing the playlists' emotions from the titles or the lyrics of their songs, such as in [56] or [55], respectively. Each playlist could be emotionally characterised from the emotions obtained, and then the results contrasted with the quadrant of interest in order to evaluate the quality of the search criteria. In any case, the result of this first task is a set of candidate songs for each Russell's quadrant.

Secondly, since the songs returned by the miner have an attribute that represents their popularity, the task *Selection by popularity* processes these songs to select the most popular. The selection is achieved by applying the inverse frequency, a numerical statistic widely used in the field of information retrieval that is intended to reflect how important a song is in the returned playlists. Those songs that have an inverse frequency greater than 2.5 are discarded (this threshold have been experimentally determined). Then, the best ranked songs are filtered to remove those that appear in more than one quadrant, and therefore that could generate confusion in the creation and training of future classification models (task *Cleaning of duplicates*). Finally, the songs of each set are annotated with their respective label (in the case of the example, they will be annotated with the label *Happy*) in order to create an unique dataset of annotated songs (task *Creation of annotations*).

The last task of the process consists in completing the data of annotated songs. The general purpose attributes (such as the artist, the album, etc) and the audio features of these songs are obtained from the *Spotify* Web data services. More specifically, the list of audio features returned by *Spotify* is: loudness, energy, tempo, acousticness, valence, liveness, speechiness, instrumentalness, danceability, key, duration, and mode. The definition, unit of measurement and representation format of these features are available in [63], as part of the specification of the *Spotify AudioFeaturesObject*. All these data are recorded jointly with the emotional annotation into the dataset as result of the task *Inclusion of data and features*.

Finally, the results of the data processing stage are briefly summarised. At the beginning of the stage, we obtained 83,078 *Spotify* songs from the *Playlist Miner*. More specifically, the number of songs for each of the four requests was: 19,092 songs that probably convey emotions located into the *Happy* quadrant, 17,661 into the *Angry* quadrant, 23,931 into the *Sad* quadrant, and 22,394 into the *Relaxed* quadrant. After applying the inverse frequency, there were selected

3,055 songs for the *Happy*, 1,817 for the *Angry*, 2,943 for the *Sad*, and 1,671 for the *Relaxed* quadrants. This selection process reduces significantly the number of available songs, but increases confidence in results concerning the users' perceived emotion. Finally, the songs located into more than one quadrant were eliminated, obtaining a final dataset composed of 1,644 songs for the *Happy*, 1,307 for the *Angry*, 1,737 for the *Sad*, and 504 for the *Relaxed* quadrant. Therefore, the *prepared data* database used for the training of the models contained a total of 5,192 songs.

V. Building of *Spotify*-based Learning Models

Once presented the dataset of annotated songs, the three stages directly involved in the building of the machine learning models are following described: the analysis and extraction of the features of interest, the application and the training of algorithms, and the validation of the models. Besides, an experiment with real users has been carried out to corroborate the validity of these models before integrating them into the *Music emotion recognition system*.

A. Analysis and Extraction of Features

As will be discussed later, we have decided to build four classification models to recognise the music emotions, one per each of Russell's quadrants. Each model will predict whether or not the emotions that the users perceive when they listen to a song belong to the corresponding quadrant. The decision of considering four different hypothesis aims at creating more accurate models. Nevertheless, it is necessary to identify first the audio features that must be involved in the building of these models. We have decided to analyse these features from the perspective of each Russell's quadrant, that is, we are supposing that a feature may be significant to identify a class of emotions, but irrelevant to others. In the literature this analysis is usually carried out by applying three different approaches [3]: by selecting the same features used in other similar research works, consulting the opinion of music experts, or evaluating and interpreting certain statistical tests frequently used in the machine learning field.

In this paper, we have applied a combination of the three approaches. Firstly, statistical tests have been calculated to evaluate the degree of features' relevance in each quadrant. Then, the test results have been contrasted with the conclusions published by other similar works in the field of MER research and refined by a group of music experts in order to determine the features to be finally selected. Table I shows the result for each quadrant after calculating the tests (first step). Additionally, the audio features selected after considering the research works and the experts' conclusions have been highlighted in green color (second and third steps).

TABLE I. Analysis of Songs' Audio Features

Happy	**Angry**	**Sad**	**Relaxed**
valence	acousticness	energy	instrumentalness
acousticness	energy	acousticness	energy
danceability	speechiness	valence	loudness
energy	loudness	loudness	acousticness
instrumentalness	danceability	liveness	valence
loudness	liveness	duration	danceability
duration	tempo	tempo	speechiness
speechiness	instrumentalness	instrumentalness	duration
tempo	mode	key	tempo
key	duration	mode	mode
mode	valence	danceability	liveness
liveness	key	speechiness	key

In more detail, three statistical tests have been calculated, specifically, the *Chi Squared*, *ANOVA F-value* and *Mutual information* tests. These tests order the features from most to least relevant. Then, a voting strategy has been applied to combine the results of the three tests, as shown Table I for each quadrant.

Then, the features considered in other *Music Emotion Recognition* systems have been reviewed [3], [44]. Most of these systems work with features extracted from the audio of the songs. In general, they are mainly interested in extracting *timbral* and *rhythmic* features and in determining the *intensity* of the songs. Each solution uses a different audio processing tool, which makes it difficult to compare their results (it is even unknown how the features are calculated by *Spotify*). Nevertheless, these conclusions can be interpreted from the *Spotify* point of view. According to our interpretation, acousticness, instrumentalness or speechness are audio features related to the songs' timbre, tempo or danceability to the rhythm, and finally energy and valence to the intensity. Therefore, those *Spotify* audio features must be included in the final selection. For example, the tests determined that the valence and danceability features could have a low relevance for the *Angry* and *Sad* quadrants, respectively. However, after analysing the existing MER proposals we have decided to include them among the selected features.

Thirdly, an activity was organised with the participation of three music experts. The goal was to gather their opinions about the importance that the *Spotify* audio features can have in the emotion recognition. The activity had two stages. In the first each expert individually studied the information published by *Spotify* about these features (definition, units of measurement, feature extraction procedures, etc.) and listened to a collection of songs for understanding the intrinsic nature of those features. Then, the second stage consisted of a discussion group in which the experts contrasted their individual opinions and collaboratively made a list of the most relevant features. They concluded that the most significant features are: energy, valence, danceability and tempo. These conclusions are consistent with those of the existing proposals [3] and reinforce the decision to include the valence and danceability features for the case of the *Angry* and *Sad* quadrants. Besides, they believed that the features key and duration are the least relevant ones. The rest of features could have a moderate influence depending on the emotion to be recognised.

Therefore, the final proposal consists of using different audio features for building of each classification model (this type of approach was already considered by [64]). As described above, the audio features that have been finally selected for each classification model are represented in green color in Table I.

B. Model Selection and Training

In this stage, the goal is to build a machine learning model for each of Russell's quadrant. The *target function* of these models is defined as: the input are the song's audio features, while the output is a pair of values (a logical value and a real value) that predicts whether the emotions perceived by the listeners are located into the corresponding quadrant. Therefore, the emotional annotation of a *Spotify* song will consist of two vectors of four values. For example, the "I want to hold your hand" song by "The Beatles" will have the following emotional annotation ([*true*, *false*, *false*, *false*], [0.765, 0.155, 0.174, 0.006]) which represents that is a *happy* song with a 0.765 probability. The angry, sad and relaxed probabilities (0.155, 0.174 and 0.006, respectively) are lower than the classification threshold and, therefore, the song is also classified as not sad, not angry and not relaxed.

For the building of the models, three types of machine learning algorithms have been considered: *Support Vector Machine* (SVM), *K-Nearest Neighbours* (KNN) and *Random Forest* (RF). These have been widely used with good results in the recognition of emotions [65], [66]. Nevertheless, we have also considered the possibility that the use

of an unique algorithm is not the best option for building the different classification models. Therefore, the best machine learning algorithm for each quadrant (its model) is also studied.

Before comparing the algorithms, there must be defined the positive and negative datasets that will be used in the training and testing of the resulting models. The starting point is the dataset of annotated songs that was created during the preprocessing stage (described in Section IV). For each quadrant, this dataset has been been divided into two parts. On the one hand, the songs that were annotated with the emotional value of that quadrant and, on the other hand, the rest of songs. For example, for the *Happy* quadrant, the first dataset is composed by the songs annotated as *happy* (positive class), and the second by those annotated as *angry*, *sad* and *relaxed* (negative class). This partitioning strategy has been replicated for the four quadrants.

Then, the three selected machine learning algorithms have been applied in the building and training of the models. The choice of input audio features is determined by the results of the previous analysis. Besides, the range of input hyperparameters has been varied in order to find the best configuration. The library *Scikit randomized search* has been used for this evaluation since it provides an efficient procedure for the analysis of the possible permutations [67].

Table II shows the results for the different combinations of algorithms and quadrants. The best combinations have been highlighted in green color. Each of these combinations has been configured with the optimal input of audio features and hyperparameters. The models have been trained by performing a *Repeated 5-fold cross validation*. The use of this validation approach is especially important when the models are built from small-sized or unbalanced datasets, as in this case. A ratio 70/30 was applied to split the original dataset into two sets, a training set and a testing set. This ratio was experimentally chosen and it seems to be a good option for this specific classification problem. The data splitting was manually made to maintain the original percentage of songs of each quadrant in the training and testing datasets. Besides, the cross validation was configured to use the *Stratified* library of *Scikit learn* to preserve the percentage of samples for the positive and negative classes. As conclusions, *Random Forest* models offer good accuracy and F1-score results for the four quadrants. These results contradict the initial assumptions of applying different algorithms for building the model of each quadrant in order to improve the models' accuracy. The mean accuracy is 88.75%, a good result compared to the other similar studies presented in Section II.

TABLE II. Comparative of Different Models/Quadrants

Algorithm	Tests	Happy	Angry	Sad	Relaxed
SVM	accuracy	0.767	0.872	0.8036	0.929
	f1	0.752	0.821	0.783	0.733
	precision	0.7475	0.8435	0.7792	0.8624
	recall	0.7715	0.8059	0.7991	0.6801
K-NN	accuracy	0.843	0.876	0.842	0.935
	f1	0.822	0.824	0.816	0.784
	precision	0.8256	0.8516	0.8185	0.8505
	recall	0.8198	0.8055	0.8142	0.7428
Random forest	accuracy	0.844	0.899	0.862	0.945
	f1	0.820	0.860	0.839	0.801
	precision	0.8307	0.8828	0.8488	0.9299
	recall	0.8083	0.8446	0.8353	0.7392

The confusion matrices of the *Random Forest* models reaffirm the good performance of the classification models, as can be seen in Fig. 3. Nevertheless, it is also important to analyse the *false positives* in order to understand where the models fails.

Table III shows a comparison of the predictions (rows) versus the true emotions (columns) for each quadrant. The diagonal of the matrix corresponds to the true positives (highlighted in grey color), while the rest of values in each row corresponds to false positives. Firstly, the results of the models *Happy* and *Angry* have been analysed. As explained, these two affective quadrants have the same arousal (the feeling's intensity), but different valence (the intrinsic pleasure/displeasure) in the Russell affective model. The model *Happy* predicts 91 false positive of which 40 were incorrectly annotated as *angry* (48% of the total false positives), and the model *Angry* predicts 50 false positive of which 39 are songs that were annotated as *happy* (78% of the total). Therefore, these wrong predictions may be due to the valence of those songs is near zero (the zero value represents the axis that separates the two quadrants), and in those cases the models are not able to classify correctly. On the other hand, the results of analysing the models *Sad* and *Relaxed* are similar (both quadrants have the same arousal, but different valence again). In this case, the model *Sad* predicts 96 false positives of which 50 were annotated as *relaxed* (49% of the total), and the model *Relaxed* predicts 7 false positives having been all these songs annotated as *sad*. As conclusion, we suppose that the songs that are mapped to a point close to the affective quadrants' axis may be wrong classified in some cases. Nevertheless, the results of models are good being the percentage of false positives very low.

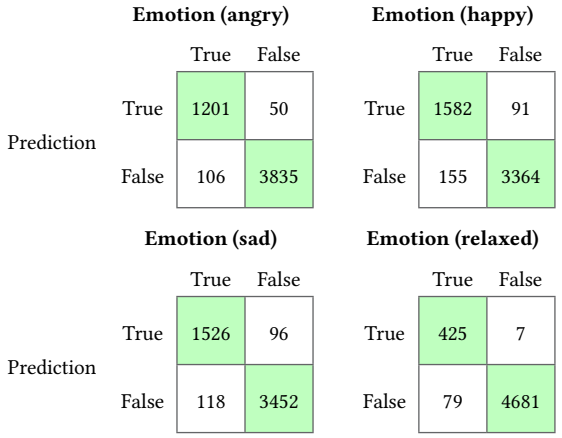


Fig. 3. Confusion matrices of the *Random Forest* models.

TABLE III. Matrix of Positive Predictions Versus True Emotions

	Happy	Angry	Sad	Relaxed
Happy	1582	40	29	22
Angry	39	1201	8	3
Sad	24	22	1536	50
Relaxed	0	0	7	425

C. Validation of the Models

The next stage is the validation of the models. From a methodological point of view, we have selected music database published by the project *AcousticBrainz* [62] for analysing the accuracy of the models built in the previous stage. This repository contains over 11 million of songs, but the version that can be downloaded is only composed by half a million (songs released before 2015). Each song has an attribute that represents the emotion conveyed by it. More specifically, this attribute is a vector of four numerical values, where each of them determines the probability of conveying an emotion belonging to a Russell quadrant. These values have been generated from users' opinions published in the music Website *Last.fm*. For that reason, these values can be especially interesting for validating the decision of creating the emotional annotations from the *Spotify* playlists (in both cases, the users' opinions and the metadata of the playlists represent the

emotional perception that the users have of the songs) and of building the recognition models using these annotations.

The downloaded dataset has been preprocessed for selecting those songs that have a high probability value in one emotion and a low probability value in the other three (in other words, a quadrant stands out from the others). After the preprocessing, the dataset size has been reduced to 60,000 songs (around 15,000 songs per quadrant in order to have a balanced sample). Then, the audio features of these songs have been obtained by invoking the *Spotify* Web data services. In this way, the features and an emotion for each song contained into the dataset are obtained. Afterwards, the goal is to validate the models using this set of *AcousticBrainz* songs.

Table IV shows the validation results. In general, the results get worse with respect to those presented in Table II: the average accuracy drops from 0,887 to 0,724, and the average f1 from 0,83 to 0,696. Nevertheless, these results were expected because two different types of annotations are "compared": the emotions deduced from the *Spotify* playlists (used for building the classification models) and the emotions extracted from users' opinions (for validating them). In any case, the most important issue is that the accuracy results are still quite good, with a mean accuracy over 72%. Besides, these results are interesting since the *Random Forest* models are particularly sensitive to changes in input data. Therefore, it is concluded that the *Random Forest* models can be a good option to recognise the emotions that the users perceive when they listen to *Spotify* songs.

TABLE IV. Results of the Model Validation

Model	Test	Happy	Angry	Sad	Relaxed
Random forest	accuracy	0.694	0.705	0.771	0.729
	f1	0.623	0.700	0.745	0.719

D. Assessment With Real Users

As a complement to the *AcousticBrainz*-based validation, an experiment with real users has been programmed to corroborate the validity of the resulting annotations. In the design of the experiment the "Pick-A-Mood" (PAM) model [68] has played a relevant role. PAM a cartoon-based pictorial instrument for representing the possible user's emotional states based on the Russell's affective model. In particular, PAM expresses eight different mood states, two for each of the four quadrants: excited and cheerful (*happy* quadrant), irritated and tense (*angry* quadrant), sad and bored (*sad* quadrant), and relaxed and calm (*relaxed* quadrant). Also, the model includes a neutral state. The added value of PAM is that its visual representation requires little time and effort of the respondents, which makes it suitable for the design of experiments in which the users must introduce their emotions.

At the beginning, a playlist composed by 12 *Spotify* songs was created, three songs of each of Russell's quadrants. These songs were selected from the dataset annotated emotionally using the *Random Forest* models, and randomly ordered in the new playlist. The experiment consisted in playing each of the songs and in asking the user what emotions she/he perceived when listening to that song. The user must listen to the entire song before responding the question since we are interested in annotating at the song level (the *Spotify* audio features used for creating the classification models are calculated processing the entire audio of songs). A *Google form* survey has been created to gather the users' responses. The survey presents a visual representation of the PAM model after playing a song and allows the user to select a maximum of two emotional states. The duration of the experiment is about 40 minutes (three and a half minutes per song, approximately).

In the experiment 25 users participated. Table V summarises the results obtained. The structure of the table is the following. It has 12 data rows, one for each song (S_1-S_{12}). Each row contains information about the emotions perceived by the users when listening to the song *Si* (these have been determined applying the recognition models built and are represented in the columns Em_{main} and $Em_{secondary}$), and about the users' responses after listening to that song (rest of columns). The column *Emmain* determines the emotion the listener is most likely to perceive and the corresponding probability value. For example, the song *S*1 ("Sorry, I'am a lady" by the duo "Baccara") was annotated as ([*true*, *false*, *false*, *false*], [0.66, 0.084, 0.014, 0.28]) which represents that is a *happy* song with a 0.66 probability (column Em_{main}). Likewise, the column $Em_{secondary}$ determines the emotional quadrant with the second highest probability value. Considering the previous example, the song *S*1 is *relaxed* with a 0.28 probability.

On the other hand, the rest of columns contains the users' responses, specifically, a column for each of the PAM states (from *Excited* to *Calm*). These columns have an integer value that represents the number of users that perceived the corresponding emotion. In green color it has been highlighted the most selected emotion, and in yellow color the second most selected. These eight columns are grouped according to the Russell quadrants, for example, the columns *Excited* and *Cheerful* correspond with the quadrant *Happy*, as is represented at the headline of the table. An extra column has been added to represent the response "Don't Know" (the column *DK*).

Following, the results obtained are briefly discussed:

- The users mostly perceived a happy emotion (*Excited* or/and *Cheerful*) when they listened to a song annotated as *happy* (songs *S*1-*S*3). The same good results are achieved when they listen to a song annotated as *angry* (songs S_4-S_6). The most of users respond that they perceive a *Tense* or/and *Irritated* emotion, the two states corresponding with the quadrant *Angry*.
- The results of the songs *sad* (songs S_7-S_9) are not as conclusive as in the two previous cases. The users mostly ascribed relaxed and/or sad emotions when listened to these songs. Although the majority of opinions correspond with these two quadrants, the responses lean towards the quadrant *Relaxed*. This fact can be due to both quadrants have the same arousal in the Russell model, but they differ in the intensity of the emotion. It could have influence in the users' responses. Besides, the high probability values of secondary emotions could have also influence in the users' opinions. For example, the songs S_7 and S_8 have high values of relaxed probability, and it could also affect to the responses. As conclusion, the results are not as satisfactory as in the previous cases, but they are not bad either.
- Finally, the high probability values of secondary emotions seems to influence the results of the songs *relaxed* (songs S_{10}-S_{12}). For example, the users mostly perceived a happy emotion when they listened to the song S_{10}. Its value of happy probability is 0.45 and, therefore, it is high value. Besides, it is important to remark that the rest of user responses concentrate on the quadrant *Relaxed* (9 users felt relaxed). The same applies to the song *S*11, but in this case the quadrants *Sad* and *Relaxed* are the most selected (the value of sad probability is also high in this case). Finally, the song S_{12} is clearly relaxed, from the users point of view. Therefore, in our opinion, the results are good and show an interesting correlation between the emotional annotations and the users' opinions. We think that we should have also included into the playlist some relaxed song in which the secondary emotion had a low probability value.

As conclusion, although the number participants and the number of songs played regarding the size of the *Spotify* catalog are low, the results obtained are very promising. And, therefore, the method of emotional labelling based on the *Spotify* playlist and the *Random Forest* models built from those annotations can be a good approach for determining the emotions that the users perceive when listen to these songs.

TABLE V. Results of Experiment With Real Users

					Happy		Angry		Sad		Relaxed		
	Em_{main}		$Em_{secondary}$		Excited	Cheerful	Tense	Irritated	Sad	Bored	Relaxed	Calm	DK
S_1	happy	0.66	relaxed	0.28	4	19	0	3	0	1	1	0	1
S_2	happy	0.73	angry	0.25	13	11	0	2	3	7	0	2	1
S_3	happy	0.68	angry	0.14	9	13	0	5	0	0	0	0	2
S_4	angry	0.61	happy	0.42	8	0	15	11	1	1	0	0	0
S_5	angry	0.62	happy	0.51	6	1	14	12	0	0	0	0	1
S_6	angry	0.59	happy	0.40	5	2	11	13	0	0	0	0	0
S_7	sad	0.66	relaxed	0.32	4	0	1	2	1	4	13	8	0
S_8	sad	0.94	relaxed	0.58	1	1	0	3	5	10	6	4	4
S_9	sad	0.56	happy	0.49	1	1	0	3	3	7	8	8	2
S_{10}	relaxed	0.61	happy	0.45	12	7	0	0	1	1	6	3	4
S_{11}	relaxed	0.66	sad	0.50	1	1	1	0	2	9	6	12	0
S_{12}	relaxed	0.59	happy	0.48	5	5	0	0	1	4	7	11	1

VI. An Automatic System for Annotating Emotionally Songs

In this section the design of the two systems involved in the annotation of songs is presented in detail: the *Music Data Retrieval* (MDR) system and the *Music Emotion Recognition* (MER) system. The goal is that these systems work automatically and are able to process efficiently a large number of songs by using the classification models previously created.

A. Description of the Annotation Process

Fig. 4 shows the stages and the data involved in the process proposed for annotating emotionally the *Spotify* songs. The green stages represent interactions with the *Spotify* data services; whereas the red stage represents the recognition actions executed by the MER system. The input is a database of artists which was previously created applying mining techniques over the data services offered by the music provider. The output is the RIADA database.

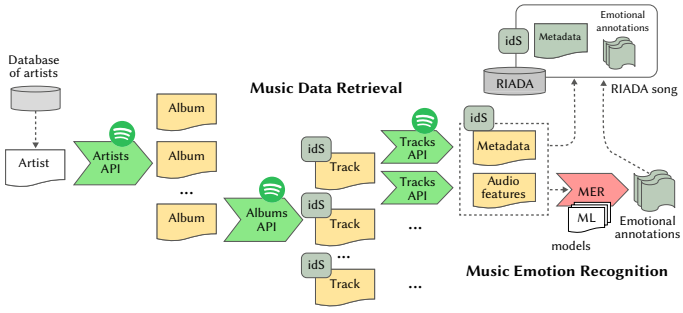


Fig. 4. Stages and data involved in the annotation process.

The MDR system is responsible for executing the first part of the process. It consists of a sequence of invocations to the *Spotify* data endpoints. Firstly, the *Artists endpoint* is invoked for getting the list of albums published by each artist. Then, each album is individually processed. A request to the *Albums endpoint* is executed for getting information about all the songs (or tracks in *Spotify* terminology) contained into that album. Each track contains a unique *Spotify ID* that will be reused to identify the song in the RIADA database. This decision facilitates the integration of the *Spotify* tools in the RIADA-based future applications. Optionally, the metadata of each song can be also obtained invoking the *Tracks endpoint*. An independent request is executed for each song of the album. Some metadata of the songs can be finally stored into the RIADA database (in grey color), if their are available, for example, the song's author, album, title, musical genre, or year of publication.

Subsequently, the MER system is in charge of annotating emotionally these songs, as shown in the right side of Fig. 4. Before, it must obtain the audio features of the songs invoking again the *Tracks endpoint* (a request for each song). Then, the MER processes each song's features and applies the four *Random Forest* models to compute the emotions that the users will perceive when listen to that song (specifically, the probability vector based on Russell's quadrants that represents those emotions). Finally, the MER creates a *RIADA song* structure which contains the song's *Spotify ID*, the emotional annotations and the metadata obtained during the retrieval phase.

Obviously, the data retrieval is a time consuming task due to it involves a large number of invocations to the endpoints and requires to process a large number of response files (in JSON format) for extracting the information of interest. These invocations are independent of each other, making possible to apply parallelism techniques to improve the efficiency of the systems involved. On the other hand, the emotion recognition also consists of a large-size bag of independent tasks (the execution time of each task is relatively small), and therefore it also requires high computing capacity for achieving an efficient processing.

B. Architectural Design of the System

The two systems involved in the annotation process have been designed according to the *master-worker architecture* [69]. It is a high-level design pattern that facilitates the parallel execution of applications composed by a set of independent tasks. The pattern consists of two class of processes: a master and a pool of workers. The former is responsible of assigning tasks to workers and guaranteeing that all of them are correctly completed; whereas the workers simply execute the assigned tasks. This architectural model is highly scalable by increasing (or decreasing) the size of pool of workers according to the execution requirements.

The master-worker architecture requires an asynchronous communication mechanism that makes possible the uncoupled

interactions between the processes involved. *Message brokers* have been usually used for this purpose, demonstrating their adaptability and effectiveness in this model of architectural solutions.

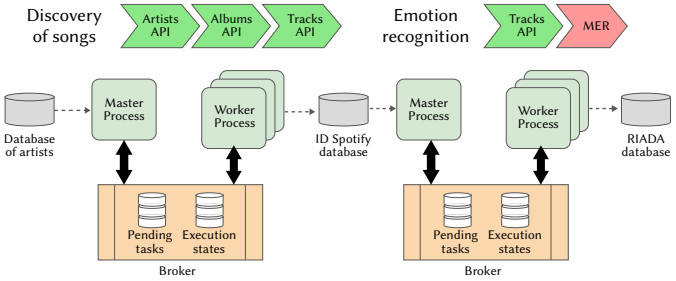


Fig. 5. Components and connectors of the architecture.

Fig. 5 shows the concrete design of the solution. It consists of two master-worker systems, one for implementing the MDR system and another for the MER system, connected between them by a shared database that contains the songs available in the music provider. Each master-worker system coordinates their processes through a broker based on message queues. Two queues have been internally declared: the *pending tasks queue*, in which the master publishes the tasks to be executed by some of the workers, and the *execution states queue*, in which the workers report to the master about the final state of executing each of their tasks (this state also includes the performance metrics concerning the execution of the task). On the other hand, both master-worker systems create their tasks from the data available into their input databases, and store the results computed by the workers into an output database. The granularity of the tasks depends on the restriction imposed by *Spotify* on the use of its services.

The process of getting the metadata of the songs published by 50 artists is an independent task in the MDR system. The master accesses to the database of artists, creates tasks that contains the identifiers of the artists to be processed (in blocks of 50), and then publishes these tasks into the broker. The workers execute the pending tasks when they are available, store the songs discovered into the output database, and finally notify the execution state of the task. These states are then used by the master for applying fault recovery strategies based on retrying the failed tasks and for generating reports of execution. On the other hand, in the MER system a task consists in annotating emotionally 50 songs, being the behavior of the system similar to that described above. In this case, the workers are responsible for getting the songs' audio features and for determining the emotional annotations applying the *Random Forest* models.

C. Cloud-based Deployment and Performance Analysis

A generic master-worker architecture has been programmed using the Python programming language. Besides, it integrates a *RabbitMQ server* as message broker in order to the processes can be executed and deployed in distributed computing environments, such as in a cloud infrastructure, for instance.

Fig. 6 shows the system configured for annotating emotionally the songs available in *Spotify*. The processes are executed on virtual machines of the *OVH cloud* (https://www.ovh.com/). Each master is running in a dedicated virtual machine in which it has been also deployed its input database. These databases have been designed and managed using *MongoDB* technology. The workers are running on a pool of machines so that these instances' computing resources are always busy. The message server has been installed as a service in the *CloudAMQP* (https://www.cloudamqp.com/), and therefore it is also deployed over cloud-based resources. Finally, the RIADA database in which the final results are stored has been installed in *mLab*, a cloud database service that hosts *MongoDB* repositories (https://mlab.com/). Therefore, the technological solution has been deployed and executed in a real environment. According to the *Technology Readiness Levels* scale (TRL, [15]) this solution has achieved a TRL-6 level, being a system prototype that may evolve into a final product.

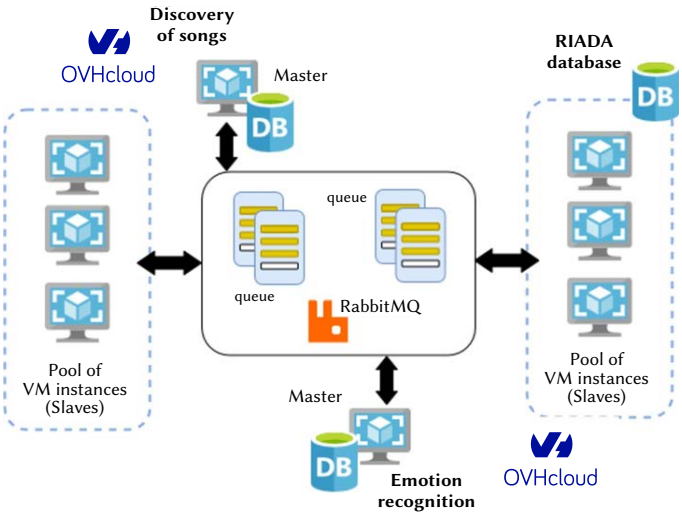


Fig. 6. Deployment over the OVH cloud resources.

A set of experiments have been also programmed for demonstrating the flexibility of the system to adapt to different resource provisioning scenarios and for analysing the scalability of the solution. Each experiment has consisted in annotating 750, 000 *Spotify* songs. Different computing instances have been hired for the execution of the master process, and different sizes of pools configured for the workers. These instances have been selected from those available in the *OVH cloud* on the basis of the authors' experience. In the future the selection criterion could be based on optimization techniques able to reduce the execution costs of the provisioning and to maximise the system performance. The use of these techniques would imply a detailed evaluation of the behavior of the deployed system, which it is out of the scope of this paper.

Table VI shows the results of the experiments. The table is structured as follows. The first column defines the type of OVH virtual machine (VM) hired to execute the master. The second column determine the number of the VM instances that compose the pool in which the workers are being executed. In all the cases, the pool is composed by *b2-7* computing instances, a general purpose virtual machine provided by OVH (2 cores at 2 GHz with 7 GB of RAM and a SSD storage of 50 GB). The MDR system executes a worker in each instance of its pool (these workers are continuously invoking to the *Spotify* data services -more than 60, 000 requests per experiment-, and the streaming provider generates response delay when two or more processes invoke it from the same machine), and the MER system executes two workers per instance (the number of interactions with *Spotify* is less, around 25, 000 requests). The third column is the total execution time needed for annotating all the songs. It is the sum of the times required to complete the execution of the MDR system and the MER system. The execution times of both systems are broken down in the fifth and sixth columns (these represent the CPU time considering all the cores involved and the real time needed to complete the execution, respectively). Finally, the last column is the mean execution time to complete a task in each of the parallel systems.

The first row of Table VI presents the results of executing sequentially the annotation system (the MDR system and the MER system are only composed by a worker). The total execution time is

TABLE VI. Performance Results of the Different Cloud-based Experiments

Master VM	Number of VM instances	Total time (hh:mm:ss)	System	CPU time (hh:mm:ss)	User time (hh:mm:ss)	Mean time per task (in seconds)
b2-7	1	11:31:35	MDR	1:19:15	1:19:15	47.55
			MER	10:12:20	10:12:20	2.51
b2-7	2	3:01:51	MDR	1:14:23	0:38:12	44.63
			MER	9:34:30	2:23:39	2.35
b2-7	3	2:00:24	MDR	1:21:35	0:28:09	48.95
			MER	9:13:08	1:32:15	2.25
b2-7	4	1:37:56	MDR	1:14:09	0:15:44	44.49
			MER	10:54:54	1:22:12	2.68
b2-7	5	1:18:58	MDR	1:08:19	0:12:34	44.49
			MER	11:01:09	1:06:24	2.70
r2-15	5	1:26:02	MDR	1:36:52	0:20:46	58.12
			MER	10:49:43	1:05:16	2.65
c2-7	5	1:17:25	MDR	1:35:01	0:19:27	57.01
			MER	9:36:59	0:57:58	2.42

more than 11 hours. Then, different experiments increasing the number of virtual machines are executed, from 2 instances to 5 instances (rows 2-5, respectively). The speedup obtained (considering this metric as the ratio between the sequential execution time and the parallel execution time of each experiment) is near to the number of workers that are being executing: 3.7X in the case of 2 instances and 4 workers, 5.7X in the case of 3 instances and 6 workers, 7.6X in the case of 4 instances and 8 workers, and finally 8.8X in the case of 5 instances and 10 workers. This behavior is a good result from the parallelization point of view. On the other hand, we have also evaluated the possibility of executing the master in other type of virtual machine, for example, in an instance with optimised CPU/RAM ratios and accelerated IOPS (specifically, a *r2-15* instance, with 2 cores with 5 GB of RAM, a SSD storage of 50 GB and a public network connection of 250 Mbps guaranteed), or in an instance for processing parallel workloads (a *c2-7* instance, with 2 cores at 3 GHz with 7 GB of RAM, a SSD storage of 50 GB and a public network connection of 250 Mbps guaranteed)). The results are shown in the two last rows of Table VI. The execution times are similar to those obtained in the experiment in which the master is executing in a *b2-7* instance (a pool of 5 instances), but a small improvement is observed in the MER execution time when a *c2-7* instance is hired.

D. A Prototype of RIADA-based Application

After executing the cloud-based system, the RIADA database contains the emotional annotations of 10 million of *Spotify* songs. As discussed in Section III, this database can be reused for developing different emotion-based applications. A Web application for searching songs applying emotional criteria has been developed as an example of RIADA-based application. The application also allows to filter the results according to the songs' musical genre or popularity, and to play a fragment of the songs found (30 seconds) through the *Spotify* music streaming service. Fig. 7 shows the interface of this application which is available in https://riada.djrunning.es/. It is hosted on *OVH hosting service* and its back-end is running on an OVH virtual private server. This back-end works directly with the RIADA database deployed in *mLab*.

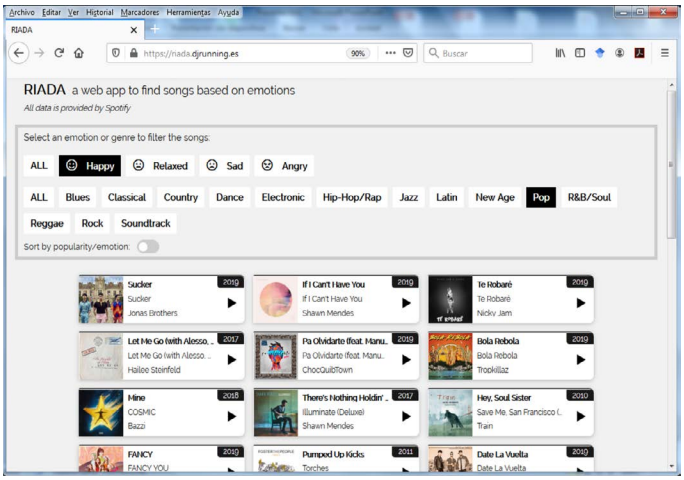


Fig. 7. Web interface of the search application based on RIADA.

VII. Conclusions and Future Work

The paper presents the systems involved into the RIADA infrastructure. These systems collaborate among them to annotate emotionally the *Spotify* catalog of songs. The processes of building the required machine learning models and of using those models to recognise the music emotions are based on the playlist and data services provided by *Spotify*. The integration of these services allows to apply the solution to a large-size catalog of songs, and it is an alternative to the usual approach based on the processing of songs' audio files. Besides, a parallel implementation of the RIADA systems has been proposed in order to improve the efficiency of the annotation processes. It is based on the master-worker architecture and has been deployed in different cloud-based environments.

On the other hand, the playlists published by the *Spotify* registered users play a relevant role in the solution. These playlists have been used to extract knowledge about what the users emotionally perceive when listen to a song, and then this knowledge has been applied in the building of the emotion recognition models. The proposal is innovative and it considers explicitly the user point of view. The resulting recognition models have been validated by using the *AcousticBrainz* dataset and by involving real users, obtaining good results in both cases. Moreover, the validation based on *AcousticBrainz* is interesting because it demonstrates that the models are only applied on the *Spotify* songs, but they can be applied successfully on other music repositories.

Although it has not been included in the paper, other alternatives to our recognition approach have been studied, for example, the possibility of building only one model able to solve a multi-class classification problem. In that case, the *target function* of this multi-class model was defined as: the input are the song's audio features,

while the output is a vector of four logical values ([*is_happy*, *is_angry*, *is_sad*, *is_relaxed*]) that determine in which Russell's quadrants could be located the emotions perceived by the listeners. We have built various models applying different machine learning algorithms and using the same dataset of songs. The *Random Forest* models are again the best option, obtaining an accuracy and f1 of 0.78 and 0.75, respectively. Therefore, the results are slightly worse than our proposal. In our opinion, the good results of our approach are due to: the splitting of the classification problem into four subproblems simplifying the classification constraints to be considered, and the adaptation of the building model stages (the selection of features and algorithms, and the training of models) to the characteristics and particularities of each quadrant.

Finally, some of the challenges that could be addressed in the future are briefly outlined:

- despite the good results obtained, to validate experimentally that the hypothesis formulated for annotating songs from playlists are really suitable in order to obtain an accurate dataset
- to publish the dataset (or a part of the dataset) so that it can be reused by other MER researchers (the *Spotify* terms of service and developer policies are being studied in order to find a viable option for its publication)
- to include the songs' lyrics in the emotion recognition in order to propose a multi-modal approach
- to explore the possibility of building *Spotify*-based accurate models able to recognise the emotions of each song's segments
- to build alternative recognition models based on fuzzy logic and to compare them with the models presented
- to analyse the execution behavior of the cloud-based system in order to optimise its configuration and to reduce the costs of its resource provisioning
- to create an emotion-aware music recommendation system based on the RIADA functionality and the content personalization and recommendation services provided by *Spotify*
- to reuse the RIADA technology for the generation of affective playlist. It is an open and interesting challenge in the field of the affective computing
- and, finally, to use wearable devices to detect the emotions induced to the listeners through the music. These devices could be used to include a new emotional dimension into the dataset or to study the correlation between the perceived emotions (the songs' annotations) and the induced emotions

Acknowledgment

This work has been supported by the TIN2017-84796-C2-2-R and RTI2018-096986-B-C31 projects, granted by the Spanish Ministerio de Economía y Competitividad, and the DisCo-T21-20R and Affective-Lab-T60-20R projects, granted by the Aragonese Government.

References

[1] G. Knox, H. Datta, "Streaming services and the homogenization of music consumption," 2020. [Online]. Available: https://research.tilburguniversity.edu/en/publications/streaming-services-and-the-homogenization-of-music-consumptionf, [Online; accessed 19-July-2020].

[2] M. Schedl, H. Zamani, C.-W. Chen, Y. Deldjoo, M. Elahi, "Current challenges and visions in music recommender systems research," *International Journal of Multimedia Information Retrieval*, vol. 7, pp. 95–116, 03 2018, doi: 10.1007/s13735-018-0154-2.

[3] X. Yang, Y. Dong, J. Li, "Review of data features-based music emotion recognition methods," *Multimedia Systems*, vol. 24, pp. 365–389, July 2018, doi: 10.1007/s00530-017-0559-4.

[4] A. Pannese, M.-A. Rappaz, D. Grandjean, "Metaphor and music emotion: Ancient views and future directions," *Consciousness and Cognition*, vol. 44, pp. 61–71, 2016, doi: https://doi.org/10.1016/j.concog.2016.06.015.

[5] J. Nam, K. Choi, J. Lee, S. Chou, Y. Yang, "Deep learning for audio-based music classification and tagging: Teaching computers to distinguish rock from bach," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 41–51, 2019, doi: 10.1109/MSP.2018.2874383.

[6] C. Gökalp, "Music emotion recognition: a multimodal machine learning approach," Master's thesis, School of Management, Sabanci University, 2019.

[7] G. Liu, Z. Tan, "Research on multi-modal music emotion classification based on audio and lyirc," in *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, vol. 1, 2020, pp. 2331–2335.

[8] Y. Dong, X. Yang, X. Zhao, J. Li, "Bidirectional convolutional recurrent sparse network (bcrsn): An efficient model for music emotion recognition," *IEEE Transactions on Multimedia*, vol. 21, no. 12, pp. 3150–3163, 2019, doi: 10.1109/TMM.2019.2918739.

[9] M. Russo, L. Kraljević, M. Stella, M. Sikora, "Cochleogram-based approach for detecting perceived emotions in music," *Information Processing & Management*, vol. 57, Sept. 2020, doi: 10.1016/j.ipm.2020.102270.

[10] R. Sarkar, S. Choudhury, S. Dutta, A. Roy, S. K. Saha, "Recognition of emotion in music based on deep convolutional neural network," *Multimedia Tools and Applications*, vol. 79, pp. 765––783, 2020, doi: 10.1007/s11042-019-08192-x.

[11] S. Zhao, S. Wang, M. Soleymani, D. Joshi, Q. Ji, "Affective computing for large-scale heterogeneous multimedia data: A survey," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, no. 3s, pp. 1–32, 2019, doi: 10.1145/3363560.

[12] M. Soleymani, A. Aljanaki, Y. Yang, "DEAM: Mediaeval database for emotional analysis in music." http://cvml.unige.ch/databases/DEAM/, 2016. [Online; accessed 19-July-2020].

[13] X. Hu, J. Downie, C. Laurier, M. Bay, A. Ehmann, "The 2007 mirex audio mood classification task: Lessons learned," 01 2008, pp. 462–467.

[14] B.-J. Han, S. Rho, S. Jun, E. Hwang, "Music emotion classification and context-based music recommendation," *Multimedia Tools and Applications*, vol. 47, no. 3, pp. 433–460, 2010.

[15] E. Commission, "Horizon 2020 work programme 2014 – 2015. european commission decision c(2015)8621," 2015. [Online]. Available: https://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/annexes/h2020-wp1415-annex-ga_en.pdf, [Online; accessed 25-July-2021].

[16] A. Gabrielsson, E. Lindstrom, *The influence of musical structure on emotional expression*, pp. 223–248. Oxford University Press, 2001.

[17] R. E. Thayer, R. J. McNally, "The biopsychology of mood and arousal," *Cognitive and Behavioral Neurology*, vol. 5, no. 1, p. 65, 1992.

[18] O. Lartillot, P. Toiviainen, "A matbal toolbox for musical feature extraction from audio," in *Proceedings of the 10th International Conference on Digital Audio Effects, DAFx-07*, Bordeaux, France, 2007, pp. 1–8.

[19] G. Tzanetakis, "Marsyas-0.2: A case study in implementing music information retrieval systems," *Intelligent Music Information Systems: Tools and Methodologies*, pp. 1–48, 2007, doi: 10.4018/978-1-59904-663-1.ch002.

[20] D. Cabrera, "Psysound: A computer program for psychoacoustical analysis," in *Proceedings of the Australian Acoustical Society Conference*, 1999, pp. 47–54.

[21] OpenSMILE, "OpenSMILE audio feature extraction." https://www.audeering.com/opensmile/, 2020. [Online; accessed 19-July-2020].

[22] D. McEnnis, C. McKay, I. Fujinaga, P. Depalle, "jaudio: An feature extraction library.," in *Proceedings of the 6th International Conference on Music Information Retrieval, ISMIR 2005*, London, UK, 01 2005, pp. 600–603.

[23] J. L. Zhang, X. L. Huang, L. F. Yang, Y. Xu, S. T. Sun, "Feature selection and feature learning in arousal dimension of music emotion by using shrinkage methods," *Multimedia systems*, vol. 23, no. 2, pp. 251–264, 2017, doi: 10.1007/s00530-015-0489-y.

[24] P. Zachar, R. Ellis, *Categorical versus dimensional models of affect: A seminar on the theories of Panksepp and Russell*. John Benjamins Publishing Company, 2012.

[25] E. Schubert, "Update of the hevner adjective checklist," *Perceptual and motor skills*, vol. 96, no. 3, pp. 1117–1122, 2003, doi: 10.2466/pms.2003.96.3c.1117.

[26] J. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[27] A. Tellegen, D. Watson, L. A. Clark, "On the dimensional and hierarchical structure of affect," *Psychological Science*, vol. 10, no. 4, pp. 297–303, 1999, doi: 10.1111/1467-9280.00157.

[28] R. E. Thayer, "Toward a psychological theory of multidimensional activation (arousal)," *Motivation and Emotion*, vol. 2, no. 1, pp. 1–34, 1978, doi: 10.1007/BF00992729.

[29] Y.-H. Chin, C.-H. Lin, E. Siahaan, I.-C. Wang, J.-C. Wang, "Music emotion classification using double-layer support vector machines," in *Proceedings of the 1st International Conference on Orange Technologies (ICOT 2013)*, 2013, pp. 193–196.

[30] J. Deng, *Emotion-based music retrieval and recommendation*. PhD dissertation, Hong Kong Baptist University, 2014.

[31] N. Nalini, S. Palanivel, "Music emotion recognition: The combined evidence of mfcc and residual phase," *Egyptian Informatics Journal*, vol. 17, no. 1, pp. 1–10, 2016, doi: https://doi.org/10.1016/j.eij.2015.05.004.

[32] R. Panda, R. Malheiro, R. P. Paiva, "Novel audio features for music emotion recognition," *IEEE Transactions on Affective Computing*, vol. Published, 2018, doi: 10.1109/TAFFC.2018.2820691.

[33] P. F. Vale, "The role of artist and genre on music emotion recognition," Master's thesis, Information Management School, 2017.

[34] Y. Ospitia-Medina, J. R. Beltrán, S. Baldassarri, "Emotional classification of music using neural networks with the mediaeval dataset," *Personal and Ubiquitous Computing*, vol. April (online), pp. 1–13, 04 2020, doi: 10.1007/s00779-020-01393-4.

[35] M. Rumiantcev, O. Khriyenko, "Emotion based music recommendation system," in *Proceedings of the 26th Conference of Open Innovations Association FRUCT*, Yaroslavl, Russia, 2020, pp. 639–645.

[36] M.-C. Chiu, L.-W. Ko, "Develop a personalized intelligent music selection system based on heart rate variability and machine learning," *Multimedia Tools and Applications*, vol. 76, pp. 15607––15639, 09 2016, doi: 10.1007/s11042-016-3860-x.

[37] K.-A. Bodarwé, J. Noack, P. Jean-Jacques, "Emotion-based music recommendation using supervised learning," in *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia*, New York, NY, USA, 2015, pp. 341–344, Association for Computing Machinery.

[38] F. Paolizzo, N. Pichierri, D. Casali, D. Giardino, M. Matta, G. Costantini, "Multilabel automated recognition of emotions induced through music," *CoRR*, vol. abs/1905.12629, 2019.

[39] J. H. Juthi, A. Gomes, T. Bhuiyan, I. Mahmud, "Music emotion recognition with the extraction of audio features using machine learning approaches," in *Lecture Notes in Electrical Engineering. Proceedings of ICETIT 2019, Emerging Trends in Information Technology*, vol. 605, 2020, pp. 318–329, Springer International Publishing.

[40] K. W. Cheuk, Y.-J. Luo, B. B, T, G. Roig, D. Herremans, "Regression-based music emotion prediction using triplet neural networks," in *Proceedings of the International Joint Conference on Neural Network, IJCNN*, Glasgow, 07 2020, IEEE.

[41] A. Ma, I. Sethi, N. Patel, "Multimedia content tagging using multilabel decision tree," in *Proceedings of the 11th IEEE International Symposium on Multimedia*, 2009, pp. 606–611.

[42] S. Das, S. Debbarma, B. Bhattacharyya, "Building a computational model for mood classification of music by integrating an asymptotic approach with the machine learning techniques," *Journal of Ambient Intelligence and Humanized Computing*, vol. May (online), pp. 1–13, 05 2020, doi: 10.1007/s12652-020-02145-1.

[43] R. Panda, R. P. Paiva, "Music emotion classification: Dataset acquisition and comparative analysis," in *15th International Conference on Digital Audio Effects, DAFx-12*, 10 2012, pp. 1–7.

[44] R. Panda, *Emotion-based Analysis and Classification of Audio Music*. PhD dissertation, Universidade de Coimbra, 2019.

[45] E. Georgieva, M. Suta, N. Burton, "Hitpredict: Predicting hit songs using spotify data," 2018. [Online; accessed 19-July-2020].

[46] M. Sciandra, I. Spera, "A model based approach to spotify data analysis: A beta GLMM," *SSRN Electronic Journal*, vol. 3, pp. 1–18, 01 2020, doi: 10.2139/ssrn.3557124.

[47] J. H. Oh, S. Ouwejan, S. T. Kim, I. Ng, "Music intelligence: Granular data and prediction of top ten hit songs," *SSRN Electronic Journal*, pp. 1–12, 05 2020, doi: 10.2139/ssrn.3585176.

[48] R. Oi, "Spotify on the streets: walking and listening to music in urban spaces," Master's thesis, Lundt University, 2019. http://lup.lub.lu.se/student-papers/ record/8976269.

[49] M. Dittenbach, R. Neumayer, A. Rauber, "Playsom: An alternative approach to track selection and playlist generation in large music collections," in *Proceedings of the Workshop of the EU Network of Excellence DELOS on Audio-Visual Content and Information Visualization in Digital Libraries (AVIVDiLib 2005)*, 2005, pp. 226–235.

[50] A. Germain, J. Chakareski, "Spotify me: Facebook-assisted automatic playlist generation," in *IEEE 15th International Workshop on Multimedia Signal Processing (MMSP 2013)*, Sep. 2013, pp. 25–28.

[51] M. Pichl, E. Zangerle, G. Specht, "Combining spotify and twitter data for generating a recent and public dataset for music recommendation," in *Proceedings of the 26th GI-Workshop Grundlagen von Datenbanken (GvDB 2014)*, Ritten, Italy, 2015, pp. 35–40.

[52] F. Fessahaye, L. Pérez, T. Zhan, R. Zhang, C. Fossier, R. Markarian, C. Chiu, J. Zhan, L. Gewali, P. Oh, "Trecsys: A novel music recommendation system using deep learning," in *2019 IEEE International Conference on Consumer Electronics (ICCE)*, 2019, pp. 1–6.

[53] M. Madathil, "Music recommendation system spotify - collaborative filtering," 2017. Reports in Computer Music. Aachen University, Germany.

[54] N. F. R. Fauzia, "The use of song titles in spotify playlists to express the affection," in *International Seminar on Sociolinguistics and Dialectology: "Changes and Development of Language in Social Life" 2017*, 2017, pp. 185–189.

[55] S. Giammusso, M. Guerriero, P. Lisena, E. Palumbo, R. Troncy, "Predicting the emotion of playlist using track lyrics," in *19th International Society for Music Information Retrieval Conference*, Paris, France, 2018.

[56] G. Subramaniam, J. Verma, N. Chandrasekhar, K. Narendra, K. George, "Generating playlists on the basis of emotion," in *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2018, pp. 366–373.

[57] H. Abderrazik, G. Angela, H. Brouwer, H. Janse, S. Lutz, G. Smitskamp, S. Manolios, C. C. S. Liem, "Spotivibes: Tagging playlist vibes with colors," in *Proceedings of the 6th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems co-located with 13th ACM Conference on Recommender Systems, RecSys 2019*, vol. CEUR, 2450, 2019, pp. 55–59.

[58] P. Helmholz, M. Meyer, S. Robra-Bissantz, "Feel the moosic: Emotion-based music selection and recommendation," in *32nd Bled eConference: Humanizing Technology for a Sustainable Society*, Bled, Slovenia, 06 2019, pp. 203–221.

[59] *Spotify* for developers, "*Spotify* web api." https://developer.spotify.com/documentation/web-api/, 2020. [Online; accessed 19-July-2020].

[60] M. D. Team, *Microsoft Application Architecture Guide, 2nd Edition (Patterns & Practices)*. Wiley, 2009.

[61] L. Liu, M. T. Özsu Eds., *n-Tier Architecture*, pp. 1924–1924. Springer US, 2009.

[62] T. M. project, "AcousticBrainz." http://acousticbrainz.org/, 2015. [Online; accessed 19-July-2020].

[63] *Spotify* for developers, "Description of the *Audio Feature Object*," 2020. [Online]. Available: https://developer.spotify.com/documentation/web-api/reference/#objectaudiofeaturesobject, [Online; accessed 26-July-2021].

[64] R. Panda, R. Malheiro, R. P. Paiva, "Novel audio features for music emotion recognition," *IEEE Transactions on Affective Computing*, vol. Published, 2018, doi: 10.1109/TAFFC.2018.2820691.

[65] C. Laurier, M. Sordo, J. Serrá, P. Herrera, "Music mood representations from social tags," in *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, Kobe, Japan, 2009, pp. 381–386.

[66] K. Trohidis, G. Tsoumakas, G. Kalliris, I. Vlahavas, "Multi-label classification of music by emotion," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2011, p. 4, Sep 2011, doi: 10.1186/1687-4722-2011-426793.

[67] J. Bergstra, Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, Feb. 2012.

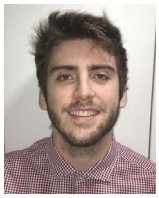
[68] P. Desmet, M. Vastenburg, V. Bel, D., N. Romero, "Pick-a-mood; development and application of a pictorial mood-reporting instrument,"

in *Proceedings of the 8th International Conference on Design and Emotion: Out of Control - Proceedings*, 09 2012, pp. 1–12.

[69] F. Buschmann, R. Meunier, H. Rohnert, P. Sommerlad, M. Stal, *Pattern-Oriented Software Architecture, Volume 1, A System of Patterns*. Wiley, 1996.

Pedro Álvarez

Pedro received the Ph.D. degree in computer science engineering from the University of Zaragoza, Zaragoza, Spain, in 2004. He works as Lecture Professor at this University, since 2000. His current research interests focus on two main aspects. First, on integration problems of network based systems and the use of novel techniques and methodologies for solving them. And, secondly, on the application of formal analysis techniques and artificial intelligence techniques to extract knowledge from logs, databases and/or IoT systems.



Jorge García de Quirós

Jorge received the B.E. degree in computer science engineering from the University of Zaragoza, Zaragoza, Spain, in 2018. He works as Researcher at this University, since 2018. His current research interests focus on affective computing and music emotion recognition. He has also interest in other topics as cybersecurity and sport data analysis.

Sandra Baldassarri

Sandra received the Ph.D. in Computer Science Engineering from the University of Zaragoza, Spain, in 2004. She is Associate Professor in Computer Science Department at the University of Zaragoza (Spain) and founder member of the AffectiveLab Research Group and member of the Engineering Research Institute of Aragon (I3A), both at the University of Zaragoza. Her research interests include affective computing, multimodal interfaces, tangible and natural interaction, virtual humans and their application in educational fields. In these areas she published a numerous papers in conferences and journals and participates as part of the scientific and organizer committees of several Human Computer Interaction national and international conferences.

Rhetorical Pattern Finding

Francisco Gómez[1]*, Manuel Tizón[2], Aitor Arronte[3], Víctor Padilla[2]

[1] Universidad Politécnica de Madrid (Spain)
[2] Universidad Internacional de La Rioja (Spain)
[3] University of Hawaii at Manoa (USA)

Received 30 June 2021 | Accepted 6 May 2022 | Published 3 October 2022

Abstract

In this paper, we research rhetorical patterns from a musicological and computational standpoint. First, a theoretical examination of what constitutes a rhetorical pattern is conducted. Out of that examination, which includes primary sources and the study of the main composers, a formal definition of rhetorical patterns is proposed. Among the rhetorical figures, a set of imitative rhetorical figures is selected for our study, namely, epizeuxis, palilogy, synonymia, and polyptoton. Next, we design a computational model of the selected rhetorical patterns to automatically find those patterns in a corpus consisting of masses by Renaissance composer Tomás Luis de Victoria. In order to have a ground truth with which to test out our model, a group of experts manually annotated the rhetorical patterns. To deal with the problem of reaching a consensus on the annotations, a four-round Delphi method was followed by the annotators. The rhetorical patterns found by the annotators and by the algorithm are compared and their differences discussed. The algorithm reports almost all the patterns annotated by the experts and some additional patterns. The algorithm reports almost all the patterns annotated by the experts (recall: 98.11%) and some additional patterns (precision: 71.73%). These patterns correspond to rhetorical patterns within other rhetorical patterns, which were overlooked by the annotators on the basis of their contextual knowledge. These results pose issues as to how to integrate that contextual knowledge into the computational model.

Keywords

Computational Models for Rhetorical Patterns, Delphi Method, Imitative Patterns, Musicological Models for Rhetorical Patterns, Musical Patterns, Pattern Annotation, Pattern Finding Algorithms, Renaissance Music, Rhetorical Patterns, Victoria's Masses.

DOI: 10.9781/ijimai.2022.10.002

I. Introduction

In this work, we study rhetorical patterns, in particular rhetorical patterns by repetition, from a musicological and computational standpoint. This kind of pattern is associated to highly imitative music of contrapuntal nature in the Renaissance and Baroque eras. To the best of our knowledge, rhetorical patterns have received very little attention in the modern musicological literature. The contribution of this research follows a path that starts by establishing a conceptual framework for the definition and characterization of rhetorical patterns, which is a question in need of further research in musicology; see Section II. After that, issues have arisen when annotating these patterns —including the problem of subjectivity and consensus-reaching—, are examined. Manual annotations are all at the same time valuable, cognitively demanding, time-consuming, and prone to error. Furthermore, when a corpus of music is relatively large, manually detecting all the patterns is not feasible and is often unrealistic. Therefore, in this study, we also took into account those patterns that the experts overlooked in their annotations. In our case, manual annotations by experts were needed to understand better the nature of rhetorical patterns as well as to provide ground truth to test the pattern-discovery algorithm. Moreover, when a set of experts annotate some music, the problem of reaching consensus among them soon arises. In order to tackle this problem, a consensus-reaching scheme, called Delphi, was implemented. Both the issues of manual annotations and consensus-reaching are addressed in Section III.A. We also built a computational models for the rhetorical patterns, which included formal computational definitions of the rhetorical patterns (Section II.B). To automatically extract and mine rhetorical patterns, we used an existing pattern-finding algorithm, the BIDE algorithm. We then proceeded to test our method on the work of Tomás Luis de Victoria (details in Section II.C) by determining whether the category of a discovered pattern matches the category assigned by annotators. A classification task —whether a discovered pattern is an annotated pattern—was set up and performance measures were computed and analyzed (Section IV). It turned out that the algorithm predicted the correct category for most of the annotated patterns (recall: 98.11%), but interestingly enough, it also found additional patterns that were overlooked by the experts (precision: 71.73%). Lastly, the paper comes to an end where conclusions about the results obtained in this research are discussed.

Musicologists and other musical scholars can benefit from this work in many ways. First, we proposed a conceptual framework for rhetorical patterns. That framework unifies different definitions given in the literature (see Section II.B) and proposes a rigorous definition of the main rhetorical patterns found in the Renaissance and Baroque music. Second, to have the possibility of searching for rhetorical patterns in large corpora of music is highly advantageous. Manual annotation of patterns is a prone-to-error and tedious task. As we will see in this paper, algorithms can detect some patterns that are diÿcult to detect by the experts. Third, we made use of a novel method to

* Corresponding author.
E-mail address: francisco.gomez@upm.es

reach consensus among experts, the Delphi method. To the best of our knowledge, such method have hardly been used in research involving consensus-reaching in musicology. Finally, the results obtained in this work open new venues for further research, especially the sub-pattern problem (the problem of having nested patterns; see Section II.B).

II. Rhetorical Patterns in Music

A. Musical Patterns

Patterns provide the musical discourse with both meaning and structure. In terms of meaning, they bear a resemblance to phonemes, the smallest units of speech that perceptually distinguish one word from another, as patterns can broadly be defined as the shortest meaningful sequences in a piece of music. Indeed, patterns carry enough relevant information to create musical statements or advance musical arguments on their own. In terms of structure, through purposeful repetition, they help construct musical syntax throughout the piece. Surprisingly enough, an agreed-upon definition of musical patterns has proved elusive. Some authors define musical patterns as sequences of notes that are repeated at least twice in the piece; for example, that is the case of the MIREX 2013 Repeated Themes & Sections task [1]; see also [2]. However, from a conceptual and musical standpoint this definition seems to be somewhat insuÿcient. Much research into musical pattern finding merely relies on repetition to find patterns in musical corpora. The rationale behind that strategy is that if a sequence of musical events is repeated, it should be because such sequence has musical entity and therefore can be classed as a musical pattern. Many pattern-finding algorithms limit themselves to track down the most frequent sequences in the music without examining the qualities of the sequence or where their occurrences appear in the piece. However, not every repeated sequence of musical events is necessarily a pattern. The famous opening theme in the first movement of Tchaikovsky's first Piano Concerto is clearly a pattern, which is remembered by all attentive listeners, but is never repeated again and certainly is not developed. It simply stands by itself as a flamboyant statement bearing no relationship to the rest of the movement. Contrariwise, the initial pattern in the first movement of Beethoven's Fifth Symphony is the backbone of the whole movement, where that pattern is developed and transformed ad infinitum.

What is then a musical pattern? Other authors ([3], [4]) use more or less vague definitions such as a musical patterns are "perceptible repetitions in a musical piece," or "an excerpt of special importance," or "a salient fragment," or "a prominent unit." For a sequence to be considered a musical pattern it needs to possess some extra attributes. A musical pattern is here defined as a musical event that constitutes a musical whole and is repeated as to create structure in a musical piece. Gestalt theory can shed light over the precise mechanisms underlying the formation of musical patterns through rules of proximity, similarity, and good continuation. For example, Deutsch [5] describes several perceptual mechanisms operating in the construction of a musical pattern as a musical whole, such as grouping of musical sounds —grouping by pitch proximity, grouping by timbre, grouping by temporal proximity, among others—, whereas Lartillot [6] identifies musical patterns in terms of style-based groupings, local boundaries, and repetition (his work in turn builds on the generative theory of Lerdahl and Jackendoff [7]). In general, the listener uses tonal-temporal hierarchies to combine notes to form patterns. Other important efforts to characterize wholeness in musical sequences are represented by Meyer [8], Huron [9], Temper [10] (theories of expectation); Lerdahl and Jackendoff [7] (generative theory of tonal music); Narmour [11](implication-realization theory); and Margulis [12] (musical tension models). Ultimately, all these mechanisms are Gestalt- and culture-based mechanisms. For an excellent account on these perceptual mechanisms, see Oxenham [13] as well as Deutsch [14] in the book *The Psychology of Music*.

The other important feature in the definition of musical pattern is repetition. How often and where in the piece a pattern is repeated definitely shapes the perception by the listener. This is what Nattiez [15] (and also Lartillot [16]) calls syntagmatic relations, which just refer to the syntax of the pattern in the piece. By musical syntax here we refer to the order in which musical elements appear in a given piece. Margulis [17], one of the few authors who has examined musical repetition in depth, states that, "music's repetitiveness is at once entirely ordinary and entirely mysterious." Repetition in language as it occurs in music would be deemed unacceptable. However, repetition in music is key as it greatly contributes to creating meaning and structure. Repetition also plays an important role in creating, denying or delaying musical expectations, which is a mechanism to produce musical meaning. Margulis [17] highlights three primary roles in repetition in music: (1) learning and level-shifting, (2) segmentation, and (3) expectation. It should be noted at this point that by repetitions we do not mean literal repetitions. Two sequences that bear certain similarities can be considered as the same sequence; for example, a sequence transposed by a fixed interval results in a similar sequence to the original one. Other perceptual pattern-preserving operations are changes of voice, change of tempo, and minor changes of duration, contour, or pitch. The extent to which two patterns can be considered the same strongly depends on the music style under study. Rolland [18] terms this relation between patterns equipollence (he claims it is more general than a similarity relation).

Notwithstanding the fact that rhetorical patterns may present musical wholeness and a certain degree of structural repetition on their own, their musical meaning stems from extramusical reasons, provoking emotions and conveying textual meanings in the musical discourse being the main ones. Therefore, we only have to look for repetitions as musical wholeness is given by the very nature of the rhetorical pattern.

Determining the importance of a musical pattern is also a significant issue in the computational analysis of music data, since the mining of frequent patterns leads to combinatorial problems [19], [20]. Closed patterns provide a succinct and robust method of reducing the existent redundancy of the set of frequent patterns by discovering a small subset of it. A pattern is said to be closed if it is not included in a larger pattern that occurs with the same frequency. As we can see by this simple definition, closed patterns reduce the search space of frequent patterns by providing at the same time a more compact representation. In this sense, [21] evaluates sequential pattern mining on a corpus of Mozarabic chant neume sequences. Their results indicate that it is possible to retrieve all known patterns with an acceptable precision using significant closed pattern discovery. In musical terms, and more specifically in this work, a closed pattern will help determine patterns that are associated with rhetorical figures.

B. Rhetorical Patterns

Among the many definitions of rhetoric in language, a very concise one is provided by Quintilian [22], who states that "rhetoric is the science of speaking well, [...] the science of correct expression." From this definition, the classical authors established three purposes of rhetoric, namely, to inform, to persuade or move, and to entertain or delight. According to the classical authors, rhetoric can be further divided as follows [23]: from the composer standpoint, invention (*inventio*), arrangement (*dispositio*), elaboration and decoration (*elocutio*); from the performer standpoint, delivery (*pronunciatio*), memory (*memoria*), and gesture (*actio*). The *Harvard Dictionary of Music* in turn defines rhetoric in music [24] as follows.

> The principles governing the invention, arrangement, and elaboration of ideas in a piece of music. Drawing on classical models of oration, music

theorists cultivated the concept of musical rhetoric in earnest during the 16th, 17th, and 18th centuries, especially for works with texts. Such activity blossomed into the so-called doctrine of figures and doctrine of affections. Although rhetorical models for music were supplanted in the 19th century, they continue to influence the various modes of musical analysis, whether or not the music in question is based on an explicit text or program.

Although similarities between both definitions can be noted, the relationship between rhetoric in language and rhetoric in music is a very complicated issue. The very analogy between music and language is already a delicate matter. We find many sources in the history of music that shed light on this connection. In the German context, theorists such as Calvisius, Lippius, Nucius, Thuringus, Kircher or Mathesson are studied and mentioned in several rhetorical analyses in contemporary sources [23], [25], [26]. The reason for focusing on German authors is their penchant for theorizing rhetorical figures of speech, contrary to what happens in Italy, where theorists do not explain these figures and just apply them in their musical compositions. During the 16th century, the discovery and absorption of the ideas in the classical rhetoric texts resulted in a new expresiveness that permeated the language and all the arts, including music. The influence of rhetoric brought about a new consciousness of musical expression. The rhetoric discourse in music has been especially important between the 16th and 18th centuries. Furthermore, in this research we mainly focus our attention on imitative figures. Most figures in musical rhetoric are divided into those of development by repetition, comparison or amplification, and surprise [23]. After combing through the relevant literature, we chose to follow López Cano [26] and Bartel [25] as our sources for defining the rhetorical figures included in this study. From a systematic musicology standpoint, little attention has been devoted to rhetorical patterns in the past few years. A notable exception is Parada-Cabaleiro [27], who studied *madrigalisms*, a composition technique that mimics the linguistic content of the lyrics.

In general, in order to determine if two sequences are the same, rhythmic and intervallic information have to be taken into account. Given the highly imitative nature of the rhetorical patterns, rhythmic information was not necessary to detect the rhetorical patterns in the corpus. In fact, due to the flexible use of the rhythm in the Renaissance period, it is very complicated to establish a rule to identify each different case (augmentations, diminutions, lengthening or shortening patterns). Therefore, only intervallic patterns will be considered here. Thus, two sequences will be considered the same if their generic-interval sequences are identical (this includes length, order, and contour of intervals in the pattern). By generic interval we refer to the interval category, e.g., third or sixth, without considering the interval quality, e.g., major or minor. That is, a minor third and a major third are the same generic interval of a third. This definition not only encompasses exact repetitions of a pattern, but also its transposed versions where modality can vary. Two additional constraints were placed to define and detect rhetorical properly:

1. The frequency of the repetition of a pattern has to be greater than 3 repetitions per work, in any voice.
2. The minimum pattern length considered has to be 3 notes. The selected figures of speech and their definitions are given below.

- **Epizeuxis**: an immediate and emphatic repetition of a pattern in any voice, which normally is short and insistent [25]. We added "short and insistent" to clarify the difference among other figures. If we draw an analogy with the speech, a repetition of a name would fit with the concept; i. e., *John, John, John, I need your help.* Since we always find this figure as a short pattern, usually in stretto, and very insistent in several voices, we have to lay down some criteria to properly identify this figure:

  (1) The maximum length of this figure is the equivalent to four onsets.

  (2) The maximum separation distance among the patterns will not be larger than half a bar. Otherwise, the insistence effect would be lost.

  (3) The patterns may or may not overlap in time.

  In the excerpt from the "Kyrie" of *Ave Maris Stella* mass shown in Fig. 1, we find an instance of an epizeuxis where a 3-note pattern is repeated throughout two voices and with less than half a bar apart between consecutive repetitions of the pattern. The two first patterns overlap in time; the second and the third do not.

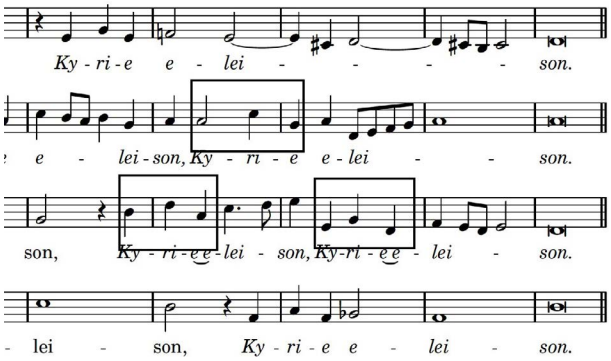


Fig. 1. Example of epizeuxis ("Kyrie" of *Ave Maris Stella*, bars 18-22).

- **Palilogy**: a repetition of a pattern transposed by an interval other than unison or octave in other voices [25]. Fig. 2 shows an instance of palilogy (beginning of the "Kyrie" of *Ave Maris Stella*).

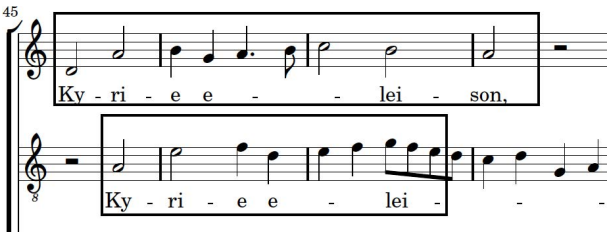


Fig. 2. Example of palilogy ("Kyrie" of *Ave Maris Stella*, bars 1-4).

- **Synonymia**: repetition of a pattern transposed by some interval in the same voice [26]. In Fig. 3 on next page, we can see a synonymia, since the imitation coincides in the same voice in different pitches; in this example two synonymias are shown. The excerpt below belongs to the 4-voice "Kyrie" of *Missa pro Defunctis* mass.
- **Polyptoton**: repetition of a pattern in other voices either in unison or at the octave [26]. This author—following other authors such as Arnold Shering, Hans-Heinrich Unger, and George Buelow—only mentions unison. We expanded the term to the octave; in vocal music it is very common to transpose patterns an octave up or down to adapt them to the vocal range. As we can see in Fig. 4 on next page, the same pitches are replicated in another voice at a higher octave (notice the suboctave clef); therefore, we are dealing with a polyptoton. This example has been selected from the "Kyrie" of *Ave Maris Stella* mass.

Notice that in some cases a rhetorical pattern of one kind may contain patterns of other kinds as sub-patterns. For instance, a palilogy may contain other shorter palilogies that are more frequent than the palilogy itself. This will be termed as the sub-pattern problem and it will be discussed in the result section. Furthermore, the rhetorical categories epizeuxis and palilogy are not mutually exclusive. A palilogy may hold hold conditions (1) and (2) in the definition of epizeuxis and therefore it can in fact be an epizeuxis. This situation will arise in the results of the experiments as we will see later.

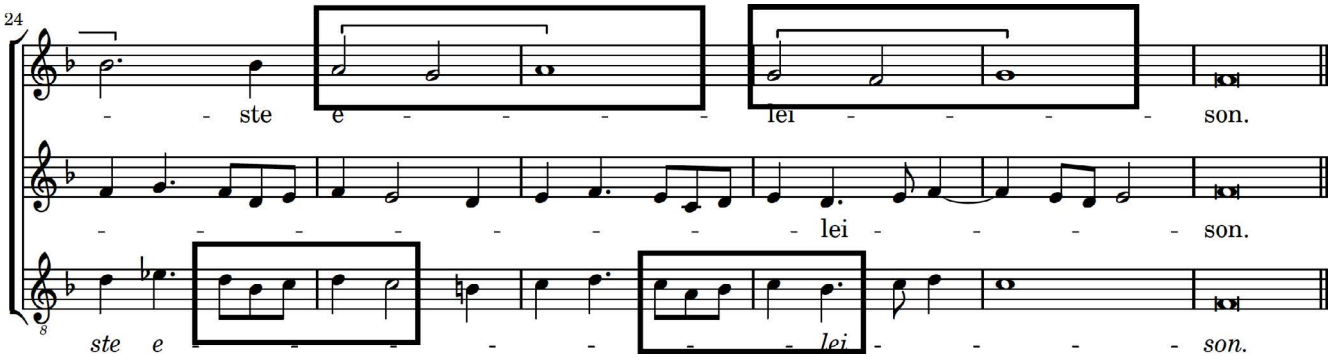


Fig. 3. Example of synonymia ("Kyrie" of Missa pro Defunctis mass, bars 24-28).

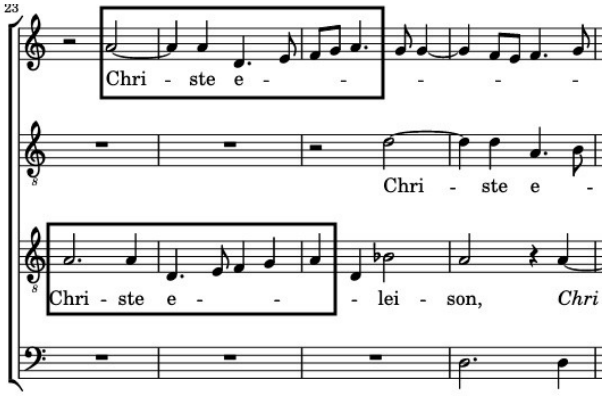


Fig. 4. Example of a polyptoton ("Kyrie" of Ave Maris Stella mass, bars 23-26).

C. A Musical Corpus of Renaissance Music

A corpus from the masses of Tomás Luis de Victoria (1548–1611) was chosen to analyze rhetorical patterns. We took 4-, 5-, and 6-voice masses in order to facilitate the data extraction process (more than six voices could have further complicated the process). Victoria is one of the main representative composers in the music of the Spanish Golden Age. With the advent of the Protestant Reformation, a Counter-Reformation is born, the Council of Trent being a pivotal event. This ecumenical council will have an important implication in music composition. Victoria's music perfectly reflects the tenets of this Council, whose musical characteristics are the following:

- The intelligible presentation of lyrics;
- The elimination of profane elements;
- Uncomplicated and understandable counterpoint, and hence, melodic lines that are streamlined and easy to remember.

As we can infer, musically speaking, text intelligibility was one of the most important issues in the Council of Trent. For that reason, the counterpoint is at the service of the text, the melodic lines are very clear, text is understandable, and the imitation is very balanced.

In addition, Victoria studied in Rome, and as stated by Wagstaff [28], it is likely that he met Palestrina in that period (from 1563-65 to 1587). But although his music is clearly influenced by Palestrina's style, Victoria's compositions are very personal and underwent a constant evolution throughout his lifetime. For instance, his mass *Ave maris stella* displays concise musical ideas, which fits with the spirit of the Council of Trent. By way of contrast, the mass *Salve regina* uses compositional procedures close to music of Giovanni Gabrieli or Monteverdi [28].

Victoria's music is characterized by imitative structures in melodic lines. Many of his compositions fit within the most popular musical settings in the Renaissance: cantus firmus, paraphrase, and parody, which are based on elaborating new compositions from pre-existent material [29]. Examples of these procedures in his works and in other Spanish composers (Francisco Guerrero and Cristobal de Morales) can be perused in Stevenson [30].

In connection with this paper, Victoria's discourse fits in the rhetorical spirit of the Renaissance era. As a matter of fact, not all figures of speech can be found in his music, we have to think that the limitations of Catholic music in this period restricted many expressive procedures, such as the treatment of consonance and dissonance. We thus contend that repetitive figures are very consistent in Victoria's music, which was the main reason to choose him for our research.

As for the corpus used in this study, a selection of four masses was made taking into consideration their relevance in the composer's opus and the time span from the earliest composed work to the latest. The masses are *Dum Complerentur*, *Ascendens Christus*, *Gaudeamus*, and *Veni Sponsa Christi*. The transcriptions were made by musicologist Nancho Álvarez and the MIDI files were taken from his web page [31]. Nevertheless, his transcriptions were checked against the original manuscripts.

III. Rhetorical Pattern Finding

A. Musical Annotations and Assessment

The methodology presented here is of deductive nature, that is, the rhetorical patterns were defined on an abstract level. To test the goodness of our definition, we set up a classification task consisting of automatically searching for those patterns in the Victoria corpus. For both tasks —providing a conceptual definition of rhetorical patterns and find them in the corpus— careful assessment was required. In order to deal with both issues, we had a small set of experts at our disposal. Five experts with different musical background, from musicologists or composers to performers, formed the group of annotators. Their task was to look at the four Victoria's masses and extract the rhetorical patterns by following the definitions provided in Section II.B. The patterns extracted by the experts would be compared to those extracted by the BIDE algorithm. As a matter of fact, not all the experts agreed upon the annotations and much heated and in-depth discussion broke out. Part of the objections raised by the experts led to a refinement of both our definition of rhetorical patterns and the computational model. A delicate issue was how to deal with the consensus-reaching problem. In order to solve that problem, we employed the Delphi method.

The Delphi method has been extensively used in several fields of health research [32] and hospitality sector [33]. This method is a systematic way to achieve consensus by a group of experts in subjects where the results are not so obvious (as it is our case). In fact, expert consensus is the ground of science in important decisions such as funding applications, publications, or different metrics of citation. Consensus methods are mainly based on the idea contained in James Surowiecki's book *The Wisdom of Crowds* [34]. This author presents a few ideas to successfully devise a consensus-reaching scheme:

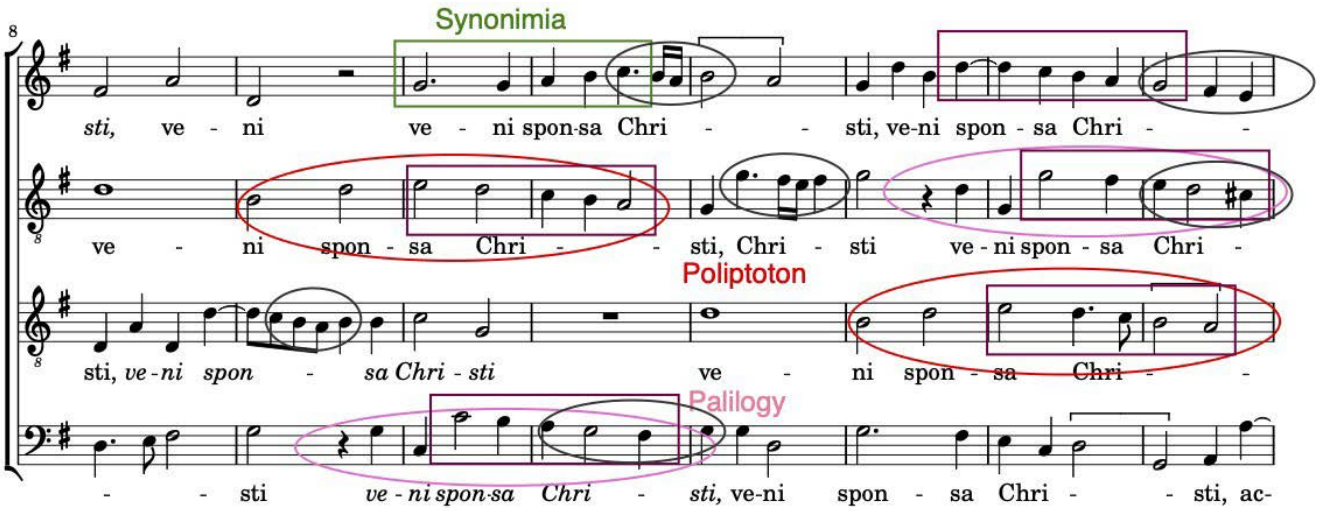


Fig. 5. Annotations of the rhetorical patterns by the experts.

TABLE I. Delphi Rounds for Consensus-Reaching

Delphi rounds	Rhetorical pattern model	Ground truth
Round 1	• Model with many low-level features both in rhythm and pitch • Based on the idea of main note • Based on changes of melodic contour • Incorporated a melodic similarity measure • Some features were not computational	• The annotations presented marked levels of discrepancy • Issues: closed patterns, overlapping patterns, different definitions of rhetorical patterns in the sources, selection of the final patterns • Experts selected patterns, but were not able to formalize their choice
Round 2	• The model is simplified, but still is low-level oriented • Still based on melodic contour and similarity measure	• Fewer differences in annotations than in the first round • Consensus on the sources and definitions grow • The issues of closed patterns and overlapping patterns persist
Round 3	• Low-level oriented model is abandoned in favor • of a simpler model • Closed patterns with constraints are introduced	• Consensus on the sources and definitions is reached • Annotators justify better their choices • Differences in annotations persist in a few cases
Round 4	• The model is only • based on intervallic content • Rhythm is not taken into account • The model is fully computational	• Very high degree of consensus on the annotations • Realization that intervallic content suffices for this study

diversity of expertise, independence of experts, de-centralized work, and a mechanism for aggregating different ideas. The Delphi method works as follows: (1) A person (the facilitator) organizes the study and recruits some experts; (2) The facilitator prepares a questionnaire and collects the responses from the group of experts; (3) The facilitator provides the experts with anonymous feedback and they review their responses; (4) After a few rounds, normally 3 or 4, the group of experts arrives at a reasoned consensus. The Delphi method has started to be used in areas other than health research. Romero [35] used it to agree upon a definition of a good musician among different experts. In education it is also taking on; see Green [36] for more information.

As we pointed out earlier, in our work, the Delphi method had two uses: (1) gather knowledge and insight from the experts to build the formal definitions of rhetorical patterns; (2) to build a ground truth for validating the pattern-finding algorithm used in our experiments. The number of Delphi rounds carried out was four. One of the authors of this paper acted as the facilitator. The experts annotated the scores in staff notation through a color and number system to keep track of the patterns. After annotating the scores, the experts would write a report where they pointed out to discrepancies or issues found during the process. In Fig. 5, an excerpt of a score annotated by an expert is shown. As the reader can see in the figure, the rhetorical patterns can overlap.

There were two main lines of inquiry at the Delphi rounds, namely, development of the rhetorical pattern model and establishing the ground truth. The main issues encountered at the beginning of the Delphi rounds were: the model was based on low-level features that in many cases could not be described in computational terms; there were discrepancies among the experts in the definition of the rhetorical patterns; experts experienced diÿculties at formalizing their annotations; it was not clear how to deal with overlapping patterns. During the rounds, these issues were resolving gradually. Closed patterns were incorporated to the model (the initial definition of pattern was not fully versatile), the equality between sequences was defined just in terms of their generic intervallic content, and the model was improved and the experts reached a high level of consensus. The resulting model was presented to the annotators and discussed with them. As a matter of fact, it was a parallel process. Another issue concerning the assessment of the algorithm was the number of pieces involved. Although four masses can seem a small number, it turns out that the number of patterns within them was high enough for our purposes. This issue was also discussed with the experts, who agreed that the number of masses was adequate. In the results section the annotations by experts are compared with the patterns mined by the algorithm[1]. Table I presents a summary of the four rounds carried out and the main points under discussion. Referring to the table, a main

[1] The reader can find an annotation of one of the sections of Victoria's masses in https://www.dropbox.com/s/ysmuc3im34rqylo/Rhetorical-patterns-Annot.pdf?dl=0. This was one of the annotations used as ground truth in this study.

note is a note that is more prominent that others, which can be due to melodic, rhythmic or harmonic reasons, among others. By very high degree of consensus, here we mean that at least four experts agreed on the annotations. However, any source of disagreement was analyzed and discussed by the experts and often there was unanimity.

B. Computational Rhetorical Pattern Finding

As stated at the outset, in this section will consider the problem of building a computational model for finding rhetorical patterns. A computational model is a description of a phenomenon, in our case rhetorical pattens, given in terms understandable by a computer. The first component of our model consists of the encoding of the musical piece in the corpus. The corpus was encoded by employing a string representation of intervals given in the software Music21 [37], and the chromatic pitch values from the MIDI files of the pieces. Each voice is divided into phrases by rests.

The second component is composed of the computational description of the definitions of rhetorical patterns given in Section II.B. Notice that all the features of the rhetorical patterns described in there can be expressed in computational terms. The third component is the concept of closed pattern, which allows eÿcient pattern-discovery in musical databases.

To computationally determine the rhetorical figure associated with a pattern P, we establish four categories that a pattern can belong to: epizeuxis (Ep), palilogy (Pa), synonymia (Sy), and polyptoton (Po). Recall that a pattern is a sequence whose frequency of occurrence is equal or greater than 3 and whose length is at least 3 notes. Their definitions are as follows:

- A pattern $P \in Ep$ if and only if there exists another sequence P' in any voice such that P and P' hold the following conditions: (1) P and P' have the same intervallic content; (2) the length of P is not greater than four onsets; (3) the maximum separation between P and P' is not greater than half a bar.
- A pattern $P \in Pa$ if and only if there exists another sequence P' in a different voice transposed by some interval other than the unison or the octave such that P and P' have the same intervallic content.
- A pattern $P \in Sy$ if and only if there exists another sequence P' in the same voice transposed by some interval other than the unison and the octave such that P and P' have the same intervallic content.
- A pattern $P \in Po$ if and only if there exists another sequence P' in a different voice transposed to an interval of unison or octave such that P and P' have the same intervallic content.

It could be expected that the rhetorical categories were disjoint. However, they are not due to the broad definition of the epizeuxis. Putting aside the length and separation of the patterns, in an epizeuxis, the patterns can be transposed by any interval and appear in any voice. This causes that palilogies, synonymias, and polyptotons may also be epizeuxis on certain occasions.

From the corpus of symbolic music encoded by using the string representation defined in this sub-section, we develop a method to obtain rhetorical patterns following the closed constrained patterns with the above-described definitions; such method is based on the BIDE algorithm [38] and a filtering mechanism. The BIDE algorithm is an eÿcient algorithm for mining frequent closed sequences without candidate maintenance. It prunes the search space more deeply compared to previous algorithms; see [38] and the references therein. The input parameters for the BIDE algorithm are the set of sequences and the minimum support threshold (minimum frequency of a pattern).

Our method works as follows:

- Given as input the set of intervallic sequences and a minimum support threshold, we obtain as output the set of all closed patterns **C** by applying the BIDE algorithm to the input.
- For each pattern in **C**, we test whether that pattern meets any of the rhetorical definitions and assign it to its corresponding categories. We thus obtain as output the set of all closed patterns that meet the rhetorical constraints.

For the sake of reproducibility of our experiments, the code is available at https://github.com/aitor-alvarez/PatternFinder.

C. Experiments

Our experiments consisted of running the BIDE algorithm on the corpus and assigning the closed patterns output by the algorithm to the rhetorical categories. This assignment was done by applying the formal computational definitions given in Section III.B. Table II lists the number of rhetorical patterns by category extracted by the experts and by the computational method proposed. In the case of palilogy and epizeuxis —since these categories are not mutually exclusive—, there were 8 patterns classed as both palilogy and epizeuxis. Those 8 patterns were not annotated by the experts. These patterns are all longer patterns that contained shorter patterns annotated by the experts.

TABLE II. Total Number of Rhetorical Patterns Extracted By Experts and by the Computational Approach Presented

	Experts	Algorithm
Palilogy	54	75
Epizeuxis	12	32
Synonymia	5	5
Polyptoton	35	33
Sum	106	145

Taking a closer look at Table II, we can see that the rhetorical figures obtained by the experts and the algorithm are exactly the same in the case of the synonymia. In the case of polyptoton, the experts found two additional patterns than the algorithm did not. Epizeuxis and palilogy show noticeable differences in number. For these two categories, the algorithm found more patterns than those annotated by the experts. Differences in polyptotons are associated with smaller length rhetorical figures such as epizeuxis, that are detected by the algorithm as closed patterns and selected in favor of longer polyptotons. This is indeed something that may require further discussion.

Closed patterns with constraints may capture sub-patterns that have a frequency of occurrence higher than the super-patterns that contains them. In the context of rhetorical figures such a definition does not always work. For instance, a palilogy may contain closed patterns that are more frequent than the palilogy itself, and in other type of musical analysis those closed patterns may be considered important motifs within a larger musical sequence or phrase. In the analysis of rhetorical patterns, such closed patterns that are incorrectly identified as epizeuxis, mostly because of the pattern length and how they overlap with each other. That is another reason to see sharp differences in the number of epizeuxis annotated by the experts and the ones discovered by the method proposed.

IV. Results

We defined a classification task by taking the annotations made by the experts as a gold standard. If a pattern was classed as in the annotation by the experts, then we considered the pattern as a correctly classified pattern. In Table III, metrics to measure the performance of the classification task are presented for both all the rhetorical patterns and by pattern individually. We denote by tp, tn, fp, and fn, true positives, true negatives, false positives, and false negatives, respectively. True positives are the annotated patterns

TABLE III. Overall Metrics and by Rhetorical Pattern

Metrics	Overall	Palilogy	Epizeuxis	Polyptoton	Synonymia
True positives tp	104	54	12	33	5
False positives fp	41	21	20	0	0
False negatives fn	2	0	0	2	0
Precision (%) = $\frac{tp}{tp+fp}$	71.73	72	37.5	100	100
Recall (%) = $\frac{tp}{tp+fn}$	98.11	100	100	94.28	100
F -score (%) = $2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$	82.87	83.72	54.54	97.05	100

correctly classified by the algorithm. Given the nature of this classification task, because the algorithm does not actually reject non-rhetorical patterns, there are no true negatives and, therefore, $tn = 0$ in all cases. False positives are those patterns identified as rhetorical patterns by the algorithm that were not annotated by the experts. Finally, false negatives are those patterns annotated by the expert that the algorithm failed to identify. In the last three rows, three metrics to assess the performance of the task were computed, namely, precision (positive predictive value), recall (true positive rate), and *F*-score.

Due to the high number of false positives in the case of palilogy and epizeuxis, precision is relatively low, especially for epizeuxis where there are many more false positives than true positives. Since there are no false negatives except for polyptoton, recall is 100%. For the polyptoton, the algorithm failed to find two annotated patterns and then recall is not 100%. The values for *F*-score are high, except in the case of epizeuxis, again due to the fact that the number of false positives was relatively high compared to the number of true positives.

An important type of implicit contextual knowledge that experts have and that the method presented in this article has problems with is identifying the palilogy. There are several instances where the algorithm labels patterns as palilogy and experts do not consider those patterns as rhetorical. These cases are related to the time of occurrence between two such patterns, that is not always clearly identifiable in a quantitative or constraint-based way. Experts tend to pay close attention to the development of musical phrases and detect rhetorical implications of patterns based on how music develops within a musical work. Context and the awareness of musical development plays an important role in filtering out such patterns by the experts.

V. Conclusions

In this article an initial investigation of rhetorical patterns in the music of Tomas Luis de Victoria was presented. A conceptual framework for rhetorical patterns was established, which in turn lead to the computational model of aforesaid rhetorical patterns. One of the main goals of this study is to test out the computational model. In order to do so, we intentionally chose a relatively small-size corpus of highly imitative music such as Victoria's masses. Annotators were asked to categorize patterns found by pattern discovery and this was compared to the categorization assigned by the computational model. It was shown that the closed-pattern mining approach with constraints produces good results that can be easily explained in musical terms. Furthermore, this approach can be a useful aid to the musicologist in the discovery of relevant rhetorical patterns. Actually, our future research will consist of designing a computer-aided system to study rhetorical patterns, which will include looking for them by using the computational models described in this paper. As seen in the results section, the computational model was robust enough to assign the same class as that assigned by the annotators for most of the patterns and discovered some other additional patterns, which were obtained mainly due to the sub-pattern problem described above. Moreover, a novel approach to reaching consensus among experts, the Delphi method, was used. This method improved inter-annotator agreement.

However, several problems of contextual knowledge were detected in this study and future works should concentrate on how to incorporate that knowledge in the pattern discovery process. In our case, rhetorical patterns contained in larger rhetorical patterns posed delicate problems as how to interpret repetition within repetition and how to resolve ambiguity so that they are clearly classified as a unique rhetorical pattern. One potential approach to deal with this issues could be considering maximal closed patterns. Another possible approach to explore could be the polyphonic modeling of harmony and counterpoint in the pieces.

Since this was an initial study, the size of the corpus was kept intentionally small. As near future work, we intend to test out the computational model on larger corpora and gain more insight into rhetorical patterns in general and the sub-pattern problem. Considering other kinds of figures of speech, not only the imitative ones, is also an interesting avenue for further research.

Acknowledgment

The authors are grateful to Rafael Martin for his participation in the musical assessment and his comments on the paper. We would like to thank the reviewers for their useful and valuable comments. We are specially grateful to Darrell Conklin for his constructive feedback.

References

[1] MIREX Wiki, "Mirex wikipedia: Discovery of repeated themes & sections." https://www.music-ir.org/mirex/wiki/2013:Discovery_of_Repeated_Themes_/6_Sections, 2013. Accessed in February 2021.

[2] A. Álvarez, F. Gómez, "Motivic pattern classification of music audio signals combining residual and lstm networks," International Journal of Interactive Multimedia & Artificial Intelligence, vol. 6, no. 6, pp. 137–163, 2021.

[3] B. Meudic, E. Saint-James, "Musical pattern extraction: from repetition to musical structure," in Proceedings of Computer Music Modeling and Retrieval, May, 26-27 2003, pp. 1–9.

[4] O. Melkonian, I. Y. Ren, W. Swierstra, A. Volk, "What Constitutes a Musical Pattern?," in Proceedings of the 7th ACM SIGPLAN International Workshop on Functional Art, Music, Modeling, and Design, 2019, pp. 95–105, Association for Computing Machinery.

[5] D. Deutsch, "Music perception," Frontiers In Bioscience, vol. 12, pp. 4473–4482, 2007.

[6] O. Lartillot, "Discovering musical pattern through perceptual heuristics," in Proceedings of the Fourth International Conference on Music Information Retrieval, Baltimore (Maryland), USA, October 26-30 2003, pp. 89–96. http://ismir2003.ismir.net/papers/ Lartillot.PDF.

[7] F. Lerdahl, R. Jackendoff, A Generative Theory of Tonal Music. Cambridge, Massachussetts: MIT Press, 1983.

[8] L. Meyer, Emotion and Meaning in Music. Chicago: University of Chicago Press, 1956.

[9] D. Huron, Sweet Anticipation. Massachusetts: MIT Press Books, 2006.

[10] D. Temperley, Music and Probability. Cambridge, Massachusetts: MIT Press Ltd, 2010.
[11] E. Narmour, The Analysis and Cognition of Basic Melodic Structures: The Implication-Realization Model. Chicago: University of Chicago Press, 1990.
[12] E. H. Margulis, "A model of melodic expectation," Music Perception: An Interdisciplinary Journal, vol. 22, no. 4, pp. 663–714, 2005, doi: 10.1525/mp.2005.22.4.663.
[13] A. J. Oxenham, "The perception of musical tones," in The Psychology of Music (Cognition and Perception), Deutsch Ed., Cognition and Perception, San Diego: Academy Press, 2012, pp. 1–34.
[14] D. Deutsch, "The processing of pitch combinations," in The Psychology of Music (Cognition and Perception), D. Deutsch Ed., San Diego: Academy Press, 2012, pp. 249–326.
[15] J.-J. Nattiez, Music and Discourse: Towards a Semiology of Music. Princeton, NJ: Princeton University, 1990.
[16] O. Lartillot, "A musical pattern discovery system founded on a modeling of listening strategies," Computer Music Journal, vol. 28, no. 3, pp. 53–67, 2004.
[17] E. H. Margulis, On repeat: how the music plays the mind. New York: Oxford University Press, 2014.
[18] P.-Y. Rolland, "Discovering patterns in musical sequences," Journal of New Music Research, vol. 28, no. 4, pp. 334–350, 1999.
[19] O. Lartillot, "Efficient extraction of closed motivic patterns in multi-dimensional symbolic representations of music," in Proceedings of the 6th International Conference on Music Information Retrieval, London, UK, September 11-15 2005, pp. 191–198. http://ismir2005.ismir.net/proceedings/1082.pdf.
[20] O. Lartillot, "In-depth motivic analysis based on multiparametric closed pattern and cyclic sequence mining," in Proceedings of the 15th Conference of the International Society for Music Information Retrieval, 2014.
[21] D. Conklin, "Mining contour sequences for significant closed patterns," Journal of Mathematics and Music, vol. 15, no. 2, pp. 112–124, 2021.
[22] M. F. Quintilian, Institutio Oratoria. Cambridge, Massachusetts: Harvard University Press, 1920. Translated by H. E. Butler. Original publication date: 95 CE.
[23] J. Tarling, Weapons of Rhetoric. Hertfordshire, United Kingdom: Corda Music, 2004.
[24] D. M. Randel, Harvard Dictionary of Music. London, England: Belknap Press, 2003.
[25] D. Bartel, Musica Poetica: Musical-Rhetorical Figures in German Baroque Music. Nebraska: University of Nebraska Press, 1997.
[26] R. López Cano, Música y retórica en el Barroco. Barcelona: Editorial Amalgama textos, 2012.
[27] E. Parada-Cabaleiro, M. Schmitt, A. Batliner, B. Schuller, "Musical-linguistic annotations of il lauro secco," in Proceedings of the 19th International Society for Music Information Retrieval Conference, (ISMIR), Paris, France, September 23-27 2018.
[28] G. Wagstaff, "El impacto del Concilio de Trento," in Historia de la Música en España e Hispanoamérica, M. del Carmen Gómez Ed., Fondo de Cultura Económica, 2012, pp. 397-462.
[29] A. W. Atlas, Renaissance Music: Music in Western Europe, 1400-1600. New York: Norton and Company, 1998.
[30] R. Stevenson, La música en las catedrales españolas del Siglo de Oro. Madrid: Alianza Música, 1992.
[31] N. Álvarez, "Tomás Luis de Victoria." https://www. uma.es/victoria/partituras.html, 2009. [Online]. Available: https://www.uma.es/victoria, Accessed in April, 2020.
[32] A. F. Jorm, "Using the Delphi expert consensus method in mental health research," Australian and New Zealand Journal of Psychiatry, vol. 49, no. 10, pp. 887–897, 2015.
[33] A. Fornells, Z. Rodrigo, X. Rovira, M. Sánchez, R. Santomá, F. Teixidó-Navarro, E. Golobardes, "Promoting consensus in the concept mapping methodology: An application in the hospitality sector," Pattern Recognition Letters, vol. 67, pp. 39–48, 2015.
[34] J. Surowiecki, The wisdom of crowds: Why the many are smarter than the few. Londres: Abacus, 2004.
[35] P. Romero Luis, El buen músico: una definición por consenso en los acervos clásico y flamenco. PhD dissertation, Facultad de Ciencias Jurídicas y Sociales, Universidad Rey Juan Carlos, Spain, 7 2017.
[36] R. A. Green, "The delphi technique in educational research," SAGE Open, vol. 4, no. 2, 2014, doi: 10.1177/2158244014529773.
[37] M. S. Cuthbert, C. Ariza, "Music21: A toolkit for computer-aided musicology and symbolic music data," in Proceedings of the 11th Conference of the International Society for Music Information Retrieval, 2010, pp. 637–642, ISMIR.
[38] J. Wang, J. Han, "Bide: Efficient mining of frequent closed sequences," in Proceedings of the 20th International Conference on Data Engineering, Boston, MA, USA, 2004, pp. 79–90.

Francisco Gómez-Martín

Francisco Gómez-Martín became Full Professor at Technical University of Madrid in 1994.He started doing research on computational geometry, computer graphics and facility location. In 2003 he switched to Music Information Retrieval and Computational Music Theory and has been doing research in this field since then. Francisco Gómez received a Ph.D. in Computer Science from the Technical University of Madrid under the supervision of Godfried Toussaint. His main interests in Music Information Retrieval and Computational Music Theory are music similarity, mathematical measures of rhythm complexity and syncopation, automated analysis of music traditions, especially flamenco music, Afro-Cuban music, Brazilian music and in general African music, teaching mathematics via the arts, and active learning methods in teaching mathematics. He has participated in several research projects funded by several Spanish agencies.

Manuel Tizón

Doctor cum laude, two master's degrees (music creation and pedagogy, 2010 and 2017) and degree in Musicology (2009), Composition (special award, 2010), Guitar (2008) and Pedagogy (2009) from the Madrid Royal Conservatory. His musical research interests are related to emotions and music, rhetoric and educational innovation. His research has been published in journals such as RECIEM (Revista Electrónica Complutense en Investigación en Educación Musical) or Nassarre (Revista Aragonesa de Musicología), and in proceedings of international conferences such as FMA (Folk Music Analysis) or ICME (International Conference on Music and Emotion). As a composer, he has premiered works for various ensembles and has been awarded in several competitions. He is coordinator of the Music Area and academic director of the Master in Musical Research at the International University of La Rioja.

Aitor Arronte Álvarez

Aitor Arronte Álvarez is a machine learning researcher specializing in Music Information Retrieval, Audio Signal Processing, and Speech Recognition. He works at the University of Hawaii at Manoa at the Center for Language and Technology as a Technology Specialist. Aitor Arronte Alvarez holds a M. Eng. in Decision Systems Engineering and finished his Ph. D. at the Universidad Politécnica de Madrid.

Víctor Padilla

Dr. Víctor Padilla Martín-Caro was born in Madrid, Spain. Currently, he is the Vice-Dean of Social Sciences and Humanities at UNIR, Universidad Internacional de La Rioja. Previously, he worked as a musical research at Lancaster Institute for the Contemporary Art at Lancaster University, UK. His research interests range from musical composition, optical music recognition (OMR), to music information research (MIR) for pattern discovery.

Mapping the Situation of Educational Technologies in the Spanish University System Using Social Network Analysis and Visualization

B. Vargas-Quesada[1]*, C. Zarco[2], O. Cordón[3]

[1] Information and Communication Department, University of Granada, Granada (Spain)
[2] Department of Market Research, Universidad Internacional de La Rioja, Logroño (Spain)
[3] Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Granada (Spain)

Received 12 February 2021 | Accepted 23 July 2021 | Published 29 September 2021

Abstract

Educational Technologies (EdTech) are based on the use of Information and Communication Technologies (ICT) to improve the quality of teaching and learning. EdTech is experiencing great development at different educational levels worldwide, especially since the appearance of Covid-19. The recent publication of a study by the ICT Sectorial of *CRUE Universidades Españolas*, the Spanish University Association, is the first report on the implementation of such technologies within Spain´s University System. This paper presents two different maps based on the data from that report. Together, they illustrate the penetration of different types of EdTech in our university system and shed light on the strategic interest behind their adoption. Our goal is to produce self-explanatory maps that can be easily and directly interpreted. The first map reflects wide granularity in terms of the global importance of technologies, while the second points to relevant conclusions given the spatial position of Spain´s universities, and the size of the nodes that represent them (directly related with their strategic interests on EdTech), as well as with the local relationships existing among them (identifying similarities on those strategic interests).

Keywords

Educational Technologies, Information and Communication Technologies, Maps, Social Network Analysis, Spanish Higher Education System, Universities.

DOI: 10.9781/ijimai.2021.09.004

I. Introduction

SINCE early 2020, the Covid-19 pandemic in Spain and its evolution have led to the adoption of extraordinary containment measures imposed by the national government, the 17 regional governments, and each university Rector´s office, with regard to the suspension of face-to-face academic activity and/or its maintenance in distance mode, for all centers of primary, secondary and higher education.

The role of Information and Communication Technologies (ICT) in every social sector has clearly grown in impact on the educational system and current approaches to education, whether formal (primary education, high schools, universities) or other means of training (business and lifelong education at home). Indeed, virtual learning has become a necessity over the past year. The Covid-19 has given rise to new educational methodologies —more focused on students— to be applied in different contexts, and the generation of new content formats. When directly related to teaching innovation, such concepts translate as the use of multimedia technologies and the Internet to improve the quality of teaching and learning, facilitating access to resources and services, while promoting collaboration and knowledge exchange. Nowadays, this is known as Educational Technologies (EdTech) [1], [2].

Higher education cannot turn its back on such transformations. Universities are busily establishing methodologies, protocols, and training programs for their staff, and progressively fomenting technical and methodological support units, so as to implement EdTech in everyday teaching activities. Given the importance of this issue, a number of reports have been put out by European university systems, including: EDUCAUSE Horizon Reports 2019 Higher Education Edition [3], Open University Innovation Report 7 [4], and Top 10 Strategic Technologies Impacting Higher Education in 2019 [5], all of them at the worldwide level, re-imagining the role of technology in higher education. Supplements refer to the National Education Technology Plan [6] for the US university system and the 2018 Survey of Technology Enhanced Learning for Higher Education in the UK [7], for the UK university system.

Unfortunately, no study of this type was available for the Spanish University System (SUS) until now. Thus, in 2017, the ICT Sectorial of the Commission of Rectors of Spanish Universities (the Spanish University Association, CRUE Universidades Españolas) tackled the matter of evaluating the application of EdTech in Spanish universities and developing a situation report, the latter task being entrusted to the Working Group in Online Training and Educational Technologies (FOLTE). In order to accurately reflect the actual state of this endeavor,

* Corresponding author.
E-mail address: benjamin@ugr.es

the members of the FOLTE group designed a large survey around 28 thematic blocks (topics) selected from a previous study based on national and international references, aiming to cover the entire spectrum of EdTech and on-line training methods and tools to date. The first situational report considered by the SUS as a result of this survey provided the response of a large sample of 47 universities [8], a high percentage (62.5%) of the 75 universities in Spain. What is more, these 47 universities represent roughly 63% of the 1.5 million students in the SUS in 2017. Such figures underline the relevance of the data dealt with in this paper and point to the global interest in the outcomes described.

Briefly, our current research efforts are aimed to design a series of maps reflecting the situation of EdTech in the SUS based on data available from the survey conducted by the FOLTE group. To do so, we pre-processed information in order to represent it as a bipartite network that links universities to their adopted technologies; we also consider advanced data visualization tools based on social network analysis (SNA) techniques [9]. Specifically, we will design two different maps that will represent respectively the penetration of the different types of technologies in our university system and the strategic interest of the universities for their adoption. In addition, the strategic interest map will be analyzed in order to identify interesting relation patterns between the different universities by both considering a manual and an automatic community detection on the associated social network.

With the firm belief that inter-university collaboration is the solution for the universal educational community (the essential meaning of University) in the current situation of contingency, our final objective is to provide a case study through a series of maps based on social networks that facilitate detailed analysis of today´s relationships among technologies/universities. The maps must be user-friendly, easily understandable for information analysts looking at the distribution and spatial location of the corresponding nodes. They may therefore stand as a technological support tool for Spanish university managers who must make strategic decisions about the adoption of EdTech in their universities. The methodology can moreover be extended to the university systems of other countries. For instance, two situation reports developed by Argentina's MetaRed University Association (http://metared.com.ar) [10], Mexico's ANUIES (http://www.anuies.mx/) and MetaRed University Associations (https://www.metared.org/mx/index.html) [11] have been recently developed following the methodology used by the FOLTE group. Situation maps derived from the data collected under the Argentinian and Mexican systems could therefore be easily obtained to analytically compare the situation of the two domains and arrive at projections for the near future.

The main contributions of the current manuscript are as follows:

1. The situation of the adoption of EdTech in the university system of an entire country, the SUS, is summarized in a single visual representation. On the one hand, it becomes a very useful analysis tool for professors, researchers, and specially university managers. On the other hand, this constitutes an interesting case study of the application of advanced data visualization tools based on SNA.
2. The proposed methodology has important advantages. It generates highly interpretable maps where the information analyst can identify inherent global relations at first glance, thus becoming a powerful knowledge discovery tool. Thus, the results of our study allow us to uncover interesting insight about the situation of EdTech in the SUS, both from the point of view of the adoption of the different EdTech and especially from the viewpoint of the strategic interests of the Spanish universities. Some of these conclusions were already drawn in the FOLTE report from which the survey data used to generate the maps was obtained, but others improve upon these or are even completely novel and highly informative.
3. As the obtained maps are based on social networks, different SNA methods can be applied on them to develop additional knowledge discovery tasks. Specifically, in the current work we apply a community detection method to supplement the expert analysis developed in the strategic interest map of the Universities.

II. Background

A. Educational Technologies

EdTech is a field of knowledge born in the 1950s and meant to provide a response to the incorporation of media and materials into education [1]. The area is very widespread given the need for innovative proposals that contribute to training processes, to improving the quality of educational services, and to renewing contents so that they respond to teaching demands [12], [13]. In the last few decades, numerous developments have stemmed from the use of ICT in teaching [14], [15]. The Internet, interactive media, mobile devices, and many other technologies have expanded the potential for teaching and learning inside and outside the classroom, both for students and for teachers [2], [16]. Novel teaching methodologies include the flipped classroom [17] and SPOC courses, personalized learning based on learning analytics [18], [19], new "disaggregated" educational formats such as MOOC courses [20], [21], the use of mobile devices, physical spaces for teaching as the Makerspaces, and methods of certification as Blockchain. These are just some examples of the great variety of methods and techniques that comprise EdTech.

B. The CRUE TIC FOLTE Working Group

Well aware of the importance of this discipline for improving the quality of university teaching in Spain, the CRUE Association created the FOLTE Working Group in December 2016, in the framework of its ICT Sectorial (http://tic.crue.org). Currently, the group is made up of more than 100 members from 55 Spanish Universities, both public and private, in addition to experts having a technical and/or academic background. Although the main focus of the group is on the ICT aspects of EdTech, because it has a multidisciplinary nature, it also analyzes methodological and instructional aspects of education. The main information, objectives and lines of work of FOLTE can be found at http://tic.crue.org/grupos-de-trabajo/formacion-online-y-tecnologias-educativas.

C. Situation of Educational Technologies in Spanish Universities

The University plays a key role in the development of ICT with academic objectives. The daily use of technology calls for an essential change in the way it is learned and taught. Although EdTech have evolved over several decades and has demonstrated its importance for university teaching, the different UNIVERSITIC reports (http://www.crue.org/SitePages/Universitic.aspx) [22] put out by CRUE in recent years reflect a substantial shift of focus, toward consolidating ICT infrastructures and university support services. Even so, their implementation in Spanish universities is not always fully developed, depending largely on the specific institution involved. There are diverse reasons behind this fact: an attitude of rejection on the part of teachers accustomed to former educational models, without these technologies; teachers' lack of technological skills or training; the great diversity, variability, speed of change, and volatility of the technologies and tools used; limited technical and human resources in the University to support the application of EdTech; the absence of a strategic commitment on the part of the institution´s governing body, etc.

D. Design and Composition of the Survey

Fifty multidisciplinary experts from more than 40 Spanish universities participated in the elaboration of the survey. It is structured around 28 thematic blocks (topics) selected from 159 preliminary topics identified from national and international references. The 28 selected topics were grouped into four dimensions from which the working group wanted to obtain information on the state of EdTech in our university system: Pedagogical/Methodological, Technological, Contents, and Management. This initiative was carried out following the TPACK Model [23], which helps identify different areas where innovation can be effectively integrated by using technology in educational contexts to enhance teaching quality. Table I reports the list of EdTech topics finally selected and the dimensions of each.

TABLE I. List of Topics and Dimensions in the FOLTE Survey

Methodological	Technological
M1. Active learning M2. Adaptive learning M3. Flipped classroom M4. Gamification M5. Mobile learning M6. MOOCs/SPOCs	T1. Blockchain for learning T2. Digital assessment/badges to accredit learning T3. Interoperability standards T4. Collaborative tools T5. Plagiarism checking tools T6. Learning analytics T7. Learning management system (LMS) T8. Makerspaces T9. Proctoring
Contents	**Management**
C1. Augmented and virtual reality C2. Production of audiovisual contents C3. Automatic production of enriched video C4. Learning object repository	G1. Digital competence evaluation G2. Teacher training G3. Teaching innovation regulation G4. Student participation in EdTech plans G5. Communication plans G6. Intellectual property rights (IPR) G7. Teacher return G8. Teaching innovation support unit G9. Usability and accessibility

Once the topics were identified, the questions associated with each of them were designed, striving for a balance between the completeness and effectiveness of the instrument. Finally, 108 questions were defined with different tradeoff levels among the topics, different types of answers, and the presence of key questions leading or not to surrogate answers. The full composition of the survey can be found in [8].

III. Methodology

We made use of SNA and visualization techniques to design two maps: one on the penetration of EdTech in Spanish universities, the other on the level of adoption of EdTech in those universities. SNA techniques have demonstrated their capacity to analyze different domains and generate high-quality schematic visualizations, with network-based representations in many fields of knowledge: system modeling [24], software debugging [25], multi-objective optimization [26], social media analysis [27], and scientometrics [28], among others.

There are many data visualization methodologies in the specialized literature that could be considered for this task, whose main objective is to make visible to the brain that which is not so visible to the human eye. In addition to SNA and its variant of Pathfinder networks, the most commonly used have been multidimensional scaling, clustering, principal component analysis, self-organizing maps, and geographic information system mapping [29]. Of all of them, the representations obtained with Pathfinder networks in combination with force layout (spring embedded) algorithms, seem to offer the best results [30], [31].

Comparing the use of different visualization methods is out of the scope of this contribution as our main goal was to design informative maps of the situation of EdTech adoption in the SUS that can be easily analyzed for domain experts as university managers. Hence, we have taken advantage of our expertise in information visualization to design the most appropriate visualization methodology for the current task. For example, we decided not to consider alternatives as multidimensional scaling or principal component analysis since the resulting maps would be less informative as they would lose the local relations. Additional advantages of the selected methodology are provided in Section III.C.

Map generation requires the pre-processing of data from a relational perspective to ensure that the network properly reflects all relevant information. To do so, a small application was programmed in Visual Basic for Excel[1]; it allowed us to build an asymmetric matrix gathering the answers given by each one of the 47 universities to the 108 questions of the survey, grouping them into the 28 topics. Our case entailed two types of analysis elements —the 28 EdTech topics considered, and the 47 universities that responded to the survey, indicating whether they had sufficient strategic interest to adopt new measures or not[2]. Our relational structure is thus an undirected bipartite network, since there could only be relations (links) between nodes of the two different types (topics and universities). There would be a link between a university and a specific EdTech topic if the university had actually adopted that technology, as reflected in the survey responses; but there could not be direct links between two universities or two technologies. The links are necessarily undirected, since when a university adopts a technology, that technology has of course been adopted by the university.

The first network is associated with the introduction of the technologies. It is made up of 28 nodes, one per topic, and its relations represent the similarity between each pair of technologies, depending on their adoption in the 47 universities. Reciprocally, the second network is associated with the strategic interest shown by the Spanish universities in adoption of these technologies. This network includes 47 nodes, one per university, and its relations represent the similarity between each pair of universities, again depending on the technologies adopted by each one.

After building the original network, we apply a network pruning method, an efficient variant of the Pathfinder algorithm [32], in order to reduce the relations between the levels of penetration of the technologies, on the one hand, and their adoption, on the other. The intention is to keep only the most significant ones at a global level (i.e. the most salient relations). To obtain the visual map, a social network layout algorithm based on forces [33] is used, tracing the network by locating the nodes in the plane.

A. Construction of the Bipartite and Projected Social Networks

To compute the network weights, a numerical value is associated with each answer: value 0 for the negative answers, and value 1 for a positive response. The "Under study" answers were deliberately ignored, so that the resulting maps reflect the current situation and not the future goals of a university. Then, the values of every answer associated with each topic were aggregated and stored in a numerical array of 28 dimensions, one per topic. Finally, the calculated values were normalized. In this way, the answers of each university regarding

[1] Available for the interested reader upon request from the authors.

[2] The source data considered is publicly available at https://tic.crue.org/publicaciones/#folte in *CSV* format.

the adoption of the EdTech under the 28 topics are represented by a 28-dimensional point in the $[0,1]^{28}$ space, while reciprocally, the penetration of each technology in the 47 universities is represented by a 47-dimensional point in $[0,1]^{47}$ space. The final result is an adjacency matrix of dimensions 28×47 that represents the bipartite network.

As a consequence, the network is weighted, i.e. the links indicate the "strength" of the adoption relationship. The weights also reflect the data of the survey questions, as explained earlier in this section, by means of the layout algorithm used, making the most related nodes attract each other and be situated closer together, while the less related ones become more separated.

We made two projections of the bipartite network to obtain the two desired maps. The first projected network is associated with the penetration of technologies and has 28 nodes, one per topic; its relations represent the similarity between each pair of technologies depending on their adoption in the 47 universities. Thus, there is only a link between two nodes (topics) if the original bipartite network included at least one university that had adopted both technologies. Otherwise, these two technologies are not at all similar to each other. In turn, the weight of the link depends on the number of universities that share the use of both technologies; the higher the number of universities, the greater the associated weight and therefore the greater the similarity between the two EdTech topics. Since there are 47 universities in the sample, the weight of the link is defined as {0, 1, ..., 47}, and then all the values are normalized by dividing them by 47. In this way, the adjacency matrix associated with the network of technologies is a symmetrical, square matrix of dimensions 28×28 with values in [0,1].

Reciprocally, the second projected network is associated with the strategic interest shown by the universities in the adoption of the analyzed technologies. This network includes 47 nodes, one per university, and their relationships represent the similarity between every pair of universities depending on the technologies adopted by each. In this case, the weight of the link depends on the number of EdTech topics adopted that are shared, and its maximum value before normalization is 28. The adjacency matrix associated with the university network is a symmetrical, square matrix of dimensions 47×47 with values in [0,1]. Fig. 1 reports an illustrative example of the procedure followed.

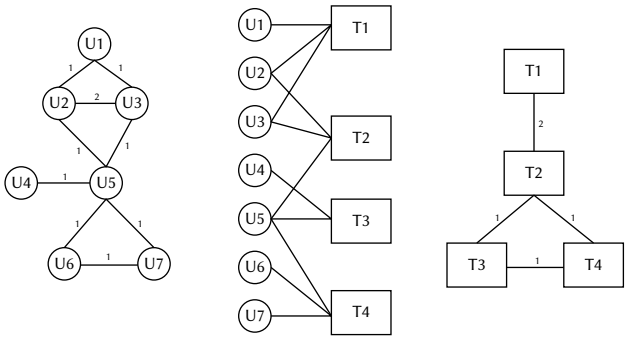


Fig. 1. Illustrative example on the creation of networks of EdTech topics and universities through the projection of the bipartite network.

B. Prunin the Social Networks

Generating graphic representations that facilitate analysis and comprehension can be difficult due to the vast dimensions of social networks. This can lead to an overload of information for the specialist, impeding interpretation of the obtained maps [34]. In our domain, the two social networks of EdTech and universities have few nodes but very high density. To achieve an attractive layout, it is necessary to reduce the number of links in the network (by pruning), thus creating a map that shows only the underlying fundamental structure, leaving the n original nodes and maintaining the most essential links.

The SNA literature offers three choices for pruning weighted networks [35]: a) prune the links having weights lower than a certain limit; b) obtain the minimum spanning tree network; or c) use Pathfinder pruning [36] to set limitations on the network paths and remove those links that do not fulfill them.

We chose the third alternative because of the advantages it provides for data visualization. The Pathfinder algorithm is known for having mathematical properties related to the conservation of triangular inequality in network paths of length q (algorithm parameter) (see [36]) for a summary of the properties). The parametric distance (the Minkowski metric) with a parameter r is used to calculate the path distances. Links not verifying the triangular inequality according to that distance metric will be removed, as they are redundant, not being included in any shortest path. When we apply the Pathfinder method with parameters q and r to a weighted network, a new pruned network called PFNET (r, q) is obtained. Therefore, the algorithm makes it possible to perform a social network sequence with decreasing complexity when the value of $q \in \{2, ..., n\text{-}1\}$ increases. The PFNET (r, q = 1) stands for the initial network, and the PFNET (r, $q = n$-1) includes the least possible number of links.

In our case, a quick version of the Pathfinder algorithm was used —Fast-Pathfinder [32], freely available from GitHub (https://github.com/aquirin/pathfinder). We applied values $r = \infty$ and $q = n$-1 to get maximum pruning, so as to derive more comprehensible maps.

C. Social Networks Layout

To draw our maps, we applied the Kamada & Kawai algorithm [30]. This method has proven to be very effective when combined with Pathfinder networks in other situations. The Kamada-Kawai algorithm allowed us to locate the nodes of the network so that their positions in the two-dimensional space could be determined by the existing global relations, maintaining the theoretical lengths of the original network paths, and also causing few cross-links; this provided for the most aesthetic and pleasant vision possible.

The visualization resulting from the combination of Pathfinder network pruning algorithm plus Kamada-Kawai layout algorithm offers a great advantage —it resembles a subway map, which facilitates its interpretation by the information analyst in metaphoric fashion. The center and the external limits of the represented information domain can be identified simply, differentiating between the center of the map (where the nodes associated with the most important analysis elements (network nodes) are located, according to the existing relationships) and periphery (where the least relevant nodes are located) [37]. In addition, one can determine the global relationships between the different analytical elements by "moving" from one node to another along the paths formed by the links. This facilitates the identification of global similarity between the analysis elements (as a function of the number of links that make up the paths) and the spatial separation between the nodes (associated with the actual length of the links). The links with higher weights, that is, greater similarity between the analysis elements in the nodes, are shorter than those with lower weights and less similarity; pruning maintains the lowest possible number of links in the original network that preserve the geodesic distances of all the global relations, clarifying the visualization. Finally, the most important nodes can be easily identified in terms of the number of links preserved in the pruned network. Consequently, these nodes act as intermediaries with other "subway lines" (network paths), i.e. as hubs or crossing points.

To enrich the visualization of our maps, the size of the nodes is

proportional to the weighted degree of the node (a measure that relates its number of links and the weights of these) in the original network. Hence the elements of analysis having stronger relations are seen at first glance, that is, the technologies used by more universities —and the universities that show strategic interest in the implementation of more technologies— present larger nodes in the map. The final visualization was obtained using the VOSviewer tool [38].

D. Community Detection

Complex networks and, consequently, social networks tend to show a community structure. Communities are simply defined as groups of similar nodes. This property usually occurs as a consequence of the global and local heterogeneity of the distribution of links in the network (the underlying graph). Based on the concept of network density, communities can be defined as groups of nodes that are densely connected internally that have scattered connections to each other. In our application domain, community detection can help us to identify interesting relation patterns between the different universities in the strategic interest social network.

There is a large number of automated community detection methods in the specialized literature [39]. In fact, this is a research area that has undergone great development in the last two decades. Louvain method [40] has been long considered as a *de facto* standard thanks to its good properties regarding accuracy and run time. Recently, a new method called Leiden has been proposed to fix a defect identified in the Louvain algorithm resulting in the derivation of disconnected communities. Leiden method extends Louvain by allowing it to split groups and not only join them through a fast local move approach. The new algorithm both increases the quality and speed of Louvain's [41]. Due to these reasons, the Leiden algorithm will be considered in the current contribution. Authors provide a free Java implementation (https://github.com/CWTSLeiden/networkanalysis) and a Gephi plugin (https://github.com/vtraag/gephi-leiden-plugin/), both of them available in GitHub.

IV. Analysis and Results

This section is devoted to analysis of the two situation maps of EdTech in the SUS, as well as the extraction of knowledge from each one of them. This information makes it possible to study, in detail, the relations existing between the different thematic blocks of technologies from the perspective of their adoption by the different universities (penetration of EdTech in the Spanish universities) and of the relationships between the different universities with respect to the application of different technologies (strategic interest of universities in the adoption of EdTech).

A. Analysis of the Educational Technologies Map

Fig. 2 shows the visualization resulting from application of the developed methodology to the projected social network of the 28 EdTech topics considered in the FOLTE report. While the original network contained 359 links and had a density D=0.95, the pruned network PFNET(r,q=n-1) in the figure has 132 links and D=0.35.

The obtained map is schematic and clear, allowing us to identify

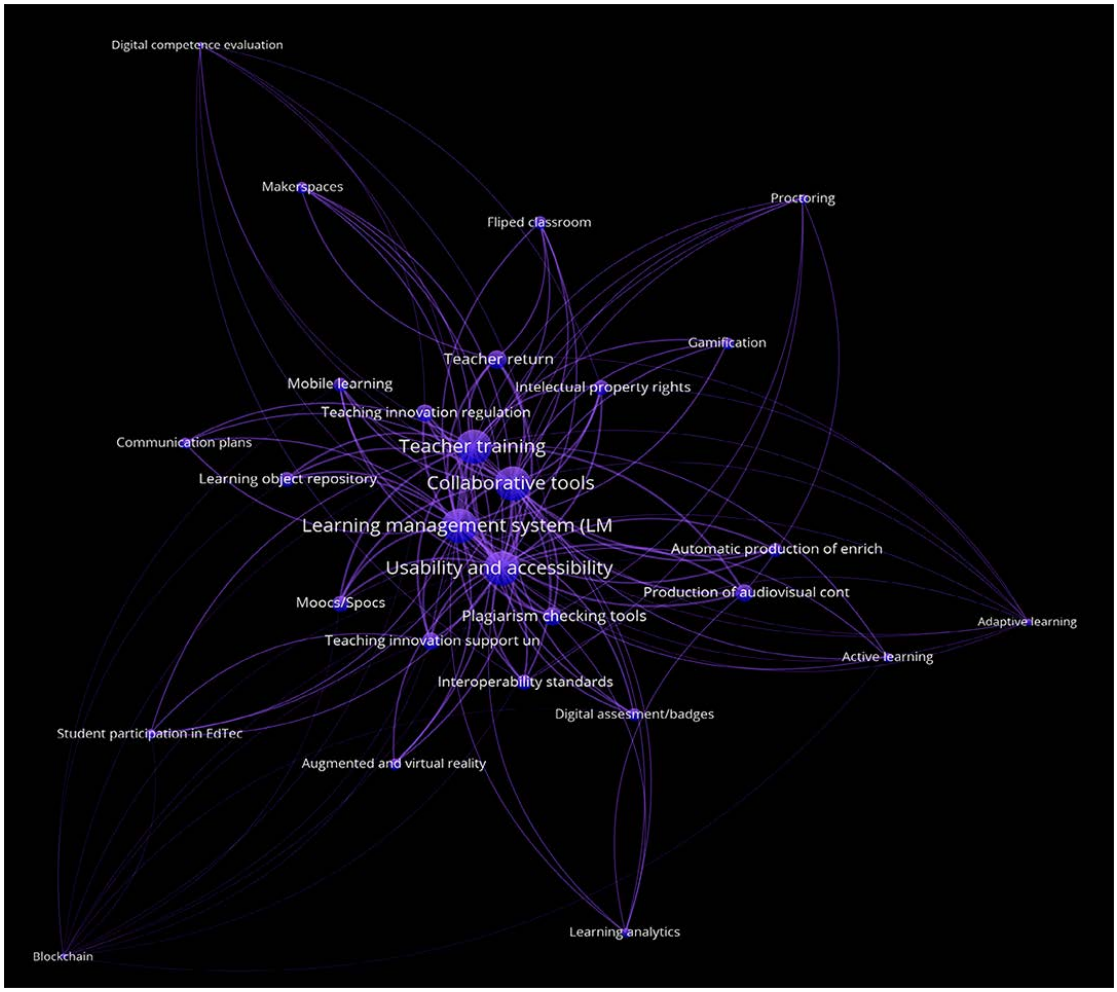


Fig. 2. EdTech Map.

important aspects of the analyzed data such as local relations. The nodes associated with two technologies have a much closer spatial position when their adoption is more similar (i.e. when many universities have jointly adopted both technologies) at a global level. Thus, the distances between the nodes, determined by the layout algorithm to make them match the global network distances, intuitively represent the similarities and differences between the different EdTech topics depending on their adoption. There is a center-periphery effect according to which the technologies located in the central part have a more similar behavior to each other (they are mostly adopted). The technologies positioned at the ends of the map show very different behavior from those in the center (lower adoption values that are reduced as we move away from the center) and consequently, also different from each other in terms of joint adoption. For example, *Adaptive learning* and *Active learning* are currently minority technologies, as reflected by their presence at one end of the map and the small size of their nodes; but they behave similarly to each other, given their joint adoption with other technologies, since they are located very near each other on the map. In contrast, *Blockchain* also shows limited adoption but behaves differently with respect to the two former technologies (and the rest), being isolated and located at a different end of the map.

In view of the map, we can corroborate several of the conclusions reached in CRUE's FOLTE report (see [8]). The 11 topics of the technologies identified there as "well established" (*Teacher training*, *Collaborative tools*, *Learning management system (LMS)*, *Usability and accessibility*, *Teaching innovation regulation*, *MOOCs/SPOCs*, *Teaching innovation support unit*, *Plagiarism checking tools*, *Learning object repository*, *Teacher return*, and *Production of audiovisual contents*) are located in the center of the map and show a large number of relations of joint adoption among them (large size of the nodes), indicating their importance in the network and their extent of adoption by Spain´s universities. The spatial distribution of this central part indicates some segmentation in their importance, distinguishing two groups: one more focused with the six most important technologies (the first six mentioned), and another with the remaining five, seen to occupy a slightly lower level of importance by being a little further away from the center of the map. This also gives us a global idea of the relationship between technologies, since a close position on the map of a group of them indicates a greater joint adoption. The other three technologies located beyond the central zone of the map are *Automatic production of enriched video*, *Intellectual property rights (IPR)*, and *Mobile learning*, which were classified in the second category in the report. Although they did not reach the adoption threshold of 70% required for the category of "well established", their position is appropriate given that they show fairly high penetration (almost 60% the first, 57.6% the second, and 53.2% the third).

The analysis of the periphery of the map is also quite similar to the hierarchy detected in the report. Two of the three technologies identified in the "low interest" category, *Digital competence evaluation* and *Blockchain*, are clearly located at two of the ends of the map, their nodes furthest from the center and smallest in size. The other technologies with a very peripheral position are *Learning analytics*, *Adaptive learning*, *Proctoring*, *Makerspaces*, and *Student Participation in EdTech plans*. All of them are located in the third category, "incipient", in the report. The map permits a greater granularity in the differentiation of the global importance of the latter five technologies, depending on the distance to the center and the joint relations with the rest, reflected by the size of the node. *Makerspaces* and *Student participation in EdTech plans* are the two most important topics in the group according to these criteria.

Finally, the remaining technologies (*Communication plans*, *Interoperability standards*, *Active learning*, *Flipped classroom*, *Augmented and virtual reality*, *Digital assessment/badges to accredit learning*, and *Gamification*) present an intermediate positioning, with different implementation levels depending on their distance to the center and the number of joint adoption relationships. This situation is well reflected in the map: the former three are more centered and have more relations of joint adoption than the latter three in all cases except *Active learning*. Although it has a good node size, it is more peripheral than the other two technologies of the second group. . This is because *Active learning* belongs to the second category, "in process", formed by the technologies where the union of the current adoption and the interest in a future adoption ("under study" response) is over a 60%. Hence, its position in the map is essentially a consequence of the "under study" answers (41%) rather than the "already adopted" ones (27.3%). Since the designed map is based on affirmative answers, the position of the node is appropriate.

B. Analysis of the Universities Map

Fig. 3 reports the map of the network generated from the data of institutional interest that the 47 universities expressed in adopting the technologies of study. In this case, the pruned network PFNET(r,q=n-1) has 136 links and D=0.13, as opposed to the 1049 links of the original network, with D=0.97.

Again, the map helps us draw a series of relevant conclusions. A direct link between two nodes indicates that those two universities share an interest in the adoption of a large number of common technologies. The presence of groups of nodes in the same area of the map and, above all, the presence of paths connecting them (the metaphor of the subway map) denotes that the universities show a strategic interest in the same technologies.

This map is seen to be more disperse than the one analyzed in the previous section, precisely because there is greater dispersion in the behaviors reflected. Logically, we find universities that decided to adopt more EdTech than others based on their strategic plans. Yet given the natural distinction between one´s own technologies (for instance, *Proctoring* might hold, *a priori*, more strategic interest for on-line universities), the center of the map does not only contain the universities implementing more technologies (larger nodes). Rather, it may include others that, having adopted fewer, incorporated the ones that turned out to be the most widespread at the global level. This gives rise to a representative pattern (average strategic profile) of the decisions made by all the universities up to now.

Apart from providing an overall view of the situation, this strategic interest map also allows us to uncover interesting relation patterns between the decisions taken by the different universities. That can be done by identifying cohesive groups (communities) in the associated social network. This SNA task is usually developed in an automatic way using community detection methods [39]. However, in our case study we will take advantage of the fact that a highly interpretable situation map has been derived. The center-periphery effect represented in the map is useful to perform a manual analysis where the expert can lead the community detection process. We will thus consider both methodologies. First, a human expert analysis will be developed by focusing on the center of the map, where the universities adopting a larger number of technologies are located, and defining the groups according to the relations of the remaining universities to these central universities. The position of each university in the map, which is related with the global relations in the system, will be considered to define the soft boundaries between the groups, resulting in an interesting analysis. We will also develop an automatic community detection by means of the robust and recently proposed Leiden algorithm and will benchmark the obtained groupings with those identified in the expert analysis.

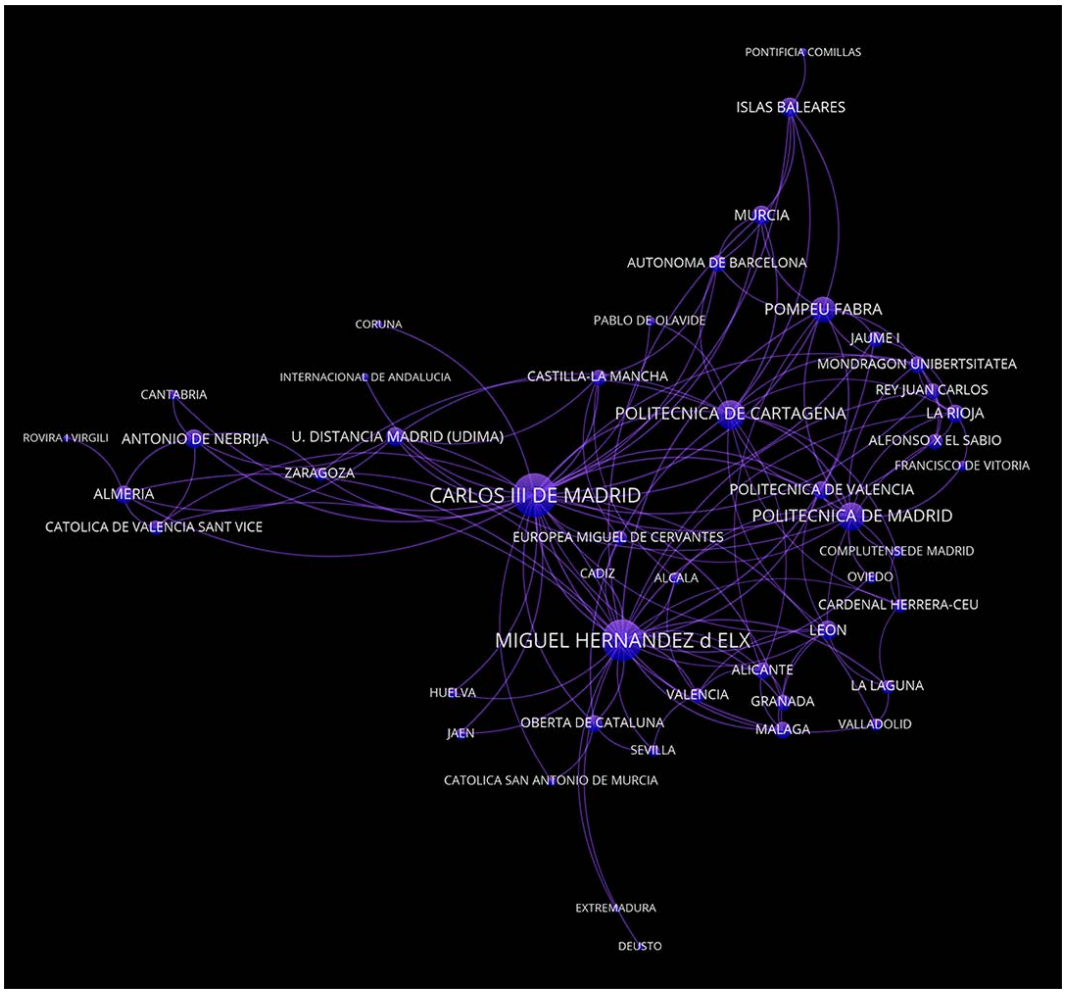


Fig. 3. Map of strategic interest of Spanish Universities in EdTech.

In the center of the map we have, on the one hand, a triangle formed by three universities that adopted many technologies: *Carlos III de Madrid*, *Miguel Hernández d'Elx*, and *Politécnica de Cartagena*. The fact that there are direct links forming this triangle means a high similarity in the chosen technologies. On the other hand, we have a series of universities that implemented fewer technologies, though some are different from the three main hubs. In this group we find: *Politécnica de Madrid*, *Politécnica de Valencia*, *Europea Miguel de Cervantes*, *Alcalá*, and *Cádiz*. The *Europea Miguel de Cervantes* university is the only one that shares links with all three hubs, indicating that despite its limited introduction of technologies (note small node size), the chosen technologies are shared with those of the three most central universities. The remaining institutions only have links to two (*Politécnica de Valencia*) or one (*Politécnica de Madrid*, *Alcalá*, and *Cádiz*) of them. This is interpreted as the average behavior (hence the central position on the map), but indicative of similarities and differences in the strategic adoption profile depending on which university they are linked to.

Moving from the center to the periphery, we may discern four large groups in the four cardinal directions, some more clearly clustered than others. To clarify these modularity relations, Fig. 4 presents the same map as Fig. 3, but highlighting in different colors the five groups of universities according to their strategic interest in adopting EdTech in Spain. The central group, depicted in blue, has already been analyzed. The other four clusters are described below. In addition to showing the internal cohesion of the groups, the inner links (i.e. those linking two universities from the same group) are seen in the group color, whereas the outer links show a blended color between the two group colors of the nodes they link.

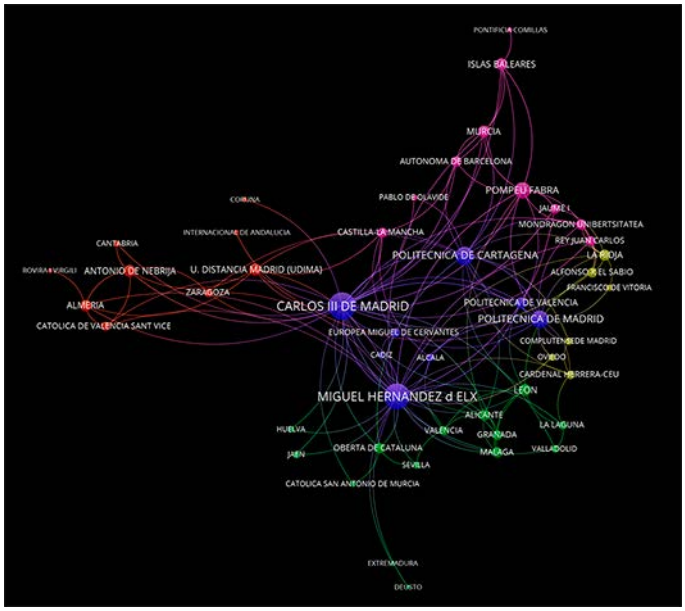


Fig. 4. Expert partitioned map of strategic interest of Spanish Universities in EdTech.

Possibly the cluster that is most easily identified is the one to the left or "west", formed by the Universities of *Coruña*, *Internacional de Andalucía*, *UDIMA*, *Zaragoza*, *Cantabria*, *Antonio de Nebrija*, *Almería*, *Católica de Valencia Sant Vicent*, and *Rovira i Virgili*. The nodes and inner links of these universities are depicted in red. They share technologies with each other as well as with *Carlos III de Madrid*, the central node that has the most connections with this group. Local relations permit the identification of patterns of similar strategic interest profiles, e.g. those of *Universidad Antonio de Nebrija* with four of the universities in the group. Indeed, we identified a new triangle between the *Antonio de Nebrija*, *Almería*, and *UDIMA* universities, the three largest nodes in the group.

Fig. 5 depicts the specific subgraph of universities in this cluster for the sake of clarity. It illustrates the strong cohesion of the group; even upon the removal of the most influential binding node of the center group, *Carlos III de Madrid* university, the group still keeps connected to all but two nodes. These two universities, *Coruña* and *Internacional de Andalucía*, become isolated, being linked only to the influential node. (We must keep in mind that the Pathfinder network preserves only the most salient relations in the original network after pruning).

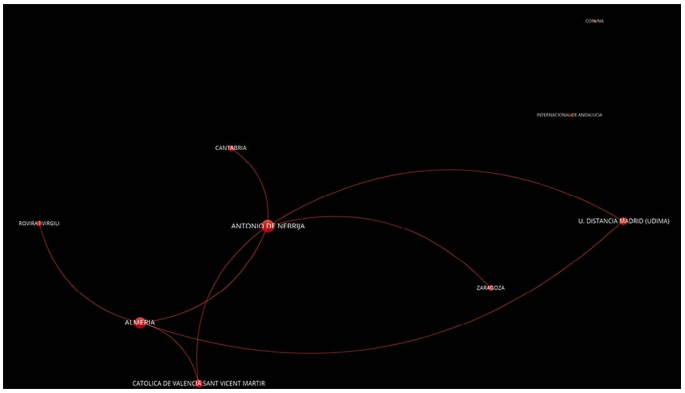


Fig. 5. "West" group of universities.

The next group identified is the "southern" sector of the map, shown in green. It includes the Universities of *Huelva*, *Jaén*, *Oberta de Catalunya*, *Sevilla*, *Católica San Antonio de Murcia*, *Extremadura*, *Deusto*, *Valencia*, *Alicante*, *Granada*, *Málaga*, *León*, *La Laguna*, and *Valladolid*. In this case, in addition to sharing technologies with each other, they share them principally with *Universidad Miguel Hernández d'Elx*. We spot two triangles between the Universities of *Valencia*, *Alicante*, and *Granada*, on the one hand, and *Alicante*, *Granada*, and *Málaga*, on the other. They show very similar profiles in terms of both the triple relation and the spatial proximity, hence the number of technologies adopted (the same, given the size of the nodes).

From a general perspective, this "southern" group has a more diverse and less cohesive behavior than the "western" group (note the diversity of relations with the main nodes in the center of the map). Fig. 6 highlights the part of the map where this can be clearly seen. Two subgroups of pairs of universities are not linked to the major component of the group: *Huelva* plus *Jaén* on the left, and *Extremadura* and *Deusto* beneath. In both cases this is because the most salient relations of these four universities are associated only with the central nodes, strongly related with them. Whereas *Extremadura* and *Deusto* are only linked to *Miguel Hernández d'Elx*, the influential central node of this group, *Huelva* and *Jaén* are also linked to *Carlos III de Madrid*. This justifies their spatial separation in the map with respect to the remaining nodes in the group.

Third, we can identify an analogous group in the upper, "northern" part of the map, represented in pink. The profiles of strategic interest

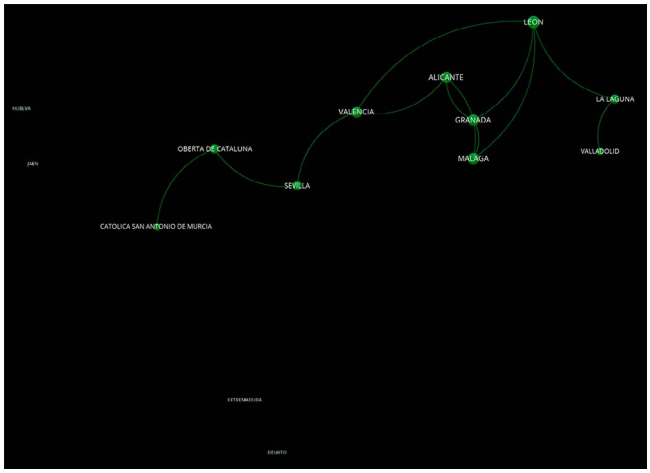


Fig. 6. "South" group of universities.

in EdTech of the universities that compose it are related to those of two universities of the main nodes, *Carlos III de Madrid* and *Politécnica of Cartagena*. Those located more to the left are related particularly to the former, those more to the right with the latter, and those of the central part with both to an equal extent. The borders of this third group are fuzzier on the right, given that the universities located in that area also have strong relations with other nodes in the center of the map such as those of the *Technical Universities of Madrid and Valencia*. This group basically includes the universities of *Castilla-La Mancha*, *Pablo de Olavide*, *Autónoma de Barcelona*, *Murcia*, *Islas Baleares*, *Pontificia de Comillas*, *Pompeu Fabra*, *Jaume I*, *Mondragón Unibertsitatea*, and *Rey Juan Carlos*, although it is difficult to draw a border or division in the right part of the group, where more cross-relations are found. There is no triangle formed only by nodes of the universities in the group. All of the above allows us to conclude that this "northern" group features the greatest diversity of strategic profiles regarding Edtech adoption.

Fig. 7 shows the north subgraph of universities. Despite the diverse behavior analyzed in the former paragraph, due to the different relational patterns involving the central nodes, this figure highlights the strong cohesion existing between the universities of the group. Only *Pablo de Olavide University* is disconnected, having just two direct links to the two influential nodes, *Carlos III de Madrid* and *Politécnica of Cartagena*.

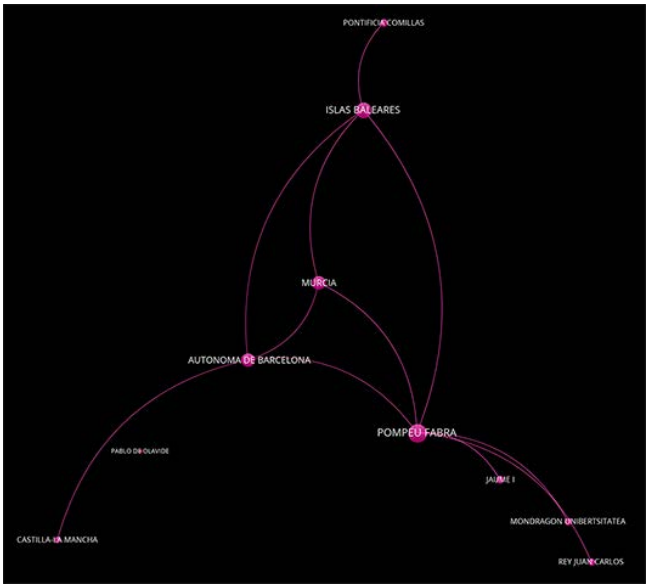


Fig. 7. "North" group of universities.

Finally, the fourth group is located to the "east", colored yellow, and its borders are the least distinct. We consider it to contain *La Rioja*, *Alfonso X el Sabio*, *Francisco de Vitoria*, *Complutense de Madrid*, *Oviedo*, and *Cardenal Herrera-CEU* universities. Their relations indicate that overall their strategic profile is similar to that of *Universidad Miguel Hernández d'Elx*, though some are also directly related to the *Universidad Politécnica de Madrid* and there are moreover links to other nodes in the center of the map. Again, no triangles are formed by this grouping.

Fig. 8 renders the final subgroup and provides additional information supporting the conclusions already drawn. When the central nodes are removed, three small subgroups can be distinguished, with the least cohesive behavior among all five of the groups analyzed. *La Rioja*, *Alfonso X el Sabio*, and *Francisco de Vitoria* are linked to four different nodes in the central part: *Carlos III de Madrid*, *Miguel Hernández d'Elx*, *Politécnica de Cartagena*, and *Politécnica de Madrid*. *Politécnica de Madrid* is the only one shared, between *Alfonso X el Sabio* and *Francisco de Vitoria*. This signals the three universities in this subgroup as gateway nodes sharing an intermediate EdTech adoption pattern. Meanwhile, *Complutense de Madrid* and *Cardenal Herrera-CEU* share a link, but their most salient relations to the central nodes are different —*Complutense de Madrid* to *Politécnica de Madrid*, and *Cardenal Herrera-CEU* to *Miguel Hernández d'Elx*. Hence, a similar case of gateway nodes. Lastly, *Oviedo* is isolated because its two main relations are specifically to these two central nodes, *Politécnica de Madrid* and *Cardenal Herrera-CEU*, thus tracing a slightly different interest pattern than the former two, and no link to either.

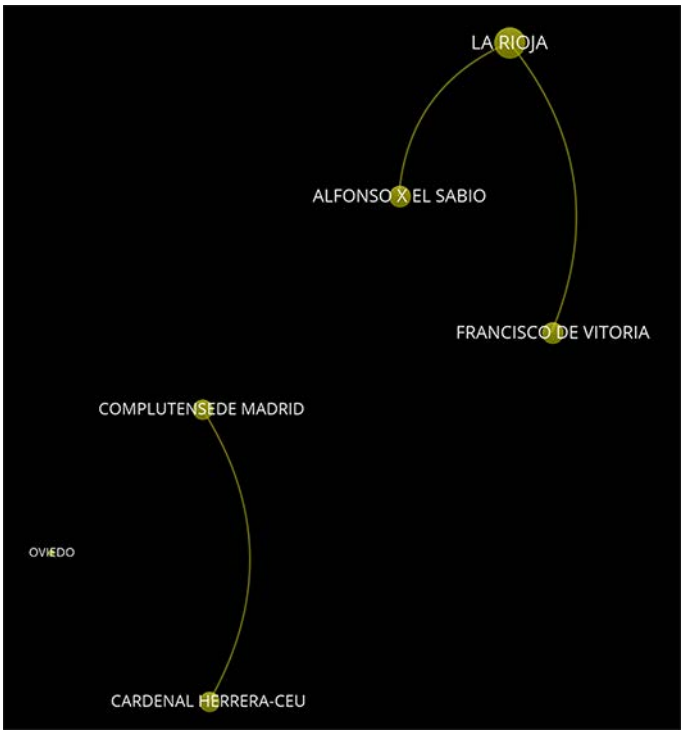


Fig. 8. "East" group of universities.

Once the expert analysis have been developed, we will validate it through automatic community detection. The Leiden algorithm has been applied setting different values for the resolution parameter in order to derive an optimal partition for different number of communities. We have considered three different number of communities, from two (which provided the optimal partition in terms of modularity) to four (which is the most similar number to that considered in the expert analysis, as we will see below). Table II collects the statistics (modularity value and number/distribution of communities) for each partition while Fig. 9 to Fig. 11 depict the different partitions in the institutional interest map.

TABLE II. Statistics of the Different Partitions Derived By the Leiden Algorithm

#Com.	Modularity	% and # nodes Com. 1	% and # nodes Com. 2	% and # nodes Com. 3	% and # nodes Com. 4
2	0,415	55,32% (26)	44,68% (21)	-----	-----
3	0,381	34,04% (16)	34,04% (16)	31,91% (15)	-----
4	0,313	31,91% (15)	29,79% (14)	25,53% (12)	12,77% (6)

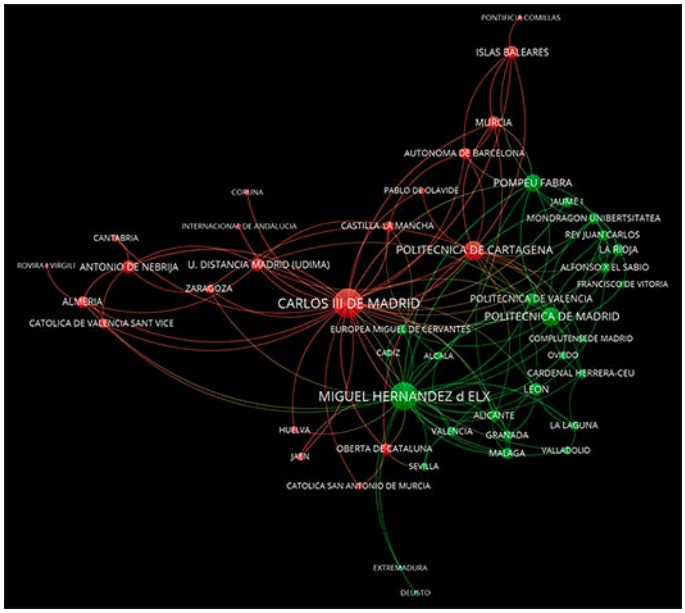


Fig. 9. Automatically partitioned map of strategic interest of Spanish Universities in EdTech: Leiden with 2 partitions.

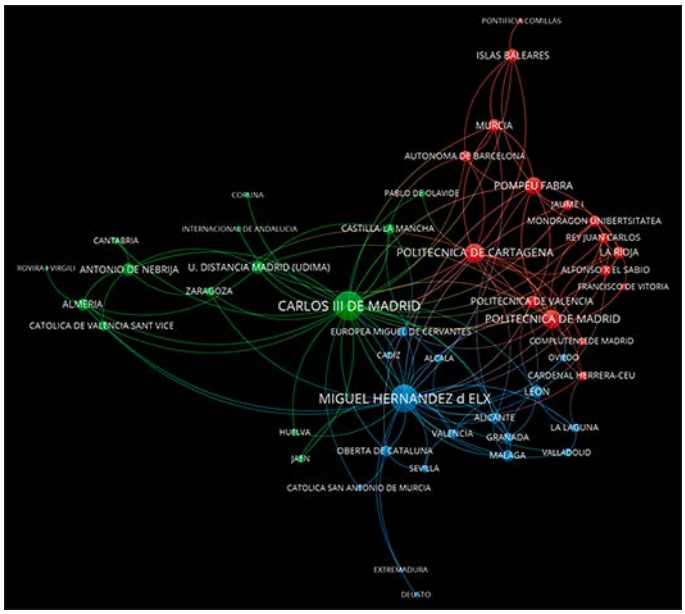


Fig. 10. Automatically partitioned map of strategic interest of Spanish Universities in EdTech: Leiden with 3 partitions.

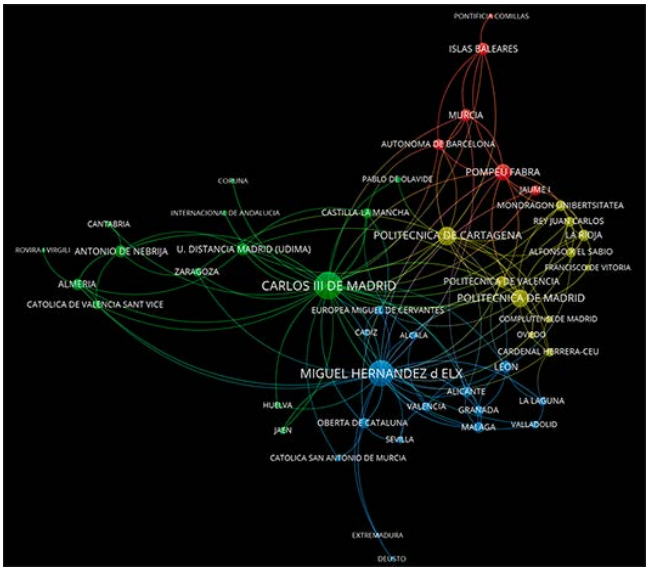


Fig. 11. Automatically partitioned map of strategic interest of Spanish Universities in EdTech: Leiden with 4 partitions.

The three partitions chosen can be considered reasonable as the modularity values are over the usual 0.3 threshold considered in the area [42]. The modularity value increases with the reduction of the number of communities. The partition with two communities shows the best value, 0,415. The main differences between the automatic and the manual partitioning is of course the fact that the nodes associated to the universities located in the center of the map were grouped in a specific community in the manual partitioning. That was done to get a single group corresponding to the universities with the highest strategic interest in EdTech adoption for the expert analysis. Then, the remaining four communities were derived from their relation to the three most central universities as defined by their positions in the map. Alternatively, automatic community detection directly creates the communities by grouping those central nodes to the most related nodes, i.e., universities having a common adoption of the different technologies, as expected.

The good performance of the map generation methodology when locating the most active universities in the center, the less active ones in the periphery, and universities sharing a similar strategic interest on EdTech spatially close can be corroborated by analyzing the composition of the groups. Note that, as soon as three different groups arise, each of them includes one of the three most central universities (Carlos III de Madrid, Miguel Hernández d'Elx, and Politécnica de Cartagena). Hence, the visualization methodology allows the information analyst to identify the inherent global relations at first sight. We can also notice that the partition with four communities is very similar to the one identified in the expert analysis (apart from the fact that in the latter a central group was voluntarily identified). Only some nodes located in the border between two different communities and with a few links to nodes in different communities are grouped in a different way (Católica San Antonio de Murcia, Huelva, Castilla La Mancha, Pablo de Olavide, ...). This again reinforces the good spatial properties of the visualization to achieve an informative representation of the analyzed domain.

Finally, note that Carlos III de Madrid and Politécnica de Cartagena universities are most similar in their EdTech adoption strategies than Miguel Hernández d'Elx university, as shown by the best partition obtained by the Leiden community detection method. This partition, composed of only two communities, groups the former two in the same community and provides us with the largest aggregation of strategic adoption plans in the SUS.

V. Conclusions

The current challenges surrounding Covid-19 are not limited to the realm of Medicine. They likewise require responses from all types of professionals, to meet a range of training needs at the core of pandemic-related priorities. The current contribution aims to visually present the situation of EdTech adoption in the SUS based on data from the survey carried out by the FOLTE Group for the CRUE TIC report [8]. We designed two different maps to respectively depict the penetration of different types of technologies in the SUS, and the strategic interest of the universities in their adoption. Our goal was to facilitate both an overall grasp as well as detailed analysis of the relations among the different technologies/universities. The maps created are easy to interpret, by looking at the distribution and spatial localization of the corresponding nodes.

The first map corroborates several of the conclusions drawn from the report, although the visual information is more granular, making it easier to distinguish the global importance of the technologies. Such additional insights depend on the distance to the center and the joint relations with the rest, reflected by the size of the nodes. The 11 topics of the most implemented EdTech topics occupy the center of the map, whose spatial distribution leads us to conclude that the six most adopted technologies are *Teacher Training*, *Collaborative Tools*, *LMS*, *Usability and Accessibility*, *Teaching innovation regulations*, and *MOOCs/SPOCs*. The two most peripheral technologies and therefore the least widespread ones are *Digital competence evaluation* and *Blockchain*.

The second map allows us to study the positioning of the different Spanish universities according to the strategic decisions made about adopting EdTech to date. Based on this map we may uncover similarities and differences, as well as identify general and specific profiles in the corresponding decision-making processes at the institutional level. The center of the map is not only composed of universities that have implemented more technologies; it likewise includes ones that, having adopted less, have incorporated precisely those most widespread at the global level. This gives rise to a representative pattern (an average strategic profile) of the relevant decisions made by all the universities until now. Moving from the center to the periphery, we can identify four large groups of universities in the four cardinal directions —some more clearly differentiated than others— depending on the similarity of the strategies applied for the adoption of technologies. That is very relevant information, pointing to similar and differential behaviors.

We should note that the maps generated are capable of showing the situation of EdTech in a full university system in a single visual representation. This makes them a powerful analytical tool for professors, researchers, and university managers involved in decision-making; indeed, they can serve as the support system requested by researchers and specialized journals [43]. In particular, the latter group can take advantage of these representations to support their strategic decisions related to the adoption of EdTech in their institutions. In addition to specific insights regarding the SUS brought out in the current manuscript, the proposed methodology can be directly applied to other university systems from different countries in the future. This would allow analysts to acquire additional knowledge about their own system and compare their situation to that of other countries. What is more, it can be applied to analyze evolution patterns by generating consecutive maps in different time periods. That is extremely interesting as the FOLTE group is currently developing a new report for CRUE that collects data on the current situation after the appearance of the Covid-19 crisis. Hence, post-pandemic maps can be generated and compared to those presented in this contribution in the future. Given that there has been a strong development of EdTech in the SUS as a consequence of the actions developed by the different

Spanish universities to face the negative effects of the pandemic, this comparison would be extremely interesting.

Acknowledgment

The authors would like to express their special thanks to the Working Group in Online Training and Educational Technologies (FOLTE), of the ICT Sectorial of the Commission of Rectors of Spanish Universities (the Spanish University Association, CRUE Universidades Españolas). Moreover, this work was supported by the Spanish Ministry of Science, Innovation and Universities, and ERDF under grant EXASOCO (PGC2018-101216-B-I00).

References

[1] T. J. Newby, D. Stepich, J. Lehman, J. D. Russell, and A.T. Leftwich, *Educational Technology for Teaching and Learning,* 4th ed., New York: Pearson, 2010.

[2] T. Gray and H. Silver-Pacuilla, Eds., *Breakthrough teaching and learning: how educational and assistive technologies are driving innovation*, New York: Springer-Verlag, 2011.

[3] B. Alexander, et al., "EDUCAUSE Horizon Report: 2019 Higher Education Edition", Louisville, CO: EDUCAUSE, 2019.

[4] R. Ferguson, et al., "Innovating pedagogy 2017: Open university innovation report #6", Milton Keynes: The Open University, 2019. Accessed: Feb. 06, 2021. [Online]. Available: https://iet.open.ac.uk/file/innovating-pedagogy-2017.pdf.

[5] J-M. Lowendahl, T-L. Thayer, and G. Morgan. "Top 10 strategic technologies impacting higher education in 2016". Accessed: Feb. 06, 2021. [Online]. Available: https://www.gartner.com/en/documents/3186323/top-10-strategic-technologies-impacting-higher-education.

[6] J. King, and J. South. "Reimagining the Role of Technology in Higher Education. A Supplement to the National Education Technology Plan", Washington: U.S. Department of Education, Office of Educational Technology, 2019. Accessed: Feb. 06, 2021. [Online]. Available: https://tech.ed.gov/files/2017/01/Higher-Ed-NETP.pdf.

[7] R. Walker, et al. "Survey of technology enhanced learning for higher education in the UK", Oxford: Universities and Colleges Information Systems Association, 2016. Accessed: Feb. 06, 2021. [Online]. Available: https://www.ucisa.ac.uk/bestpractice/surveys/tel/tel.aspx

[8] J. Gómez Ortega, et al. "Informe de situación de las Tecnologías Educativas en las universidades españolas 2018" (in Spanish). Madrid: Crue Universidades Españolas, 2019. Accessed: Feb. 06, 2021. [Online]. Available: https://tic.crue.org/publicaciones/#folte.

[9] S. Wasserman, and K. Faust, *Social network analysis: methods and applications (structural analysis in the social sciences),* Cambridge: University Press, 1994.

[10] A. Santos, et al., "Estado de situación de las Tecnologías Aplicadas a la enseñanza y el aprendizaje en la Educación Superior argentina" (in Spanish). Buenos Aires: Metared, 2019, Accessed: Feb. 10, 2021. [Online]. Available: https://www.metared.org/argentina/wp-content/uploads/sites/11/2019/10/Estado-de-Situacion-TAEA-Educacion-Superior_Metared-Argentina.pdf.

[11] J. L. Ponce-López, C. M. Vicario-Solórzano, and F. López-Valencia. "Estado Actual de las Tecnologías Educativas en las Instituciones de Educación Superior en México" (in Spanish). ANUIES, México, 2021. Accessed: Feb. 21, 2021. [Online]. Available: https://estudio-tic.anuies.mx/#estado_te.

[12] M. Miller, *Ditch That Textbook: Free Your Teaching and Revolutionize Your Classroom*, Michigan: Dave Burgess Consulting, 2015.

[13] P. A. Rodríguez, V. Tabares, N. D. Duque, D. A. Ovalle, and R. M. Vicari, "BROA: An agent-based model to recommend relevant Learning Objects from Repository Federations adapted to learner profile", *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 2, no. 1, pp. 6-11, 2013, doi:10.9781/ijimai.2013.211.

[14] G. Allen. *The New Pillars of Modern Teaching*, Bloomington: Solution Tree Press, 2016.

[15] L. Kolb. *Learning First, Technology Second: The Educator's Guide to Designing Authentic Lessons*, Portland, Oregon: International Society for Technology Education, 2017.

[16] L. De-Marcos, E. García-Lopez, A., and García-Cabot. "On the effectiveness of game-like and social approaches in learning: Comparing educational gaming, gamification & social networking", *Computers & Education*, vol. 95, pp. 99-113, 2016, doi:10.1016/j.compedu.2015.12.008.

[17] J. Nouri. "The flipped classroom: for active, effective and increased learning - especially for low achievers", *International Journal of Educational Technology in Higher Education*, vol. 13, no. 33, 2016, doi: 10.1186/s41239-016-0032-z.

[18] Á. Martínez Navarro and P. Moreno-Ger. "Comparison of clustering algorithms for learning analytics with educational datasets", *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 2, pp. 9-16, 2018, doi: 10.9781/ijimai.2018.02.003.

[19] R. Klamma, P. de Lange, A.T. Neumann, and P. Nicolaescu, "An Integrated Learning Analytics Approach for Virtual Vocational Training Centers", *International Journal of Interactive Multimedia and Artificial Intelligence,* vol. 5, no. 2, pp. 32-38, 2018, doi:10.9781/ijimai.2018.02.006.

[20] J. M. Spector, "Remarks on MOOCS and Mini-MOOCS", *Educational Technology Research and Development*, vol. 62, no. 3, pp. 385-392, 2014.

[21] S. Margoum, R. Bendaoud, K. Berrada, and A. Idrissi. "UC@MOOC's Effectiveness by Producing Open Educational Resources", *International Journal of Interactive Multimedia And Artificial Intelligence*, vol. 5, no. 2, pp. 58-62, 2018, doi:10.9781/ijimai.2018.02.007.

[22] J. Gómez Ortega, et al., "UNIVERSITIC 2017, Análisis de las TIC en las universidades españolas" (in Spanish). Madrid: Crue Universidades Españolas, 2017. Accessed: Feb. 06, 2021. [Online]. Available: http://tic.crue.org/publicaciones/informe-universitic-2017.

[23] M. Koehler, and P. Mishra, "What is technological pedagogical content knowledge (TPACK)?" *Contemporary Issues in Technology and Teacher Education*, vol. 9, no 1, pp. 60-70, 2009.

[24] D. P. Pancho, J. M. Alonso, O. Cordón. A. Quirin and L. Magdalena. "FINGRAMS: visual representations of fuzzy rule-based inference for expert analysis of comprehensibility", *IEEE Transactions on Fuzzy Systems*, vol. 21, no 6, pp. 1133-1149. 2013.

[25] E. Serrano, A. Quirin, J. Botia, and O. Cordón, "Debugging complex software systems by means of pathfinder networks", *Information Sciences*, vol. 180, no 5, pp. 561-583, 2010.

[26] Trawinski, M. Chica, D. Pancho, S. Damas, and O. Cordón, "moGrams: a network-based methodology for visualizing the set of non-dominated solutions in multiobjective optimization", *IEEE Transactions on Cybernetics.*, vol. 48, no 2, pp. 474-485, 2018.

[27] C. Zarco, C. E. Santos, and O. Cordón, "Advanced visualization of Twitter data for its analysis as a communication channel in traditional companies", *Progress in Artificial Intelligence*, vol. 8, no 3, pp. 327-333, 2019.

[28] B. Vargas-Quesada, Z. Chinchilla-Rodríguez, and N. Rodriguez, "Identification and visualization of the intellectual structure in graphene research", *Frontiers. Research. Metrics. Analytics*, vol. 2, no 7. 2017. Accessed: Feb. 06, 2021. [Online]. Available: https://doi.org/10.3389/frma.2018.00013.

[29] A. Quirin, O. Cordón, J. Santamaría, B. Vargas-Quesada, and F. Moya-Anegón, "A new variant of the pathfinder algorithm to generate large visual science maps in cubic time", *Information Processing & Management*, vol. 44, no 4, pp. 1611-1623, 2008.

[30] X. Lin, H. D. X. White, and J. Buzydlowski, J. "Real-time author co-citation mapping for online searching", *Information Processing & Management*, vol 39, no. 5, pp. 689-706. 2003, doi.org/10.1016/S0306-4573(02)00037-7

[31] N. J. van Eck, and L. Waltman, L. "Visualizing bibliometric networks", *In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), Measuring scholarly impact: Methods and practice*, pp. 285–320. Springer, 2018.

[32] Author 3, 2008.

[33] T. Kamada, and S. Kawai, "An algorithm for drawing general undirected graphs", *Information Processing Letters*, vol. 31, no 1, pp. 7-15. 1989.

[34] A. Unwin, M. Theus, and H. Hofmann, Graphics of large datasets: visualizing a million, New York: Springer Science & Business Media, 2008.

[35] C. Chen, and S., "Visualizing evolving networks: Minimum spanning trees versus pathfinder networks", in Proceedings of IEEE Symposium on Information Visualization, Seattle, USA, pp. 67–74, 2003, doi:10.1109/INFVIS.2003.1249010.

[36] R. Schvaneveldt, F. Durso, and D. Dearholt, "Network structures in proximity data", *Psychology of Learning and Motivation*, vol. 24, pp. 249-284. 1989.

[37] S. P. Borgatti, and M. G. Everett, "Models of core/periphery structures", *Social Networks*, vol. 21, no 4, pp. 375-395, 1999.kk

[38] N. J. Van Eck, and L. Waltman. "Software survey: VOSviewer, a computer program for bibliometric mapping", *Scientometrics*, vol. 84, no 2, pp. 523-538. 2010.

[39] S. Fortunato. "Community detection in graphs". *Physics Reports*, vol. 486, no 3-5, pp. 75–174, 2010, doi: 10.1016/j.physrep.2009.11.002.

[40] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre. "Fast unfolding of communities in large networks". *Journal of Statistical Mechanics: Theory and Experiment*, vol. 10:10008, 2008, doi:10.1088/1742-5468/2008/10/P10008.

[41] V. A. Traag, L. Waltman, and N. J. van Eck. "From Louvain to Leiden: guaranteeing well-connected communities". *Scientific Reports*, vol. 9: 5233, 2019, doi:10.1038/s41598-019-41695-z.

[42] A. Clauset, M. E. J. Newman, and C. Moore. "Finding community structure in very large networks". *Physical Review E*, vol. 70:066111, 2004, doi: 10.1103/PhysRevE.70.066111.

[43] C. Romero, and S. Ventura, "Guest Editorial: Special Issue on Early Prediction and Supporting of Learning Performance", *IEEE Transactions on Learning Technologies*, vol. 12, no. 2, pp. 145-147, 2019, doi:10.1109/TLT.2019.2908106

Benjamín Vargas-Quesada

Benjamín Vargas-Quesada obtained his Ph.D. degree in Scientific Library Science and Information from the University of Granada (UGR), Spain, in 2005. He is a Professor and the Dean of the Faculty of Communication and Library Science at the University of Granada, as well as a founding member of the SCImago group. His lines of research focus on information visualization, scientometric analysis of different scientific domains (geographic and disciplinary), and the analysis of networks of scientific collaboration, having published more than 100 peer-reviewed scientific publications —including a research book on information visualization, two book chapters on the same subject, and 38 JCR-SCI/SSCI-indexed journal papers 26 in Q1— and having directed five Ph.D. dissertations and participated in 15 research projects and contracts (amounting to ~2M€ overall). As of January 2021, his publications received 758 citations (h-index=14) in the Clarivate Analytics' Web of Science, and 2943 citations and h-index=27 according to Google Scholar.

Carmen Zarco

Carmen Zarco is currently an Associate Professor at the International University of La Rioja (UNIR), Spain. She joined UNIR in 2013 and coordinated the Master on Digital Marketing and e-Commerce for two years (2015-16). Before, she was the responsible of the knowledge management unit at the European Centre for Soft Computing (2007-2011) where she was the link between the different research units and the support and management staff. In her professional career, she has enjoyed different positions of responsibility in private companies while combining such activity with research and teaching. She first had a research scholarship in Baratz, S.A. Teledocumentation Services (1999), was the documentation head at the pulevasalud.com health web portal (1999-2003), and was later promoted as responsible for the Knowledge Management unit in the Puleva Food, S.L. business (2003-2006). In her research career she has published several journal articles (some of them in journals indexed in the JCR), book chapters, and contributions to national and international conferences. She has also participated in different research projects and contracts. Her main research lines are related to information retrieval, communication management, and marketing analysis using tools based on artificial intelligence and social network analysis.

Oscar Cordón

Oscar Cordón is Professor with the University of Granada (UGR), Spain. He was the founder and leader of its Virtual Learning Center (2001-05) and Vice-President for Digital University (2015-19). He also was one of the founding researchers of the European Centre for Soft Computing (2006-2011), being later contracted as Distinguished Affiliated Researcher (2011-15). He has been, for almost 30 years, an internationally recognized contributor to R&D Programs in fundamentals and real-world applications of computational intelligence. He has published more than 380 peer-reviewed scientific publications including a research book on genetic fuzzy systems (with ~1400 citations in Google Scholar) and 110 JCR-SCI-indexed journal papers (66 in Q1, 37 in D1), advised 19 Ph.D. dissertations, and coordinated 23 research projects and 15 research contracts (with an overall amount of >9M€). By February 2021, his publications had received 5357 citations (h-index=39), being included in the 1% of most-cited researchers in the world (source: Clarivate Analytics' Web of Science. 14427 citations and h-index=57 in Google Scholar). Besides, he is co-inventor of an international patent on an intelligent system for forensic identification under exploitation by Panacea Cooperative Research, distributed in Mexico and South Africa. He has developed intelligent marketing models with Zio Analytics for big brands as The Coca Cola Company, Jaguar-Land Rover L'Oréal Mexico, Samsung, Telefónica, and El Corte Inglés. Prof. Cordón received the UGR Young Researcher Career Award in 2004; the IEEE Computational Intelligence Society (CIS) Outstanding Early Career Award in 2011 (the first such award conferred); the International Fuzzy Systems Association (IFSA) Award for Outstanding Applications of Fuzzy Technology in 2011; the National Award on Computer Science ARITMEL by the Spanish Computer Science Scientific Society in 2014; the IEEE Fellow grade for his contributions to genetic and evolutionary fuzzy systems in 2018; and the IFSA Fellow in 2019. He also was a member of the Working Group that developed the Spanish R+D Strategy for Artificial Intelligence by the Spanish Ministry of Science (2018-19). He was member of the ICT Executive Board of the Association of Spanish Universities and President of the Working Group on Online Learning and Educational Technologies (2016-2020, in particular during the Covid-19 lockdown, being involved in the design of contingency plans for the Spanish Ministry of Universities). He is currently or was Associate Editor of 18 international journals, and was recognized as IEEE Transactions on Fuzzy Systems Outstanding AE in 2008 and IEEE Transactions on Evolutionary Computation AE in 2019. Since 2004, he has taken many different representative positions with the European Society for Fuzzy Logic and Technologies (EUSFLAT) and the IEEE CIS.

Local Model-Agnostic Explanations for Black-box Recommender Systems Using Interaction Graphs and Link Prediction Techniques

Marta Caro-Martínez, Guillermo Jiménez-Díaz, Juan A. Recio-García

Department of Software Engineering and Artificial Intelligence, ComplutenseUniversity of Madrid, Madrid (Spain)

Received 16 March 2021 | Accepted 29 September 2021 | Published 14 December 2021

Abstract

Explanations in recommender systems are a requirement to improve users' trust and experience. Traditionally, explanations in recommender systems are derived from their internal data regarding ratings, item features, and user profiles. However, this information is not available in black-box recommender systems that lack sufficient data transparency. This current work proposes a local model-agnostic, explanation-by-example method for recommender systems based on knowledge graphs to leverage this knowledge requirement. It only requires information about the interactions between users and items. Through the proper transformation of these knowledge graphs into item-based and user-based structures, link prediction techniques are applied to find similarities between the nodes and to identify explanatory items for the user's recommendation. Experimental evaluation demonstrates that these knowledge graphs are more effective than classical content-based explanation approaches but have lower information requirements, making them more suitable for black-box recommender systems.

Keywords

Black-box Recommender Systems, Explainable Artificial Intelligence, Graph Knowledge, Graph Representation, Link Prediction Techniques.

DOI: 10.9781/ijimai.2021.12.001

I. Introduction

Recommender systems are one of the essential tools on the Internet today [1]. They are set up on many platforms of e-commerce (Amazon, eBay) and entertainment (Netflix, Spotify), among others. They are necessary to help users find the most interesting products according to their interests. This task can be difficult for them due to the wide selection of products that they can access with new technologies [2]. However, recommender systems may not be as effective as we would expect. Many times, users do not trust this kind of system since they do not understand how a recommender system works and the reasons behind the recommendations [3], [4]. As a consequence, users do not put much attention on them. Because of this, explanations in recommender systems have appeared to solve this problem. Explanation systems try to clarify why a recommendation was provided for a target user [5].

In the literature, several approaches implement recommender systems based on classical techniques. Traditionally, these techniques involve collaborative filtering and content-based systems. Collaborative filtering systems use knowledge extracted from user ratings [6]. In the case of content-based systems, the recommendations are generated with the information about item descriptions and user preferences [7]. However, we must acknowledge that this useful knowledge is not always available when making explanations for a recommendation, sometimes because this information does not exist and other times because we cannot obtain it. For example, we find this problem when we want to provide explanations to black-box recommender system users [8], [9]. A black-box system is a system where users do not know how the method works, and they do not understand it. Therefore, as developers, we cannot access the recommendation process and use it as knowledge source to obtain explanations. [10]. In recent surveys [11], [12], authors focus on the wide use of knowledge representations and reasoning based on graphs to solve complex tasks. Furthermore, the authors classify recommender systems as a type of knowledge-aware application where the integration of knowledge graphs can enhance the reasoning behind the recommendation and, therefore, its interpretability and explainability. This is the reason why we have decided to focus our work on graphs. Our hypothesis is that graph-based explanations can achieve more effective explanations than other classical techniques.

Consequently, we propose a local model-agnostic surrogate explanation system for recommender systems that can be included in this knowledge-aware applications group. A local model-agnostic system is a post-hoc approach that tries to explain a black-box model's behaviour, focusing on a portion of the complete knowledge to provide explanations. In return, users can better understand how the system works because local models are more interpretable than global models, which use all the knowledge available and may be too complex [13], [14], [15], [16]. To tackle the problem of explaining recommendations

* Corresponding author.

E-mail addresses: martcaro@ucm.es (M. Caro-Martínez), gjimenez@ucm.es (G. Jiménez-Díaz), jareciog@ucm.es (J. A. Recio-García).

in black-box systems, we infer and use the knowledge within the interactions between users and items and represent them in graphs to provide explanations. We extend our previous work [17] defining two different approaches according to the entities that represent the graph nodes: the item-based approach and the user-based approach. We apply link prediction techniques on this knowledge to find similarities between the nodes [18], [19]. Among all link prediction techniques, we only consider those with which our proposal becomes a local model. These similarities are used to retrieve explanatory items to show to the user as an explanation for a recommendation. The explanatory items are a set of items with which the target user has interacted before. The user can compare these items with the recommended item and assess if the recommendation is suitable and interesting for them. Accordingly, the explanation is personalized for this user.

Moreover, our graph-based explanation method is independent of the recommender system: it does not require information about how the recommender works in obtaining the explanations, therefore it is suitable as a surrogate model for black-box recommender systems. The experimental evaluation performed in this paper compares the performance of both approaches -the user-based and item-based explanation approaches- to a content-based explanation method, which needs additional information in order to provide explanations. Although we can apply our graph-based methods to any recommendation system, including any black-box recommender, we have provided explanations for matrix factorization recommender systems in the evaluation because they are an excellent example of black-box recommender systems that achieve good results. Results demonstrate that our knowledge graph approach provides better explanations than content-based approaches, while having a remarkably lower information requirement.

The paper is structured as follows. First, Section II shows a review of the literature about explanation approaches in recommender systems and graph-based works. In Section III, we introduce our proposal: the item-based and the user-based graphs. We also present the similarity measures based on link prediction techniques that we have used in our approaches. Afterwards, in Section IV, we present the evaluation that was performed: the dataset, the experimental setup, and a discussion of the results that we obtain. Finally, we present the conclusions of this work and some future research in the last Section V.

II. Related Work

There are many state-of-the-art research studies about recommender systems. Many of them are reviews on this topic, and others are proposals of new techniques to make recommendations [1], [2], [20], [21]. In these works, we can observe two main classical techniques in recommender systems: collaborative filtering and content-based algorithms [22].

On the one hand, recommender systems based on collaborative filtering use the users' ratings to compute the recommendations [23]. On the other hand, the content-based systems take into account the item features and the user profiles to find the most interesting items to be recommended [7]. There are many research works focused on these two techniques and their effectiveness in recommendation tasks [24], [25], [26], [27], [28], [29], [30]. For example, we have some proposals related to our work in the works by Bobadilla et al. [31] and Cordobés de la Calle et al. [32] because they present recommender approaches that use knowledge about the user past interactions without considering rating values. Moreover, the work [32] uses graphs to represent the information and get the recommendations.

Explanation system research is a growing field in studies on recommender systems. When users do not understand why an item is suitable for them according to the recommendation system, they usually stop using these systems or decreasing their use because they do not feel confident with their results [1]. This issue is critical in some contexts and fields such as health care [33]. However, providing explanations in recommendations of daily activities, such as e-commerce or entertainment, is also remarkable because it increases the system's credibility and user's loyalty [34]. As a consequence, it increases users' trust and their use of recommender systems [1]. Therefore, one goal of explanation system research on recommendation systems is to increase the users' trust [35]. There are already many research reviews on and approaches to explanations in recommender systems. In the work by Zhan and Chen [36], we can see that the explanation approaches are based on users, items, and features, traditionally. Therefore, the knowledge extracted from collaborative filtering and content-based systems plays an essential role in classical techniques to provide explanations. The work by Nunes and Jannach [37] describes a taxonomy of explanations in recommender systems. To develop this model, the authors delve into a large amount of research in this field. It was also an essential reference for our previous work [38], where we built a theoretical model to classify explanation approaches and an ontology of explanation approaches in recommender systems: ExRecOnto[1]. We have found other taxonomies about explanation approaches that inspired our previous work, such as the works by Friedrich and Zanker [9] and Papadimitriou et al. [39]. In other works, we can see new proposals for explanation approaches. In the paper by Herlocker et al. [10], we find a classic work that proposes some different types of visualization of explanations in collaborative filtering recommender systems. They use the information extracted from collaborative filtering to provide explanations. Gedikli et al. [40] also present new ways to visualize explanations based on the work of Herlocker et al. In the recent work by Kouki [41], we observe different styles of explanation approaches: user-based, item-based, content-based, social-based, and item popularity.

Moreover, there are more innovative explanation approaches. For example, in the work by Quijano-Sanchez et al. [42], an explanation approach for group recommender systems is proposed based on social information. Andjelkovic et al. [43] describe a music recommender system that includes information about the recommendation through a graph. The approach allows the users to interact with the interface to change their preferences. In the work by Wang et al. [44], the authors describe a new proposal named the Tree-enhanced Embedding Method (TEM). TEM uses models based on embeddings and trees to provide explanations using knowledge extracted from collaborative filtering and latent factors. In our previous work [45], we propose a new way to explain recommendations based on matrix factorization. We use the information from latent factors to build a case-based reasoning [46] system that retrieves explanatory cases [47].

Knowledge representation in graphs and reasoning based on these structures are useful techniques to solve challenging problems [11], [48], [49], [50], [51], [52], [53]. The work by Ji et al. [11] describes a complete review on this topic. They categorize the work on knowledge graphs, and according to the classification proposed in this work, we find four types of graph-based knowledge research works: knowledge acquisition, knowledge representation learning, temporal knowledge graphs and knowledge-aware applications. The last group is the most interesting for this paper because it includes recommender systems as applications that can be enhanced using knowledge graphs.

Link prediction techniques are one of the essential bases of our work. We use the metrics from link prediction techniques to find similarities between our system's items and users. Some research works review these techniques [18], [19], [54], [55] and their application on social networks and recommender systems [56]. There are some approaches that make recommendations and that use graphs with or without link

[1] Available at: https://gaia.fdi.ucm.es/ontologies/#exreconto

prediction techniques. For example, in the work by Chiluka et al. [57], the authors describe an approach based on a user-item graph that employs link prediction techniques to collect the recommendations in User-Generated Content systems (UGCs) such as YouTube or Flickr. In the case of the work in [58], the authors do not use these link prediction techniques. However, they present a graph-based approach that combines content-based and collaborative knowledge for digital libraries. Chen et al. [59] propose a recommender system based on interaction graphs and collaborative filtering. Wang et al. [60] present a new system that uses graph representation to enrich news recommendations. We also find interesting research in the work by Shahmohammadi et al. [61]. The authors define the new concept "collaborative path", which refers to the use of collaborative filtering based on the user interaction background. They use "collaborative path" to create new proximity measures and recommendation algorithms based on link prediction techniques for online social networks, such as Facebook. Another example is the work [62], where the authors employ a bipartite network projection to provide recommendations.

On the topic of explanation approaches, there are a few research works about knowledge graphs. We encounter some recent works that review the role of knowledge graphs in the Explainable Artificial Intelligence (XAI) field, identifying the necessities that they cover [12], [63], [64]. We can also find explanation approaches based on knowledge graphs, which are different from our proposal. With our proposal, we provide explanations-by-example without considering the knowledge from the recommender system, only information about interactions. It requires a minimum amount of knowledge, while other similar systems use additional information. For example, in the work by Barbieri et al. [65], an explanation system using graphs and link prediction techniques is proposed. They provide an explanation using the reason why a link exists. To do this, the authors include latent factors that represent the user's preferences. Xian et al. [66] describe a new method called Policy-Guided Path Reasoning (PGPR), which uses a knowledge graph to generate recommendations. PGPR takes into account real paths in the graph to create explainable recommendations. Therefore, it uses additional information from the white-box recommender system to provide explanations. We also have to mention the work by Wang et al. [67], which introduces a new model Knowledge-aware Path Recurrent Network (KPRN). KPRN also uses knowledge graphs to make recommendations and to collect better results than other models such as collaborative knowledge base embedding or the neural factorization machine. One remarkable feature of the model is that it is also interpretable. Again, this method needs to use data from the recommendations to obtain explanations. However, there is much work to do in the field of explanations for recommender systems using graphs. These structures can represent a wide and varied knowledge, which is likely to be useful to justify recommendations.

III. The Knowledge Graph Explanation System

In this paper, we propose a novel knowledge-light, explanation system for black-box recommender systems. Our method only requires the knowledge inferred from the interactions between users and items within the recommender system to generate example-based explanations. Because our proposal is independent of the recommender process, it is suitable to support black-box recommender systems, where its working and knowledge is not available to obtain explanations. Thus, it can be considered a model-agnostic surrogate method, as depicted in Fig. 1.

Example-based explanations use previous items that the user liked and and similar to the recommended one. Additionally, theses explanations can also present items that users similar to the target user liked. Thus, every explanation is personalized for the target user, who will check the provided recommendation's suitability compared with the explanatory items.

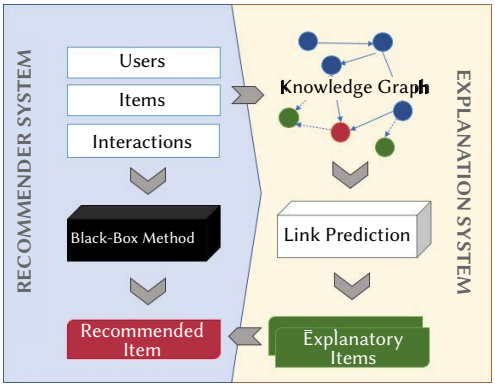


Fig. 1. General overview of the explanation approach using interaction graphs.

It is important to note that our method is designed to be knowledge-light, not requiring any other additional information such as rating values, descriptions, etc. Although additional information may increase the explanation's accuracy, it also increases the dependency on the recommendation algorithm. Our goal is to propose a model-agnostic explanation method with a minimum knowledge requirement but that achieves an acceptable performance compared with other explanation algorithms that require a similar or even higher knowledge level regarding the underlying recommendation process. It is also remarkable that our method proposed is totally independent of the recommendation system. It does not require any knowledge or reasoning that the recommender method uses. Therefore, we can apply our explanation method to any recommendation system. This is the reason why our graph-based explanation system is suitable to be applied on black-box recommender systems.

The general overview of our explanation method is presented in Fig. 1 and is as follows. First, we create knowledge graphs (described in Subsection A) using the interactions performed by the user within the recommender system. We define two different approaches to select the explanatory examples: *item-based* and *user-based* knowledge graphs.

The item-based graph represents connections between items, where links between two nodes are created when at least one user has consumed both items. The weight of the link is the number of users who have interacted with both items. Then, given a recommended item i for target user u, we compute the similarity between i and potential explanatory items (items that u liked previously), which is calculated using link prediction techniques (Subsection B). Finally, the most similar items will be presented as explanatory items. This approach is described in Subsection C.

Alternatively, the user-based graph describes connections between users who have interacted with at least one item in common, where link weights represent the number of common items consumed by the two users. Here, given a target user u, link prediction techniques are used to find other potential similar users that may consume items related to recommendation i. Next, the items already consumed by these similar users and u are aggregated and selected as explanatory items for u. This alternative approach is described in Subsection D.

We have provided the functional description of our method, and the following subsection describes the acquisition process to create the required item-based and user-based knowledge graphs.

A. Knowledge-Graph Acquisition

We can define an interaction for a recommender system as an action that a user has carried out with an item, such as watching a movie, rating a book or buying a product.

The most common interaction used in recommender systems is the rating of items. A rating action can be represented as a tuple $R = (t, u, i, x)$, where t is the timestamp when the interaction was performed, u is the user that went through with the interaction, i is the item with which u interacted, and x is the value associated with the rating. For instance, x could be the rating value (such as "5 stars") with which user u has assessed item i. However, this representation is still valid for any other kind of interaction where x is empty, and no additional information is associated with the interaction (such as "watching a movie"). This way, we assume a "minimal knowledge scenario" for our proposal, as it does not distinguish between positive, negative or neutral interactions. For example, it is positive if the user has rated a movie 4 or 5 stars. In contrast, a user has negatively interacted with a movie if they rated it less than 4 stars. A neutral interaction is defined as when x is empty. However, our model does not make this distinction and represents all of them equally: the x value is not required.

Taking negative interactions into account may seem useless or even harmful, but we think that we may lose important knowledge if we delete them. There are three main goals to include negative interactions in the model:

1. **Help users to find a correlation between the recommnedations and past interactions**. Items, which the target user did not like, have attributes that she may like. The user had at least a minimum interest in this item. Therefore, they could find recommendations similar to it helpful. For example, a user has watched a horror movie that she did not like. It does not mean that she does not like this type of movie necessarily. Maybe this particular movie was not great for her. Therefore, if the system recommends a new horror movie, then the movie that she had watched previously can be a reasonable explanation for this recommendation. She can see why the system provided this recommendation and the connections between the movies according to her preferences.
2. **Help to discard recommendations**. It could be valuable for the target user to discard recommendations that are not interesting. Although the recommender system had used the target user preferences to get a recommendation, this item might not be a good recommendation for her. Users have many different interests, and they need different products depending on their context [42]. For example, a target user wants to watch a movie on Halloween with her friends. They want to watch a horror movie. Although the target user likes Disney movies, these movies are not suitable for her in this context. With negative interactions, she can assess the recommendations better.
3. **Negative interactions are helpful to provide trust and loyalty**. Users need to trust the system, and we can only provide trust if users know how the system works [35]. The target user needs to know why the system provides a recommendation that is not interesting for her. She can understand how the system works, even if the system is mistaken. This information provides trust and loyalty, increasing user satisfaction [34].

Therefore, an interaction represents a relation between entities from *user set U* and *item set I*. This relation can be represented in an adjacency matrix $A = A_{ui}$, where A_{ui} represents if an interaction has occurred between user $u \in U$ and item $i \in I$. If the interaction has occurred, then $A_{ui} = 1$; otherwise, $A_{ui} = 0$, so the link does not exist. The graph built using this adjacency matrix is a bipartite graph: the nodes represent both user (from user set U) and item (from item set I) entities, and the relation is always from u to i. The semantic description of this relation is as simple as "user $u \in U$ has interacted with the item $i \in I$".

However, we can transform this graph by applying a bipartite network projection (Fig. 2) to create two different types of knowledge graphs that will ease our task of generating explanations: an *item graph* and a *user graph*.

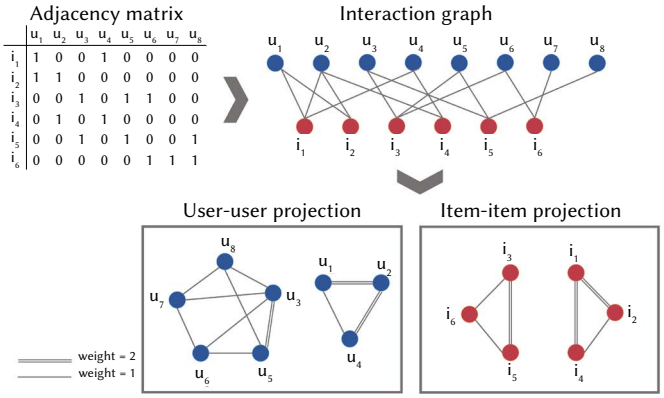


Fig. 2. Transformation of a bipartite graph into a nonbipartite graph through *bipartite network projection*.

On the one hand, in the item-based graph, nodes represent entities from the items set *I*. A link between two items i and j represents that "at least one user in common has interacted with both i and j". The link's weight is the number of common users that interacted with i and j.

On the other hand, the user-based graph only contains entities from user set U. In this graph, a link between two users u and v represents a relation between users, whose semantic description is "u and v have interacted with at least one item in common". As in the previous graph, the weight of the link represents the number of items in common.

These knowledge graphs are the primary building block of our explanation method, as they are used to find the most relevant items to show to the target user as explanatory items. To identify these items, we apply link prediction techniques to compute the similarities between the nodes in the graphs. We have considered and evaluated several similarity metrics from the link prediction literature, which are described in the following section.

B. Link Prediction Metrics

Link prediction techniques are algorithms from social network analysis that predict new links that will appear in a graph [18], [19], [54], [55]. There are several types of metrics to predict these links, but we focus on the similarity-based approaches due to our proposal's nature. The similarity-based metrics are, in turn, divided into four groups: node-based (they use the node properties), neighbour-based (they take into account the features between neighbours of the nodes), path-based (they define paths between the nodes), and walk-based (they use transition probabilities between nodes and neighbours). We choose to stress node-based and neighbour-based approaches because they are local models. These models focus on a local section of the knowledge, which is suitable for providing a concrete explanation without considering the whole knowledge represented in the graph [13], [14], [15], [16]. Local models are more interesting than global ones because they are more interpretable for target users, and we do not need to reach explanatory cases far from our target node. Therefore, the explanatory examples collected with the link prediction methods can be more useful and suitable for our proposal than other global link prediction techniques. The metrics used in the current study are the ones proposed in our previous graph-based recommender and explanation approaches [17], [68], [69], [70]. They are a variation of the classic link prediction metrics, and some of them can be divided into two versions: weighted and unweighted. Although similarity

metrics are commonly measured in the range of [0,1], our approach defines these similarity metrics more as a scoring function to rank similar items.

To clarify the description of the similarity metrics described here, we give some notation:

- $N(i)$ represents the set of neighbours of node i.
- $|N(i)|$ represents the number of neighbours (or node degree) of node i.
- L_{ij} represents the weight w of the link between nodes i and j.
- $W(i) = \sum L_{ix}: x \neq i \in I$ represents the *weighted node degree* of node i, which indicates the sum of the weights of the links directly connected with node i.

The link prediction metrics used in our explanation system are as follows:

Edge Weight (EW). This metric measures the similarity between two nodes as the weight of the link between them. $L_{ij} = 0$ represents that node i and node j are not connected. An unweighted version of this metric exists ($L_{ij} = 1$ if the link exists, 0 otherwise), but we have not used it because it is too simple.

$$EW(i,j) = L_{ij} \tag{1}$$

Common Neighbours (CN). Using this metric, the similarity between two nodes is the number of neighbours they have in common. The greater the intersection of the neighbour sets of any two nodes is, the greater the chance of a future association between them. Weighted Common Neighbours (WCN) is the weighted version of this metric.

$$CN(i,j) = |N(i) \cap N(j)| \tag{2}$$

$$WCN(i,j) = \sum_{z \in N(i) \cap N(j)} L_{iz} + L_{jz} \tag{3}$$

Jaccard Neighbours (JN). This metric is an improvement of $CN(i, j)$, as it measures the number of common neighbours of i and j compared with the number of total neighbours of both nodes. It does not have a weighted metric version.

$$JN(i,j) = \frac{|N(i) \cap N(j)|}{|N(i) \cup N(j)|} \tag{4}$$

Adar/Adamic (AA). This metric also measures the intersection of neighbour sets of two nodes in the graph, emphasizing the smaller overlap. The weighted version of this metric is Weighted Adar/Adamic (WAA).

$$AA(i,j) = \sum_{z \in N(i) \cap N(j)} \frac{1}{log|N(z)|} \tag{5}$$

$$WAA(i,j) = \sum_{z \in N(i) \cap N(j)} \frac{L_{iz} + L_{jz}}{log(1 + W(z))} \tag{6}$$

Preferential Attachment (PA). This metric is based on the consideration that nodes create links, with higher probability, with those nodes that already have a larger number of links. The probability of creating a link between nodes i and j is computed as the product of the degree of nodes i and j; therefore, the higher the degree of both nodes is, the higher is the probability of linking. This metric has the drawback of leading to high probability values for highly connected nodes to the detriment of the less connected nodes in the network. Weighted Preferential Attachment (WPA) is the weighted version of this metric. It is an improvement of PA in which the link weights are taken into account when computing the degree of nodes i and j.

$$PA(i,j) = |N(i)| \cdot |N(j)| \tag{7}$$

$$WPA(i,j) = W(i) \cdot W(j) \tag{8}$$

We have described the link prediction techniques, and the following sections present the contributions of this paper.

C. The Item-Based Explanation Method

The explanation process is defined for a target user u, who accepts the recommendation of item i. The goal of our system is to retrieve the best list of explanatory items $E = [e_1, e_2, ..., e_k]$ that helps u understand why the black-box recommender system recommended i. Therefore, this explanation-by-example method consists of displaying items similar to i that u previously interacted with.

This method uses the item-based knowledge graph to find explanatory examples. We define our item graph as $G_i = \langle I, L \rangle$, where I is the set of nodes representing the items, and L is the set of links that connect the nodes. We can define links as $L = \{(i, j, w) | i \neq j \in I\}$. Nodes i and j represent the items connected by a link, and w is the weight of the link. As described before, the weight w is the number of common different users that have interacted with both items. Due to the graph's high density, we decided to apply a preliminary filter to remove low representative links. Therefore, we define a threshold parameter δ_w to remove all links whose weight w is lower than its value.

The process for creating the list of explanatory items E is as follows (Fig. 3):

- **Step 1**. We build a similarity matrix S with the similarity scores between all nodes in the graph using the link prediction metric lp. Thus, the $S(i, j) = lp(i, j)$ value corresponds to the similarity between items i and j computed by the link prediction metric lp, as defined in Section B.
- **Step 2.** We build a set of candidate explanatory items $E' = \{(e_1, s_1), (e_2, s_2), ..., (e_n, s_n)\}$ that includes the items most similar to i using the similarity values in S. Value $s_x = S(i, e_x)$ represents the similarity between i and the explanatory item e_x.
- **Step 3**. We filter the candidate explanatory items already consumed by the target user by removing from E' all items in this set with which u has not interacted yet.
- **Step 4**. We rank E' in decreasing order using the similarity scores of the items (s_x). Finally, the top k items in this sorted list are returned as the explanatory items $E = [e_1, e_2, ..., e_k]$ for recommendation i and target user u.

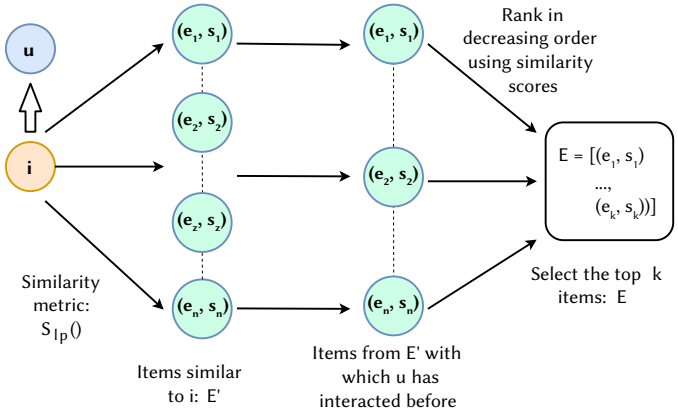


Fig. 3. Process of selecting the explanatory items with our item-based approach.

D. The User-Based Explanation Method

This alternative method is based on the user-to-user graph. The graph $G_u = \langle U, L \rangle$ represents the set of the user entities U as nodes, and the set of the links L is noted as $L = \{(u, v, w) | u \neq v \in U\}$. Analogously to the previous section, a link connects two nodes u and v when they

have interacted with at least one item in common. The weight of link w is the number of items with which both have interacted. Again, as it is a high-density graph, we apply a threshold parameter δ_w to remove all links whose weight w is lower than its value.

The process for creating the list of explanatory items E for target user u and a recommended item i is as follows (Fig. 4):

- **Step 1**. We build a similarity matrix S that stores all similarity scores between every pair of nodes of the graph using the link prediction metrics lp. Therefore, the $S(u,v) = lp(u,v)$ value corresponds to the similarity value between users u and v using the similarity metric lp. Again, lp is one of the link prediction metrics proposed in Section B.
- **Step 2**. From S, we build the set $V = \{v_1, v_2, ..., v_n)\}$ containing the n most similar users to u.
- **Step 3**. For every related user v, we obtain the set of items that has interacted with: $I_v = \{(e_1, s), (e_2, s), ..., (e_m, s\}$. Here, the similarity s associated with each item is the similarity between the target user u and the user v, that is, $s = S(u, v)$. Therefore, all of the items in I_v have the same similarity value.
- **Step 4**. Next, we build the set of candidate items for the target user by joining the items that similar users have interacted with: $E' = \bigcup_{v \in V} I_v$. Duplicated items are stored only once, and their associated similarity is the highest value found among all repetitions in the set.
- **Step 5**. We filter E' by removing all of the items that the target user has not interacted with yet.
- **Step 6**. Finally, the list of explanatory items $E = [e_1, e_2, ..., e_k]$ is created by sorting E' in decreasing order according to the similarity values associated with each item, and selecting the first k elements.

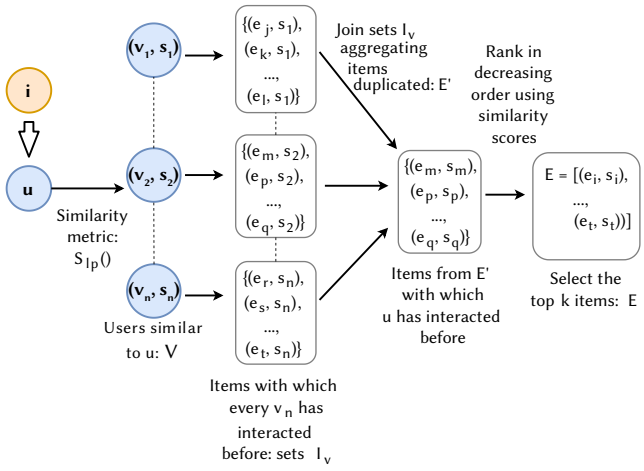


Fig. 4. Process of getting explanation items with our user-based approach.

We have presented our two surrogate methods, and the following section presents their experimental evaluation.

IV. Evaluation

Our experimental evaluation goal is to demonstrate that our method achieves similar performance as other explanatory approaches while having a lower knowledge requirement. The code implementation of the evaluation carried out with the graph-based methods is on a GitHub repository[2] to make more accessible the reproducibility of our proposal.

[2] https://github.com/martcaro/GraphBasedExplanations

In the literature, we can encounter that the items descriptions and item features are one of the knowledge used the most to generate explanations in recommender systems [36]. Therefore, we designed a comparative evaluation against a content-based system [46] that uses the item descriptions to find explanatory examples. Similar to our methods, this system will provide explanatory cases for a target user and a recommended item. However, it requires information about the features of the items in order to retrieve the explanatory examples. In addition, it is a global model because it uses all knowledge available. The evaluation hypothesis is that our proposal can find explanatory items that are as useful as the ones retrieved by the content-based system while using less information about items, only a portion of the knowledge represented by the interaction graphs.

We also need to use a recommender system to evaluate the models. Although we do not use knowledge about the recommendation process, we need to have a list of recommendations based on actual data. The models evaluated will try to explain this list. As we are dealing with model-agnostic surrogate models, we do not need information from the recommendation process and we can provide explanations to any recommender system. However, we consider that evaluating our model with a black-box recommender is more interesting. Because we do not need information about the recommendation process, we can validate our hypothesis that the graph-based explanation methods can retrieve useful explanatory examples without using rating values, descriptions or other additional information. Taking this into account, we decided to use a matrix factorization recommender system in our evaluation. Matrix factorization is one of the most e˙ective algorithms to make recommendations nowadays [47]. Nevertheless, it is not transparent for target users, and it is not easy to understand why an item has been recommended to them. Hence, matrix factorization recommendations are a good example that fits the need for an explanation system. In the evaluation process described next, we use the recommender as a black-box system, and our graph-based methods do not use any other information from the recommender to obtain the explanatory examples. Furthermore, we have used a dataset from the movie domain, as it is one of the most widely used to evaluate recommender systems. Next, we explain the evaluation process. In Section A, we describe the dataset that we have used to perform the evaluation. In the following Section B, we relate the experimental process itself. Finally, Section C discusses the results of the evaluation.

A. Dataset

In the experiment, we combine two different datasets for building our evaluation test set. On the one hand, we use the 100K MovieLens dataset[3] because it is a common choice to evaluate recommender systems. The MovieLens dataset contains 100K ratings in a set of tuples $R = (t, u, i, x)$, where u is the user who has watched movie i, t is the timestamp when u has rated i and x is the rating provided by u for i. Only the $\langle u, i \rangle$ pair is the information required by our explanation method. However, this evaluation aims to demonstrate that the quality of the explanations is similar to that for a content-based approach with a higher knowledge requirement. Therefore, we require an additional dataset with extra information about the recommended items. The chosen dataset is the IMDB dataset[4]. This dataset contains feature information about 5,000 movies such as genres, actors, and directors. This information will complement the MovieLens dataset to implement the content-based explanation system. However, not all movies in the MovieLens dataset appear in the IMDB one. Hence, we filter the MovieLens dataset to retain the movies that also appear in the IMDB dataset. We denote the resulting dataset as D. We divide D into the training set D_t with 90% of the interactions and the test set D_e with the remaining 10%.

[3] https://grouplens.org/datasets/movielens/100k/

[4] https://www.imdb.com/

Before performing the evaluation, we run an exploratory analysis on the dataset, following the model proposed by Dooms [71] that we use in previous works [17], [68]. Table I shows the results of this analysis.

TABLE I. Analysis of the Datasets Used in the Evaluation. *ML* is the Original MovieLens Dataset With 100K Interactions

Metric	ML	D	D_t	D_e	B_e
# Ratings	100,000	11,477	10,330	1,147	280
# Items	1,682	164	164	145	109
# Users	943	587	584	394	134
Density	0.06	0.12	0.11	0.02	0.02
Items					
Maximum # ratings per item	583	329	305	30	10
Median # ratings per item	27	43.5	39	5	2
Average # ratings per item	59.45	69.98	62.99	7.91	2.57
Minimum # ratings per item	1	1	1	1	1
Users					
Maximum # ratings per user	737	128	113	15	11
Median # ratings per user	65	12	11	2	1
Average # ratings per user	106.05	19.55	17.69	2.91	2.09
Minimum # ratings per user	20	1	1	1	1
Ratings					
% Ratings ≥4	55.38	52.54	52.66	51.44	37.50
% Ratings <4	44.62	47.46	47.34	48.56	62.50

From this analysis, we found a significant bias in dataset D_e: it is unbalanced regarding the number of items associated with each rating value. As a consequence, we decided to create a stratified test set, B_e. B_e avoids this bias because it contains the same number of items (35) for each rating value. We have selected this amount of items because it is the minimal amount for a rating value (2.5 stars).

B. Experimental Setup

The experimental process starts by building the graph-based and content-based explanation methods on the training set. We implemented several versions of the graph-based methods regarding the similarity metrics described in Section B: AA, CN, EW, JN, PA, WAA, WCN, and WPA. We also configured the threshold $\delta_w = 5$ since we considered this value sufficient to reduce the density without removing essential knowledge.

The content-based explanation system retrieves the items most similar to the recommended item i_r taking into account the movie features in the IMDB dataset. In this case, we evaluate Euclidean, Cosine, and Jaccard methods as similarity metrics. Both explanation methods generate the list E of explanatory items sorted in decreasing order. The list size is adjusted by the k parameter, which has been evaluated within the range $k \in [1,10]$.

The experimental process continues by measuring the effectiveness of both explanation systems against two test sets: D_e and B_e. Each evaluation is repeated 100 times, where we randomly obtain B_e from D_e for each iteration. The explanation methods' effectiveness is evaluated from the similarity between the ratings of recommended item i_r and the explanatory items in E_k. To do this, we employ the Root Mean Square Error (RMSE) metric to compare the rating for i_r predicted by the matrix factorization recommender system and the actual ratings of the explanatory items in E_k retrieved by each method. It is important to note that, for each k, we have removed the users who do not interact with at least k movies in the test set. Thus, if a target user u_t has only rated four movies, then she is not suitable to be evaluated when the list of explanatory items has $k \geq 5$.

C. Discussion

Table II and Table III report the results obtained with the original dataset and the stratified dataset, respectively. For each table, we show the performance of the graph-based and content-based methods. It is remarkable that there are similar scores for both datasets, but not among the methods being evaluated.

TABLE II. Results of the Evaluation With the Original Test Set (D_e). The Column Values Correspond to k From 1 to 10. The Best Results Among Similarity Metrics Are in Bold. The Best Evaluation Metric Values Are Underlined

	1	2	3	4	5	6	7	8	9	10
I - AA	1.095	0.865	0.792	0.769	0.747	0.723	0.702	0.686	0.678	0.664
I - CN	1.035	0.821	0.750	0.713	0.683	0.666	0.649	0.640	0.636	0.635
I - EW	1.087	0.878	0.797	0.754	0.734	0.706	0.679	0.666	0.651	0.639
I - JN	**0.961**	**0.734**	**0.658**	**0.624**	**0.599**	**0.573**	**0.562**	**0.549**	**0.540**	**0.534**
I - PA	1.126	0.897	0.806	0.789	0.782	0.762	0.741	0.728	0.718	0.706
I - WAA	1.113	0.968	0.908	0.855	0.822	0.794	0.775	0.747	0.723	0.705
I - WCN	1.113	0.968	0.907	0.855	0.821	0.796	0.778	0.749	0.723	0.704
I - WPA	1.115	0.970	0.910	0.852	0.821	0.793	0.779	0.747	0.723	0.703
U - AA	0.865	0.798	0.759	0.741	0.720	0.705	0.696	0.690	0.688	0.685
U - CN	0.877	0.788	0.758	**0.734**	**0.715**	0.701	**0.692**	0.688	0.685	0.681
U - EW	0.871	**0.778**	**0.744**	0.736	0.718	**0.699**	0.695	**0.684**	0.688	0.684
U - JN	0.874	0.805	0.777	0.745	0.728	0.711	0.704	0.703	0.695	0.686
U - PA	0.865	0.791	0.752	0.742	0.720	0.706	0.695	0.690	0.689	0.685
U - WAA	**0.864**	0.793	0.753	0.743	0.720	0.706	0.695	0.690	0.690	0.684
U - WCN	**0.864**	0.793	0.753	0.743	0.720	0.706	0.695	0.690	0.690	0.684
U - WPA	0.865	0.791	0.752	0.743	0.720	0.706	0.695	0.690	0.690	0.684
Cosine	0.973	1.036	1.064	**1.078**	**1.087**	1.100	1.101	1.104	1.108	1.111
Euclidean	**0.966**	**1.032**	**1.063**	1.079	1.092	**1.092**	**1.099**	**1.100**	**1.102**	**1.105**
Jaccard	0.974	1.037	1.064	**1.078**	**1.087**	1.099	1.101	1.104	1.109	1.111

TABLE III. Results of the Evaluation With the Stratified Test Set (B_e). The Column Values Correspond to k From 1 to 10. The Best Results Among Similarity Metrics Are in Bold. The Best Evaluation Metric Values Are Underlined

	1	2	3	4	5	6	7	8	9	10
I - AA	1.164	0.947	0.861	0.839	0.817	0.796	0.774	0.764	0.754	0.738
I - CN	1.084	0.893	0.811	0.761	0.739	0.721	0.696	0.695	0.695	0.698
I - EW	1.142	0.938	0.863	0.827	0.802	0.766	0.743	0.736	0.714	0.702
I - JN	**1.004**	**0.748**	**0.673**	**0.643**	**0.629**	**0.597**	**0.584**	**0.573**	**0.566**	**0.566**
I - PA	1.182	1.000	0.902	0.883	0.873	0.852	0.835	0.819	0.809	0.793
I - WAA	1.156	1.061	1.002	0.944	0.922	0.894	0.874	0.838	0.810	0.790
I - WCN	1.149	1.054	0.994	0.937	0.917	0.890	0.873	0.837	0.807	0.787
I - WPA	1.152	1.065	1.011	0.940	0.921	0.889	0.873	0.837	0.812	0.789
U - AA	0.883	0.830	0.794	0.765	0.741	0.731	0.717	0.716	0.715	0.713
U - CN	0.898	0.820	0.787	0.758	0.740	0.727	0.715	0.715	0.713	0.710
U - EW	0.888	**0.804**	**0.758**	**0.749**	**0.733**	**0.710**	**0.707**	**0.698**	**0.706**	**0.704**
U - JN	0.881	0.833	0.811	0.771	0.755	0.738	0.733	0.733	0.725	0.716
U - PA	0.883	0.820	0.778	0.765	0.740	0.727	0.709	0.708	0.711	0.708
U - WAA	0.879	0.821	0.779	0.766	0.740	0.728	0.710	0.709	0.712	0.708
U - WCN	0.881	0.823	0.779	0.767	0.741	0.729	0.711	0.710	0.713	0.709
U - WPA	0.881	0.820	0.778	0.767	0.742	0.730	0.712	0.712	0.715	0.711
Cosine	1.117	1.130	1.125	1.125	1.125	1.120	1.129	1.125	1.121	1.121
Euclidean	1.090	1.105	1.102	1.110	1.105	1.106	1.106	1.111	1.112	1.112
Jaccard	**1.052**	**1.052**	**1.054**	**1.067**	**1.084**	**1.097**	**1.100**	**1.099**	**1.102**	**1.100**

In the Table II, we can observe the differences among the similarity metrics used in the evaluation and their performance when applied to the original dataset D_e.

For the item-based method, the scores are similar regarding the k parameter. However, outlier values correspond to the JN similarity

metric. With this similarity metric, we come across a better result for the RMSE for all values of k. JN always improves the performance of the other similarity metrics with a difference of approximately 10% to 20%. Our explanation for this behaviour is that JN considers the number of common neighbours compared with the number of total neighbours to obtain the items most similar to the recommended one. This indicates that this metric considers the similarity and the diversity of the sets. It may also indicate that the knowledge from negative interactions is useful for explaining recommendations.

However, we do not find this pattern in the user-based approach. The results of all the metrics are very similar. Moreover, there is no obvious best similarity metric. When $k = 1$, WAA and WCN are the metrics that perform better. They take into account the weight of the links. Therefore, they achieve higher performance when $k = 1$ because they exploit that information. For the rest of the values of k, the best similarity metric varies between CN and EW. With EW, we achieve the best results four times, while with CN, we reach the lower scores for five setups. Therefore, we can conclude that CN may be the best similarity metric for the user graph-based method. We can conclude that EW provides better results because it considers the links' weight to obtain explanatory items.

Regarding the content-based results, the best similarity metric is also clear. It is not as obvious as in the item-based method, but we can conclude that we obtain the best results with the Euclidean distance. Moreover, we can observe that the difference between the results is even lower than the difference found for the previous methods. These di˙erences are not significant.

We can obtain additional conclusions by comparing the best scores. This analysis is shown in Fig. 5. In this figure, we report the results of each approach with its best similarity metric: JN for the item-based approach, CN for the user-based approach, and the Euclidean distance for the content-based approach. On the one hand, we can see a heterogeneous behaviour regarding the k parameter. In the case of the content-based system, the RMSE value becomes higher with increasing k. However, in the case of the graph-based approaches, the performance improves. The error value stabilizes because the algorithm retrieves a larger amount of explanatory items; therefore, it is more diÿcult to make a significant mistake. On the other hand, the best performance values are always achieved by graph-based methods. When $k = 1$, the best results are achieved with the user-based proposal. In the rest of the cases, we have achieved the best results with the item-based proposal. We can also conclude that the item-based approach performs better because the recovery of similar items is straightforward and target users are familiarised with items with which they have interacted before. Moreover, explanatory examples using a justification based on similar users who they do not know can be less helpful.

Table III reports the results of the evaluation with the stratified dataset B_e. The trend in the results is similar to the trend that we come across in Table II. With the stratified dataset, we obtain worse results, but the difference is not remarkable: the bias does not have a relevant impact. However, the results are slightly worse with the stratified dataset because we remove the bias that we found in the original dataset.

In the item-based approach, we can see that JN is the similarity metric that performs better. Therefore, the bias of the dataset does not change the comparison among the imilarity metrics. For the user-based method, the discussion is also very similar to the previous one. We can observe that EW is the best similarity metric for all values of k except for $k = 1$. With $k = 1$, WCN, WPA, and JN achieve the best performance with the same value (0.881). Here, we can conclude that WCN acts as the best similarity metric with both the original and stratified datasets. For the rest of the values of k, EW is still the best metric, but CN worsens, although it achieves suÿcient results. Therefore, we can say that the similarity metrics that work with knowledge about the weight provide better results with the user-based approach. In the case of the content-based system, the conclusions change. We observe that the best result is achieved with the Jaccard similarity metric for all values of k.

To compare the performance of the three approaches when applied to the stratified dataset B_e, we have created Fig. 6. We have also decided to represent in this chart the best similarity metric for each method: the JN similarity metric for the item-based method, EW for the user-based method, and Jaccard for the content-based method. Again, we can see the same trend that we saw with the original dataset. The chart shapes are almost the same, although we do not achieve the best results with the same similarity metrics for the user-based approach and content-based system.

Considering the results analysed, we can also discuss the parameters chosen in the evaluation. On the one hand, we split the dataset into the training set and test set with the 90% and 10% of the interactions, respectively. It can a˙ect the results in terms of performance. If we decrease the number of interactions in the training set and increment the interactions in the test set, we may get lower RMSE values. The explanation models' performance depends on the amount of knowledge that we use to build them. It can a˙ects to both graph models and content-based methods. In the graph-based models, we will have a graph with a lower amount of nodes and links. Therefore it is more challenging to get a correct answer if we have a less amount of candidates. Equally, the content-based method

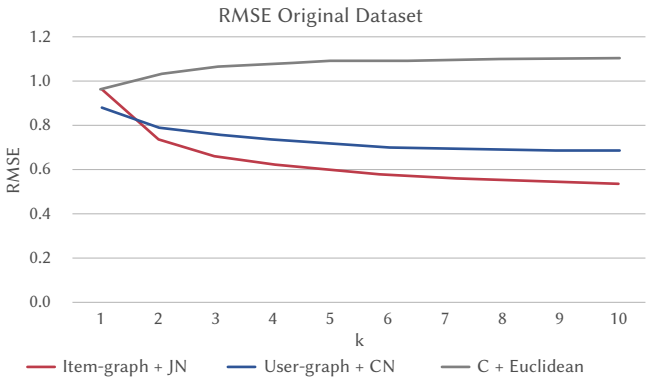


Fig. 5. Chart which represents the results got with the original dataset (D_e). For each approach studied, we have chosen the similarity metrics which performed better. In the axis Y, we represent the RMSE value. We consider the number of explanatory items retrieved in the axis X.

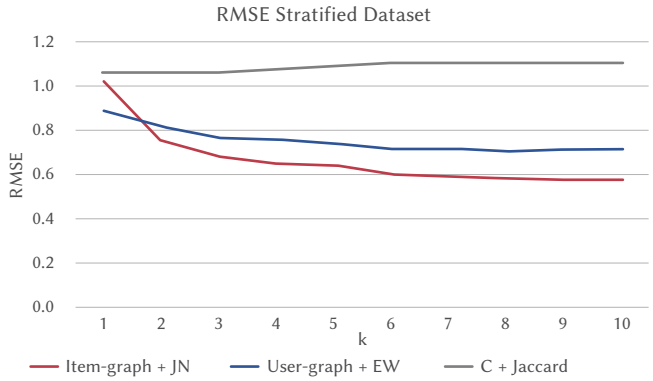


Fig. 6. The results got with the stratified dataset (B_e). For each approach, we have chosen the similarity metrics which performed better. In the axis Y, we represent the RMSE value. We consider the number of explanatory items retrieved in the axis X.

would have a smaller list of explanation example candidates. Then, the probabilities of finding similar items according to their attributes are lower. On the other hand, we delete the dataset's bias removing all the ratings whose value is lower than 2.5 stars. We have shown in the results above that, removing the bias, we can get better results. However, we did not analyse how the models' behaviour is if we had chosen other minimal value. When choosing a lower value, our theory is that we will get similar values to the evaluation carried out with the original dataset because the bias would not be removed. However, using higher values than 2.5, we would obtain worse results because we would be deleting knowledge.

Finally, we have decided to show the percentage of improvement of the graphs concerning the content-based system in Table IV. The two first rows correspond to the comparison between the item-based model using JN and the user-based model using CN with the content-based system using Euclidean distance, which are the models that performed the best in Table II. The last two rows show the comparison between the item-based model using JN and the user-based model using EW with the content-based system using Jaccard, which are the models that performed the best in Table III. We can see that the graph models enhance the performance of the explanations, becoming 50% better than the content-based ones.

TABLE IV. Percentage of Improvement of the Graph-based Methods Regarding the Content-based Model. The Bestpercentage of Improvement in Every Row Is Marked in Bold

	1	2	3	4	5	6	7	8	9	10
I-JN (D_e)	0.49	28.83	38.09	42.18	45.17	47.50	48.83	50.10	50.99	**51.64**
U-CN (D_e)	9.26	23.65	28.67	31.94	34.50	35.83	36.99	37.46	37.84	**38.35**
I-JN (B_e)	4.56	28.90	36.15	39.74	41.97	45.58	46.91	47.86	**48.64**	48.55
U-EW (B_e)	15.56	23.54	28.10	29.82	32.34	35.29	35.72	**36.45**	35.91	36.03

As a conclusion of this evaluation, the previous discussion validates our hypothesis. We achieve better results with the graph-based methods with both the original and stratified datasets than with the content-based method, regardless of the type of graph or the similarity metric used. Thus, we can consider that our graph-based proposals perform better than content-based approaches to provide explanations using a less amount of knowledge. Furthermore, we can conclude that the item-based approach performs better because it finds the explanatory items in a straightforward way.

V. Conclusions and Future Work

The current work proposes a novel local, model-agnostic, surrogate method to provide explanations for black-box recommender systems using knowledge graphs. This proposal is an alternative solution for when classical explanation techniques cannot be applied due to their requirements regarding the recommender system's input data or internal behaviour. Traditionally, these techniques involve collaborative filtering that requires ratings as the input knowledge or content-based methods that take into account item features or user profiles. However, in many scenarios, the knowledge required by these techniques is not available.

The minimum knowledge that can be obtained from a recommender system is the previous interactions between users and items. In this work, we propose only to use this knowledge to implement a surrogate explanation-by-example method for recommender systems. This proposal does not need information about ratings or descriptions or any other additional knowledge from the recommender system. Therefore, our proposal is suitable to support any type of recommender system, including black-box recommenders whose information is not available to obtain explanations. We hypothesize that we can provide explanations for black-box recommender systems using a minimum amount of knowledge while achieving the same or even better performance than the classical techniques.

We represent the interaction knowledge as a graph, where nodes are users and items and the links represent that the user has interacted with the item. Then, we apply a bipartite network projection obtaining two di˙erent knowledge graphs: an item-based graph and a user-based graph. The item-based graph has items as nodes, and their connections represent the number of users that have interacted with both items. Alternatively, the user-based graph represents users as nodes, and the weight of the link is the number of items with which the users have both interacted. Thus, we have two di˙erent graph structures to provide explanations.

In our method, we apply link prediction techniques to find similar nodes that lead to the discovery of explanatory examples. It is important to note that these link prediction techniques turn our approach into a local model. This implies that our approach is easier to interpret for target users.

One of this work's major novelties is that we consider all interactions performed by users (positive, negative and neutral). As a consequence, the explanation examples provided to the target users can be items that they did not like; therefore, they can decide better if the recommended item is of interest to them or not.

From the item-based graph, the identification of the explanatory examples is very straightforward, directly applying link prediction metrics in order to find the items most similar to the recommended one. Then, we filter these items, removing the ones with which the target user has not interacted yet. In the case of the user-based approach, the process is slightly more complicated. We apply the similarity metrics to find the users most similar to the target user. Then, we compute the items with which this set of users has interacted, removing those the target user has already interacted with. These items will be the explanatory examples to show to the target user.

Therefore, the explanation is personalized for each target user. However, it is important to note that our approach has the cold start problem, similar to many recommender and explanation systems. If there are not su˙cient interactions, then we cannot provide personalized explanations. Solving this problem could be a future line of research.

To validate our method, we performed an experimental evaluation. Its goal was to compare our method's performance with that of a global classical explanation-by-example technique with a higher knowledge requirement, that is, a content-based explanation system.

The evaluation dataset was created from the MovieLens and IMDB datasets. We used the RMSE metric to compare the performance achieved with the three approaches: item-based graph, user-based graph, and content-based. After a complete analysis of the evaluation results, we conclude that the hypothesis is correct, as the graph-based approaches achieve a higher performance than the content-based approach while requiring a lower level of knowledge. Furthermore, the more the list of explanatory items grows, the better the graph-based system's performance, in contrast with the behaviour of the content-based system. Globally, the item-based graph seems to be the most effective method when configured with the Jaccard Neighbours similarity metric.

For future work, we can outline some research areas, apart from solving the cold start problem. We want to validate our hypothesis with different datasets from other domains. For example, we believe that recommenders for the music domain can take advantage of our approach because their datasets usually lack ratings. We could compare our approaches with additional techniques, such as collaborative filtering or machine learning techniques, to confirm the hypothesis

that we verified in this work. We can also evaluate our models applying our explanation models to another kind of black-box recommender systems, apart from matrix factorization. Another research area is to apply new aggregation methods to the graph-based approaches, which we have already performed in previous recommender system proposals [68]. We can also use global link prediction techniques and compare their performances in our graph-based methods with this paper's results. Another essential area for future work is to evaluate with real users because they can provide a more accurate opinion about the graph-based approaches' effectiveness. Moreover, they can provide an analysis regarding their explanation goals, such as user trust or user satisfaction. To perform this evaluation, we would need to develop a visualization method for the explanations. For example, we could provide explanations based on textual justifications or more innovative and visual interfaces that use graph representation.

Founding

Supported by the UCM (Research Group 921330), the Spanish Committee of Economy and Competitiveness (TIN2017-87330-R) and the funding provided by Banco Santander at UCM (CT42/18-CT43/18).

References

[1] C. C. Aggarwal, *et al.*, *Recommender systems*. Springer, 2016.

[2] J. Bobadilla, F. Ortega, A. Hernando, A. Gutiérrez, "Recommender systems survey," *Knowledge-based systems*, vol. 46, pp. 109–132, 2013.

[3] D. Jannach, M. Jugovac, I. Nunes, "Explanations and user control in recommender systems," in *Personalized Human-Computer Interaction*, De Gruyter Oldenbourg, 2019, pp. 133–156.

[4] N. Tintarev, J. Masthoff, "A survey of explanations in recommender systems," in *2007 IEEE 23rd international conference on data engineering workshop*, 2007, pp. 801– 810, IEEE.

[5] R. Sharma, S. Ray, "Explanations in recommender systems: an overview," *International Journal of Business Information Systems*, vol. 23, no. 2, pp. 248–262, 2016, doi: 10.1504/IJBIS.2016.078909.

[6] M. D. Ekstrand, J. T. Riedl, J. A. Konstan, *et al.*, "Collaborative filtering recommender systems," *Foundations and Trends® in Human–Computer Interaction*, vol. 4, no. 2, pp. 81–173, 2011, doi: 10.1561/1100000009.

[7] P. Lops, M. De Gemmis, G. Semeraro, "Content-based recommender systems: State of the art and trends," in *Recommender systems handbook*, Springer, 2011, pp. 73– 105.

[8] R. Sinha, K. Swearingen, "The role of transparency in recommender systems," in *CHI'02 extended abstracts on Human factors in computing systems*, 2002, pp. 830–831, ACM.

[9] G. Friedrich, M. Zanker, "A taxonomy for generating explanations in recommender systems," *AI Magazine*, vol. 32, no. 3, pp. 90–98, 2011.

[10] J. L. Herlocker, J. A. Konstan, J. Riedl, "Explaining collaborative filtering recommendations," in *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, 2000, pp. 241–250, ACM.

[11] S. Ji, S. Pan, E. Cambria, P. Marttinen, S. Y. Philip, "A survey on knowledge graphs: Representation, acquisition, and applications," *IEEE Transactions on Neural Networks and Learning Systems*, 2021, doi: 10.1109/TNNLS.2021.3070843.

[12] Q. Guo, F. Zhuang, C. Qin, H. Zhu, X. Xie, H. Xiong, Q. He, "A survey on knowledge graph- based recommender systems," *IEEE Transactions on Knowledge and Data Engineering*, 2020, doi: 10.1109/TKDE.2020.3028705.

[13] M. T. Ribeiro, S. Singh, C. Guestrin, ""why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[14] J. Singh, A. Anand, "Exs: Explainable search using local model agnostic interpretability," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 2019, pp. 770–773.

[15] V. Arya, R. K. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilovic, *et al.*, "One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques," 2019.

[16] M. T. Ribeiro, S. Singh, C. Guestrin, "Anchors: High- precision model-agnostic explanations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[17] M. Caro-Martinez, J. A. Recio-Garcia, G. Jimenez-Diaz, "An algorithm independent case-based explanation approach for recommender systems using interaction graphs," in *International Conference on Case-Based Reasoning*, 2019, pp. 17–32, Springer.

[18] P. Wang, B. Xu, Y. Wu, X. Zhou, "Link prediction in social networks: the state-of-the-art," *Science China Information Sciences*, vol. 58, no. 1, pp. 1–38, 2015.

[19] D. Liben-Nowell, J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American society for information science and technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[20] F. Ricci, L. Rokach, B. Shapira, "Introduction to recommender systems handbook," in *Recommender systems handbook*, Springer, 2011, pp. 1–35.

[21] F. Isinkaye, Y. Folajimi, B. Ojokoh, "Recommendation systems: Principles, methods and evaluation," *Egyptian Informatics Journal*, vol. 16, no. 3, pp. 261–273, 2015, doi: 10.1016/j.eij.2015.06.005.

[22] L. Sharma, A. Gera, "A survey of recommendation system: Research challenges," *International Journal of Engineering Trends and Technology (IJETT)*, vol. 4, no. 5, pp. 1989–1992, 2013.

[23] J. B. Schafer, D. Frankowski, J. Herlocker, S. Sen, "Collaborative filtering recommender systems," in *The adaptive web*, Springer, 2007, pp. 291–324.

[24] J. Bobadilla, A. Hernando, F. Ortega, J. Bernal, "A framework for collaborative filtering recommender systems," *Expert Systems with Applications*, vol. 38, no. 12, pp. 14609–14623, 2011, doi: https://doi.org/10.1016/j.eswa.2011.05.021.

[25] B. M. Sarwar, G. Karypis, J. A. Konstan, J. Riedl, *et al.*, "Item-based collaborative filtering recommendation algorithms.," *WWW*, vol. 1, pp. 285–295, 2001.

[26] J. Bobadilla, F. Ortega, A. Hernando, J. Alcalá, "Improving collaborative filtering recommender system results and performance using genetic algorithms," *Knowledge-based systems*, vol. 24, no. 8, pp. 1310–1316, 2011.

[27] J. L. Herlocker, J. A. Konstan, L. G. Terveen, J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems (TOIS)*, vol. 22, no. 1, pp. 5–53, 2004.

[28] M. De Gemmis, P. Lops, C. Musto, F. Narducci, G. Semeraro, "Semantics-aware content-based recommender systems," in *Recommender Systems Handbook*, Springer, 2015, pp. 119–159.

[29] M. De Gemmis, P. Lops, G. Semeraro, P. Basile, "Integrating tags in a semantic content-based recommender," in *Proceedings of the 2008 ACM conference on Recommender systems*, 2008, pp. 163–170, ACM.

[30] C. Musto, G. Semeraro, M. de Gemmis, P. Lops, "Learning word embeddings from wikipedia for content-based recommender systems," in *European Conference on Information Retrieval*, 2016, pp. 729–734, Springer.

[31] J. Bobadilla, F. Ortega, A. Gutiérrez, S. Alonso, "Classification-based deep neural network architecture for collaborative filtering recommender systems.," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 6, no. 1, 2020, doi: 10.9781/ijimai.2020.02.006.

[32] H. Cordobés de la Calle, L. F. Chiroque, A. Fernández Anta, R. García, P. Morere, L. Ornella, F. Pérez, A. Santos, "Empirical comparison of graph-based recommendation engines for an apps ecosystem," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 3, no. 2, pp. 33–39, 2015.

[33] K. W. Darlington, "Designing for explanation in health care applications of expert systems," *Sage Open*, vol. 1, no. 1, p. 2158244011408618, 2011.

[34] B. Walek, V. Fojtik, "A hybrid recommender system for recommending relevant movies using an expert system," *Expert Systems with Applications*, vol. 158, p. 113452, 2020.

[35] N. Tintarev, "Explanations of recommendations," in *Proceedings of the 2007 ACM conference on Recommender systems*, 2007, pp. 203–206, ACM.

[36] Y. Zhang, X. Chen, "Explainable recommendation: A survey and new perspectives," *Foundations and Trends in Information Retrieval*, vol. 14, no. 1, pp. 1–101, 2020, doi: 10.1561/1500000066.

[37] I. Nunes, D. Jannach, "A systematic review and taxonomy of explanations in decision support and recommender systems," *User Modeling and User-Adapted Interaction*, vol. 27, no. 3-5, pp. 393–444, 2017, doi: 10.1007/s11257-017-9195-0.

[38] M. Caro-Martinez, G. Jimenez-Diaz, J. A. Recio-Garcia, "A theoretical model of explanations in recommender systems," *ICCBR 2018*, p. 52, 2018.

[39] A. Papadimitriou, P. Symeonidis, Y. Manolopoulos, "A generalized taxonomy of explanations styles for traditional and social recommender systems," *Data Mining and Knowledge Discovery*, vol. 24, no. 3, pp. 555–583, 2012, doi: 10.1007/s10618-011-0215-0.

[40] F. Gedikli, D. Jannach, M. Ge, "How should I explain? a comparison of different explanation types for recommender systems," *International Journal of Human-Computer Studies*, vol. 72, no. 4, pp. 367–382, 2014.

[41] P. Kouki, J. Schaffer, J. Pujara, J. O'Donovan, L. Getoor, "Personalized explanations for hybrid recommender systems," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 2019, pp. 379– 390, ACM.

[42] L. Quijano-Sanchez, C. Sauer, J. A. Recio-Garcia, B. Diaz-Agudo, "Make it personal: a social explanation system applied to group recommendations," *Expert Systems with Applications*, vol. 76, pp. 36–48, 2017.

[43] I. Andjelkovic, D. Parra, J. O'Donovan, "Moodplay: Interactive music recommendation based on artists' mood similarity," *International Journal of Human- Computer Studies*, vol. 121, pp. 142–159, 2019.

[44] X. Wang, X. He, F. Feng, L. Nie, T.-S. Chua, "Tem: Tree-enhanced embedding model for explainable recommendation," in *Proceedings of the 2018 World Wide Web Conference*, WWW '18, Republic and Canton of Geneva, Switzerland, 2018, pp. 1543–1552, International World Wide Web Conferences Steering Committee.

[45] J. Jorro-Aragoneses, M. Caro-Martinez, J. A. Recio-Garcia, B. Diaz-Agudo, G. Jimenez-Diaz, "Personalized case-based explanation of matrix factorization recommendations," in *International Conference on Case-Based Reasoning*, 2019, pp. 140–154, Springer.

[46] F. Sørmo, J. Cassens, A. Aamodt, "Explanation in case- based reasoning–perspectives and goals," *Artificial Intelligence Review*, vol. 24, no. 2, pp. 109–143, 2005, doi: 10.1007/s10462-005-4607-7.

[47] J. Bennett, S. Lanning, *et al.*, "The netflix prize," in *Proceedings of KDD cup and workshop*, vol. 2007, 2007, p. 35, New York, NY, USA.

[48] M. Nickel, K. Murphy, V. Tresp, E. Gabrilovich, "A review of relational machine learning for knowledge graphs," *Proceedings of the IEEE*, vol. 104, no. 1, pp. 11– 33, 2015, doi: 10.1109/JPROC.2015.2483592.

[49] L. Qiao, L. Zhang, S. Chen, D. Shen, "Data-driven graph construction and graph learning: A review," *Neurocomputing*, vol. 312, pp. 336–351, 2018.

[50] Q. Wang, Z. Mao, B. Wang, L. Guo, "Knowledge graph embedding: A survey of approaches and applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 12, pp. 2724–2743, 2017.

[51] H. Paulheim, "Knowledge graph refinement: A survey of approaches and evaluation methods," *Semantic web*, vol. 8, no. 3, pp. 489–508, 2017.

[52] Y. Lin, X. Han, R. Xie, Z. Liu, M. Sun, "Knowledge representation learning: A quantitative review," 2018, https://arxiv.org/abs/1812.10901.

[53] Y. Chong, Y. Ding, Q. Yan, S. Pan, "Graph-based semi- supervised learning: A review," *Neurocomputing*, 2020.

[54] B. Furht, *Handbook of social network technologies and applications*. Springer Science & Business Media, 2010.

[55] L. Lü, T. Zhou, "Link prediction in complex networks: A survey," *Physica A: statistical mechanics and its applications*, vol. 390, no. 6, pp. 1150–1170, 2011, doi: 10.1016/j.physa.2010.11.027.

[56] N. N. Daud, S. H. Ab Hamid, M. Saadoon, F. Sahran, N. B. Anuar, "Applications of link prediction in social networks: A review," *Journal of Network and Computer Applications*, p. 102716, 2020.

[57] N. Chiluka, N. Andrade, J. Pouwelse, "A link prediction approach to recommendations in large-scale user- generated content systems," in *European Conference on Information Retrieval*, 2011, pp. 189–200, Springer.

[58] Z. Huang, W. Chung, T.-H. Ong, H. Chen, "A graph- based recommender system for digital library," in *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, 2002, pp. 65–73, ACM.

[59] H. Chen, X. Li, Z. Huang, "Link prediction approach to collaborative filtering," in *Digital Libraries, 2005. JCDL'05. Proceedings of the 5th ACM/IEEE-CS Joint Conference on*, 2005, pp. 141–142, IEEE.

[60] H. Wang, F. Zhang, X. Xie, M. Guo, "Dkn: Deep knowledge-aware network for news recommendation," in *Proceedings of the 2018 world wide web conference*, 2018, pp. 1835–1844.

[61] A. Shahmohammadi, E. Khadangi, A. Bagheri, "Presenting new collaborative link prediction methods for activity recommendation in facebook," *Neurocomputing*, vol. 210, pp. 217–226, 2016.

[62] T. Zhou, J. Ren, M. Medo, Y.-C. Zhang, "Bipartite network projection and personal recommendation," *Physical Review E*, vol. 76, no. 4, p. 046115, 2007, doi: 10.1103/PhysRevE.76.046115.

[63] I. Tiddi, *et al.*, "Foundations of explainable knowledge- enabled systems," *Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges*, vol. 47, p. 23, 2020.

[64] F. Lecue, "On the role of knowledge graphs in explainable ai," *Semantic Web*, no. Preprint, pp. 1–11, 2019.

[65] N. Barbieri, F. Bonchi, G. Manco, "Who to follow and why: link prediction with explanations," in *20th ACM SIGKDD International Conference on Knowledge discovery and data mining*, 2014, pp. 1266–1275, ACM.

[66] Y. Xian, Z. Fu, S. Muthukrishnan, G. De Melo, Y. Zhang, "Reinforcement knowledge graph reasoning for explainable recommendation," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 285–294.

[67] X. Wang, D. Wang, C. Xu, X. He, Y. Cao, T.-S. Chua, "Explainable reasoning over knowledge graphs for recommendation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 5329–5336.

[68] M. Caro-Martinez, G. Jimenez-Diaz, "Similar users or similar items? comparing similarity-based approaches for recommender systems in online judges," in *International Conference on Case-Based Reasoning*, 2017, pp. 92–107, Springer.

[69] G. Jimenez-Diaz, P. P. Gómez-Martín, M. A. Gómez- Martín, A. A. Sánchez-Ruiz, "Similarity metrics from social network analysis for content recommender systems," *AI Communications*, vol. 30, no. 3-4, pp. 223– 234, 2017.

[70] G. Jimenez-Diaz, P. P. G. Martín, M. A. G. Martín, A. A. Sánchez-Ruiz, "Similarity metrics from social network analysis for content recommender systems," in *International Conference on Case-Based Reasoning*, 2016, pp. 203–217, Springer.

[71] S. Dooms, A. Bellogín, T. D. Pessemier, L. Martens, "A framework for dataset benchmarking and its application to a new movie rating dataset," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 7, no. 3, p. 41, 2016.

Marta Caro-Martínez

Marta Caro-Martínez is a Ph.D. Student at the Complutense University of Madrid. She got a degree in Computer Science in 2015 and a Master's degree in the same field in 2017. Her research work focuses on recommender systems and Explainable Artificial Intelligence. She uses graphs and Social Network Analysis techniques to implement recommendations and explanations. She has also researched interactive visualization and augmented reality applications for museums. Her publications have appeared at several International Conferences.

Guillermo Jiménez Díaz

Guillermo Jiménez Díaz is Computer Research Scientist and Associate Professor at Universidad Complutense Madrid. He received his Ph.D. Universidad Complutense Madrid in Computer Science. His Thesis focused on using virtual environments and active-based learning to teach Object-Oriented Programming. His research is concerned to recommender systems and its combination with techniques from social network analysis. Our research is applied in two different domains: tourism and e-learning. He is also interested in the application of augmented reality technologies in Museums.

Juan A. Recio García

Juan A. Recio García is Associate Professor at the Department of Software Engineering and Artificial Intelligence at the Computer Science Faculty at the Complutense University of Madrid, where he held the position of Head of Department from 2015 to 2019. Currently he holds the BOSCH-UCM Honorary Chair on Artificial Intelligence applied to Internet of Things and is Board Member of the IAAA Student Chapter in Mexico for the promotion of Artificial Intelligence. He is lead investigator of several national-founded projects and he has conducted several contracts with companies in the area of Artificial Intelligence. His research has focused on the confluence of Software Engineering and Case-Based Reasoning, developing the COLIBRI platform for building CBR systems. He has also worked in the areas of Context-aware and social Recommender Systems. Currently his research is focused on eXplainable Artificial Intelligence (XAI).

OntoInfoG++: A Knowledge Fusion Semantic Approach for Infographics Recommendation

Gerard Deepak*, Adithya Vibakar, A. Santhanavijayan

Department of CSE, National Institute of Technology, Tiruchirappalli (India)

Received 8 April 2021 | Accepted 6 October 2021 | Published 22 December 2021

Abstract

As humans tend to improvise and learn on a constant basis, the need for visualizing and recommending knowledge is increasing. Since the World Wide Web is exploded with a lot of multimedia content and with a growing amount of research papers on the Web, there is a potential need for inferential multimedia like the infographics which can lead to an ultimate new level of learning from most viable information sources on the Web. The potential growth and future of technology have called for the need of a Web 3.0 compliant infographic recommendation system in order to be able to visualize, design and develop aesthetically. The trend of the Web has asked for better infographic recommendations in the attempt of technological exploration. This paper proposes the OntoInfoG++ which is a knowledge centric recommendation approach for Infographics that encompasses the amalgamation of metadata derived from multiple heterogenous sources and the crowd sourced ontologies to recommend infographics based on the topic of interest of the user. The user- clicks are taken into consideration along with an Ontology which is modeled using the titles and the keywords extracted from the dataset comprising of research papers. The approach models user topic of interest from the Query Words, Current User-Clicks, and from standard Knowledge Stores like the BibSonomy, DBpedia, Wikidata, LOD Cloud, and crowd sourced Ontologies. The semantic alignment is achieved using three distinct measures namely the Horn's index, EnAPMI measure and information entropy. The resultant infographic recommendation has been achieved by computing the semantic similarity between enriched topics of interest and infographic labels and arrange the recommended infographics in the increasing order of their semantic similarity to yield a chronological order for the meaningful arrangement of infographics. The OntoInfoG++ has achieved an overall F-measure of 97.27 % which is the best-in-class F-measure for an infographic recommendation system.

Keywords

Horn's Overlap Index, Infographics Recommendation, Knowledge Centric, Ontologies.

DOI: 10.9781/ijimai.2021.12.005

I. Introduction

In this contemporary Business world where people are running for money, time management has become a very important issue for everyone. There is lot of important information that people come across daily from posters on the road to product description on e-commerce websites. Remembering and processing extensive information is not only time consuming but requires techniques which are computationally expensive. To address this issue, the mode of communication or knowledge transfer has to be reformed where the amount of information communicated must be more with respect to the time consumed and the information must provide deep insights and must be easy to remember. There is so much information around us that it is impossible to understand and recall anything in a short span of time. However, using infographics to show data and statistics in the form of graphs, pictures, images, bulletins, etc., can help individuals absorb data in a much more efficient way.

Infographics are graphic portrayals of data, facts, information, or knowledge designed to easily and clearly display complex information. By incorporating graphic images, the perception of designs and patterns in the human visual system can be improved to an extent such that Infographics tell a tale as they help in organizing details and make it visually appealing and catchy, such that the audiences and other users can process, analyze, and interpret information quickly. Infographics display vast quantities of data and knowledge in the form of an information graph, flowchart or an image. They are used for many reasons such as they are fun to use and make learning enjoyable, eye-catching, succinct, and all the details they contain are easily absorbed by the reader, which makes them beneficial.

While the Web 2.0 is currently in use, a lot of additional technologies is continuously being added, which paves a way to the Web of data on the semantic standards of the Web continuous research, also referred to as Web 3.0. Web 2.0 was powered by Social Networks and cloud services while Web 3.0 is primarily based on newly developed technologies like Open Data Networks (ODN) and Semantic Intelligence. While Web 2.0 was powered by the emergence of smartphone, social and cloud services, Web 3.0 is based on three new levels of technical innovation: Ontology focused Computing, Knowledge based Computing and Semantic Inference. Semantic Web is an advancement of the existing Web which comprises of organized layers into a framework which are further modeled into Open Linked Data. The word Semantic means "processable information", and it is mandatory for the Semantic Web

* Corresponding author.
E-mail address: gerard.deepak.cse.nitt@gmail.com

to have a vocabulary with which both data and rules are articulated for Data Justification, which permits the export of entities for Knowledge Representation and Reasoning by information systems.

A unified data is always considered to be the uppermost priority for representation. A data or certain information which is not unified because of certain traits or features can cause clogging of thoughts while representing it. It is easy to understand and drive flexible solutions from a coalesced form of knowledge from heterogeneous sources. The key factor of a semantic paradigm is how the information is represented and reused. In the era of Web 3.0 where all entities are labelled, it is necessary to have some grounds where information is derived as useful auxiliary knowledge to be processed by information systems. In this era where there is limitless data, representation should be such that it allows machine to process the available information genuinely and provide accurate answers based on the queries imposed. Trending technological affairs drives towards a new approach of building solutions and developing Ontologies to represent information on the Semantic Web. The need for processable definitions and terms has now become a great requirement so that information extracted can be used and be manipulated according to the needs of the user and requirement of information systems. Knowledge Representation is a sub-basket of Artificial Intelligence dealing with interpreting, developing, and applying ways of expressing information on a machine such that programs can use this information for various purposes.

The semantically driven knowledge centric approach is entirely based on the paradigm of inferencing based on semantic similarity measures and diversity indices. Since the data on the World Wide Web is exponentially growing on everyday basis, it is almost impossible to learn from the contents on the World Wide Web. However, the Learning Based Approaches such as Machine Learning and Deep Learning Strategies, learn only from a sub-set of data, namely the dataset used in the approach. This transforms the problem as a closed domain problem, and specifically in Machine Learning, there is a need to choose or devise a feature selection approach which should perform well. However, in the Deep Learning paradigm, the entire approach is a Black Box where step by step computations are not visible. Moreover, owing to the large amount of Linked Open Data on the Web 3.0, it is highly complex to train a Learning Algorithm by accounting and preserving the Links between the entities. The semantic-based approach transforms the entire problem into an inferential scheme which is suitable for highly linked cohesive environment with a high data density like the Semantic Web.

OntoInfoG++ paves a way for eXplainable AI as it is based on semantic intelligence driven inferential paradigm. Learning Based Schemes do not promote eXplainable AI as eXplainable AI deals with breaking the Black Box involved in Machine Learning and Deep Learning Algorithms. Machine Learning Approaches however facilitate manual feature selection which can be configured separately in the algorithm and the strength of the algorithm can be improved. However, in Deep Learning even the feature selection is auto-handcrafted, and the computations for a specific set of data is a completely Black Box. eXplainable AI deals with solving a specific problem by formulating algorithms in which step by step computability can be reasoned out by human minds. OntoInfoG++ does not encompass Machine Learning or Deep Learning Algorithms, rather it makes use of Semantic Intelligence Driven Reasoning Schemes and transforms the problem into an inferential open domain problem. The entire computations which happen in the proposed approach is seen as a white-box and ensures human minds to reason out thus supporting eXplainable AI.

A. Motivation

Research on infographics over the past decade has been mainly focused on the role of the use graphical representation as an attention-grabbing strategy. Infographics have been a helpful tool in many areas of domain from education to advertisements. Nowadays for specific topics search engine yields a lot of infographics for the input topics, but how much of this is relevant? In order to solve this issue, a proper recommendation model has to be devised to furnish relevant content. Though infographics exists as images, traditional image processing or visual similarity-based recommender techniques cannot be used, as the queries are in the form of text and an annotation-based approach is the need of the hour. This can be achieved using a Semantically Driven approach, which would arrange the infographics relevant to that topic of user interest in a chronological order. The Semantically driven approach is responsible for giving practical and logical representations that can give more sensible solutions when compared to the conventional recommender systems that use only basic feature extraction techniques, where the real-world knowledge will not be taken care of. The traditional recommendation systems give solutions that are non-practical and may not generalize easily to all the topics of the same problem as learning the huge volumes of data from the web is infeasible. In a semantically driven approach, as the real-world knowledge is integrated from several heterogeneous sources, entities will be populated such that there will be more context terms with a high information density that will be added. The World Wide Web houses several Knowledge Sources which when incorporated increase the density of the knowledge, and thereby facilitate enriching the supplementary knowledge such that the synonymy, polysemy, cold-start, serendipity, context irrelevance, and cross domain data sparsity problems can be solved.

B. Contribution

A semantic approach for an Annotations Based Infographic recommendation has been proposed in this paper. OntoInfoG++ is a Knowledge-centric approach which uses real-world knowledge from various heterogenous sources. The OntoInfoG++ uses both user query and user clicks which are preprocessed and are formulated as a query word set and are collated to form a user initial set of user topic of interest. A Knowledge Graph is formulated by subjecting the query word set to topic enrichment using BibSonomy, DBpedia, Wikidata, LOD Cloud, and Crowd Sourced Ontologies. The titles and keywords extracted from the dataset are utilized to formulate Ontologies which facilitates in semantic concept alignment with the formulated knowledge graph, to yield the enriched topic of interest knowledge graph. The semantic similarity is computed with the help of EnAPMI, Horn's Index and information entropy between the enriched topic of interest knowledge graph and keywords extracted from infographics extracted from research papers. A story of infographics is created and it is recommended based on the scores obtained from the semantic similarity measures.

C. Organization

This paper is organized as follows. Section II addresses the relevant research work done related to this area. Section III depicts the Problem Definition and Assumptions. The proposed methodology is represented in Section IV. Section V composes the implementation details. The performance evaluation and results are depicted in Section VI. The paper is concluded in Section VII.

II. Related Work

Siricharoen et al. [1] have briefly explained how infographics were used in journalism, and also how infographics serves as better mode for effective communication in the digital age and have put forth the history, significance, and benefits of infographics and tools for making infographics more beneficial and effective. They have also addressed the suggestive guidelines of infographics creation. Sujia

Zhu et al. [2] have reviewed and classified automatic tools that cater to visual recommendations for visualizing storytelling, visualizations of graphs, visualization of annotations, and visualization of information networks in several varied perspectives. They have also posed many obstacles and directions for potential work in the field of automated infographics and visual recommendations. Wilkinson et al. [3] have performed content analysis obtained evaluating Diet-Related Infographics and have used it for Behavior Change Theories. This was implemented by pin creation that makes use of both pictures and textual descriptions to portray elements that convey information about nutrition.

Featherstone [4] has proposed how infographics is used as a primer and supported it with visual data and briefly explained about the tools used for the same. Mohd Noh et al. [5] have discussed how Infographics is implemented as a training tool to assist teachers in education and learning sessions to allow student and teachers understand and interpret concepts with ease. Siricharoen et al. [6] have addressed the critical aspects assessment approach for infographics, which is discussed briefly with questions. Murray et al. [7] have studied and discussed about some basic concepts for the design of effective infographics and have proposed some suggestions for the development of engaging infographics. Cifci et al. [8] have studied how Infographics affect students' achievements. The analysis is very significant which leads to designing instructional materials which can be used in classrooms. They have implemented the research as a quasi-experimental study, one of the quantitative methods of study. They have also inferred that the use of infographics in geography lesson improved academic performance.

Nuhoglu et al. [9] have researched how infographics can be generated as a scheme for visualizing content for interactive learning of scenarios by incorporating a design for infographics which caters to a collaborative technology based Bridge21 learning model which has been proposed to foster learning. Chen et al. [10] have proposed a deep learning driven strategy that retrieves a timeline template from the images which are quite magnanimous. They have adopted a deconstruction and reconstruction technique. Cui et al. [11] have put forth a strategic infographic generation technique and have built a system that synthesizes statements relevant to statistics to a potential infographic which is obtained from previous studies. However, there is an emphasis on the aspect of modeling infographics for statistics as a potential domain. Mackinlay et al. [12] presented an approach for designing graphical presentation on the basis of the perspective, that graphical representations are outcome based phrases resultant from graphics-based languages. They also introduced a model of a presentation method called APT with AI techniques focused on algebra and graphic design requirements.

Deepak et al. [13] have developed a Web 2.0 complaint RDF driven model that focuses on decreasing the irrelevance and promoting diversity in the results from semantic search. Indicator terms were yielded by computing dyadic RDF entities from a set of webpages for which an RDF polarization vector is derived from the inferencing of the modeled term-frequency matrix and term co-occurrence matrix. Middleton et al. [14] have built ontological models for profiling recommendation, namely, Quickstep and foxtrot. An ontological interference model has been employed for the improvement of the performance and also encompasses external ontological entities for achieving profile base bootstrapping. Furthermore, the visualization of user profiles to yield relevant feedback has been proposed. David Werner et al. [15] have designed an ontology based multi-layer recommendation system for economic articles based on a client's profile, that produces a magazine per customer composed of a set of daily produced articles. The main aim of the developed system was to reduce the overload of useless issues.

Peis et al. [16] have focused on reviewing semantic recommender systems based on classification criteria, ontological and conceptual diagrams, which have been proved to be effective for research and experimentation. Pazahr et al. [17] have designed an advertisement recommender system that has been semantically enhanced, and at the same time produces recommendations in a simplified manner. The proposed architecture uses semantic logic to showcase the recommended products and this in turn can differentiate between the recommender unit from the classical recommender methods. Prafulla et al. [18] have proposed an approach that makes uses of semantic clustering for task recommender systems to identify right personnel. The method utilizes a feature extraction scheme which is based on generation of synsets and a strategy of semantic clustering which is iterative is nature. The approach also cognitively maps synonyms which helps in yielding a better performance. The approach also solved the issue of scalability with reduced entropy. Huijsduijnen et al. [19] have designed a model, Bing-CSF-IDF+, a content-based RS for news which is semantically driven. They have compared the performance with a previously designed version, Bing-SF-IDF+, and found the former to outperform the latter by statistical means like F-measure and kappa. The approach uses concepts and relationships from domain ontology, synsets and synset relationship from semantic lexicon from WordNet. Tymchenko et al. [20] have proposed a multifactor selection scheme for the design of infographics. This is a hierarchical approach that uses a pair comparison model for the evaluation of higher levels of interconnection elements using factor comparisons between a pair of elements for infographics design.

Ontologies have played a pivotal role in improving the recommendations in web search systems. Owing to the reason that Ontologies provide a good amount of supplementary knowledge in improving the context and the scope of the query, they can be employed in recommending infographics. Ontologies have significantly increased the recommendation relevance in [21], [22], [23], [24], [25] and have solved the problem of polysemy, ambiguity, and context irrelevance. However, Ontologies, once modeled, need to be constantly monitored for quality and must be updated. Also, there can arise a scenario where the Ontology would not be able to dispatch the auxiliary knowledge for a specific query. In such cases, dynamically modeled metadata or dynamically extracted relevant entities to the query need to be supplied which is addressed in this paper by infusing entities from a wide array of knowledge sources and building a knowledge graph from the user query and the knowledge sources.

Berkani et al. [26] have proposed a semantically driven approach that depends on social representation of user profiles for recommendations of user's profiles and have also employed two categorization strategies in order to optimise the performance of the recommendations: using the K-mean algorithm (originally utilised for everyone) and K-Nearest Neighbors method (applied to newly added users). Javed et al. [27] have proposed a context-aware recommender system for filtering things related to the user interest, as well as a context-based recommender system for recommending those things. Their context-based recommender system extracts patterns from the World Wide Web based on the user's previous interactions and delivers recommendations for future news. Houari et al. [28] have proposed a domain specific tool for recommending experts using N. S., & PROMETHEE II and Negotiation in support of industrial maintenance.

Bobadilla et al. [29] have incorporated Deep Learning schemes for enhancing the quality of recommendations and in Recommender Systems. However, there is a need to arrive at techniques that either eliminate learning paradigms or break the Blackbox in machine learning and deep learning strategies in order to provision eXplainable AI.

Analysis of the Literature clearly points out that the infographics are quite useful in rendering knowledge and there is a need for

an infographics recommendation system. However, the existing infographics recommendations are either based on simple ontologies or semantic logics or even focus on very few parameters like the user-clicks and the user query or infuse learning algorithms from the Web. They either make the system complex when learning algorithms are infused or depict a lacuna as enough knowledge is not infused into the system. When clustering alone is the focus, then the approach results in high amounts of error rates. However, this can be solved when the right semantic similarity techniques are infused via semantic agents, and numerous entities are dynamically infused into the system.

III. Problem Definition and Assumptions

A. Problem Definition

Given a query which specifies at least one real-world entity, a dataset comprising research papers, an access to several real-world knowledge sources, and the user-click information, the first objective is to model the user topics of interest from the query words, user-click information and the entities from several real-world knowledge sources. The second objective is to model Ontologies using titles and keywords extracted from research papers from the dataset. The third objective is to achieve semantic concept alignment between the formulated topic of interest knowledge graph and the initially formulated Ontologies. The final objective is to furnish the infographics by computing the semantic similarity and arranging them in a chronological order in the increasing order of the semantic similarity and recommend to the user, until there are no further user-clicks recorded.

B. Assumptions

The Ontologies modeled from the keywords and titles from the research papers must be strictly adherent to the papers present in the dataset and must be free of inconsistencies. The modeled Ontologies must be at least a strong representation of the upper-level ontologies. The user-clicks used in the approach must also be adherent to the domains in the dataset. The Queries must have at least one strong entity and must not exceed to more than 12 strong entities. It is a mandatory requirement that the dataset be categorical in nature and the Infographics must be strictly labelled or annotated.

IV. Proposed Methodology

The architecture of the proposed Semantically Driven Infographics Recommendation System, the OntoInfoG++, is depicted in Fig. 1 which is knowledge centric and constitutes a large amount of real-world knowledge from varied heterogenous sources. The OntoInfoG++ is also driven by the user query and imbibes the current user-clicks into the system. The user query is subject to initial pre-processing which constitutes of the Tokenization, Lemmatization, Stop Word Removal, and elimination of special characters. The preprocessed user queries are formulated as a query word set which is collated along with the current user clicks of the user to formulate an initial set of User-Topic of Interest.

The initial set of User-Topic of Interest is subject to topic enrichment by aggregation of auxiliary domain knowledge from real-world data stores, namely the BibSonomy, DBPedia, Wikidata, LOD Cloud, and Crowd Sourced Ontologies. Since the infographics recommendation comprises of recommending knowledge containing graphics and diagrams, there is a need for integration of knowledge from BibSonomy. The DBPedia and Wikidata further remove anomalies and integrate detailed auxiliary knowledge into the existing entities and facilitate linking of newer entities. The reason for including Crowd centric domain ontologies is to further enhance the density of information linked to the exiting entities and to yield a humanized perspective to the user query. The inclusion of domain centric ontologies helps in adding a broader human centric perspective to the query and also enhances the density of knowledge which enhances the diversity of results without much deviation from the user-interests and the query topic.

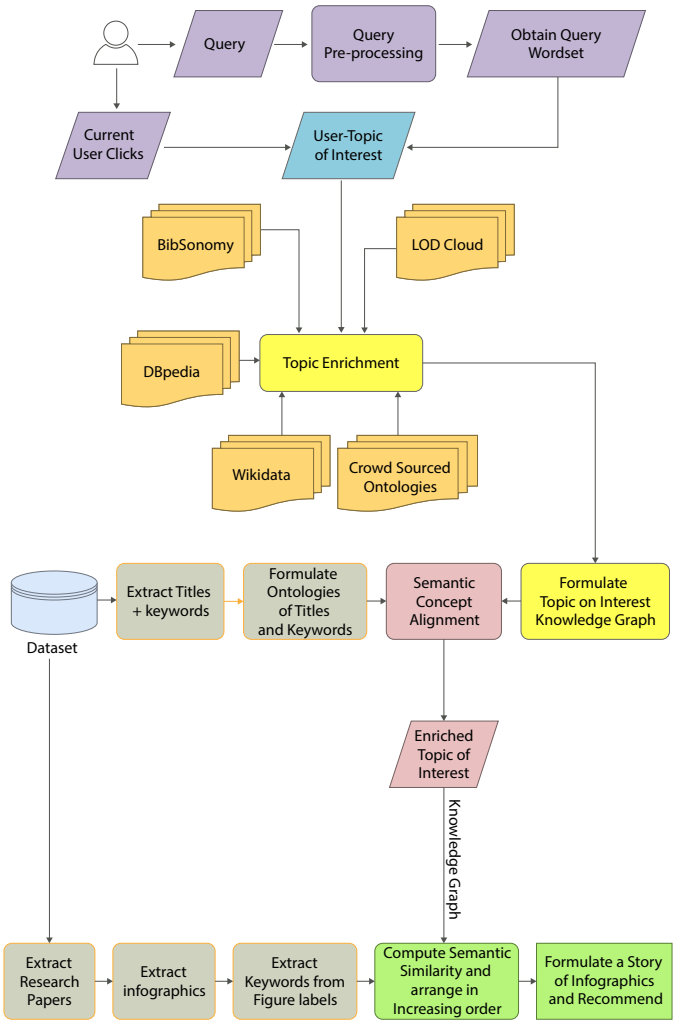


Fig. 1. Proposed System Architecture.

The Enriched Topic of Interests which is inclusive of cognitive real-world knowledge from several knowledge stores is modeled into Knowledge Graphs which is undirected and at least a single link is established between its constituent entities. The knowledge Graph simulates query relevant metadata which has been loaded from the Web 3.0. The User Topic of Interest Knowledge Graph serves as a liaison between the dataset and the user and integrates the categories in the dataset. OntoInfoG++ is a knowledge centric annotations-based infographics recommendation system where annotations play a vital role in formulating a sequential story comprising of the infographics and thereby recommend a collection of similar infographics which are relevant to the query and is suitable for satisfying the needs of the user. The titles and the keywords from the research papers are extracted from the dataset and an Ontology is formulated by creating links among the keywords in the research papers and meaningful words in the title of the paper.

A strategic Semantic Alignment of Concepts from the Ontology constituted from the Title and Keywords with that of the User Topic of Interest is carried out to further Enrich the amount of Knowledge and the Ontology- Knowledge Graph pair is associated with each

other by creating links between its concepts formulating a larger Knowledge Graph constituting Topic of Interest. Further, the research papers from the dataset are loaded, parsed, and figures are extracted based on an Agent which recognizes a specific figure is an infographic or not. There is also a term identification agent which eliminates performance graphs and keeps the infographics intact. System Architectures or diagrams describing a specific scenario or schematic or block diagrams are considered as infographics. Language Parsing Rules and matching with a set of terms is incorporated in the agent for extraction of infographics. The Semantic Similarity is computed between the Enriched Topic of Interest Knowledge Graph and the terms extracted from figures in the research paper. This is continued for all the infographics extracted. The terms are extracted in the increasing order of the Semantic Similarity and the infographics are also rearranged in the same order to formulate a story and arrange the infographics in a meaningful manner.

$$H_0 = \frac{\Sigma(p_{ij}+p_{ik})\log(p_{ij}+p_{ik}) - \Sigma p_{ij} log p_{ij} - \Sigma p_{ik} log p_{ik}}{2log2} \quad (1)$$

The Horn Index Ho as depicted in Eq. (1) is used to compute the Semantic Similarity or the Overlap of entities by computing the Ho between the instances in the Knowledge Graph and the keywords from the figure labels. When adapted into a semantic environment comprising of an Open Linked Data as knowledge graphs, the pij refers to the semantic similarity of terms i & j and pik refers to the semantic similarity between the terms i & k, such that 'i' corresponds to a term in label set of the infographic images while 'j' & 'k' correspond to the terms that are having a single link in the knowledge graph, which signifies they are adjacent to each other. Traditionally the Horn's Index used the proportion resource utilized by species of a type but when it is adapted in an Environment of the semantically driven information systems, it is substituted by the Semantic Similarity without having to modify the exact index. It must also be noted that any base of the logarithm can be used but it also must be ensured that a uniform logarithm base must be maintained throughout. The Horn's Index furnishes a value between 0 and 1, and the threshold is assumed as 0.5 for the Horn's Index for accepting and rejecting entities for recommendation.

The Semantic Similarity within Horn's Index is computed using the Normalized Pointwise Mutual Information (NPMI) measure by considering only the magnitude of NPMI measure which lies between 0 and 1. If the NPMI furnishes a negative value, only the magnitude is considered while the negative number is ignored. The reason for choosing NPMI over EnAPMI within the Horn's Index is due to the fact that NPMI is computationally less expensive than the EnAPMI. Moreover, since the entities after being subjected to NPMI are further being passed into the Horn's Index, as a result the less stringent NPMI would be sufficient at this case. However, the threshold for NPMI is empirically considered as 0.75 to allow only the entities that are highly relevant through the NPMI measure into the Horn's Index. Furthermore, this is also the reason to keep the threshold of the Horn's Index to 0.5 as NPMI is much more stringent and Horn's index need not be as stringent as already relevant entities are passed into it for a further approval. Eq. (3) depicts the NPMI measure which is based on the Pointwise Mutual Information score (PMI). PMI is a knowledge analysis and statistics indicator used to measure the relationship between terms. The PMI score is based on the Eq. (2). The PMI Score is normalized such that the values occur between [-1, +1], resulting in -1 for uncooperative incidents, 0 for isolated events and +1 for co-occurrence and the equation of NPMI is as shown in the Eq. (3).

$$\text{PMI (X, Y)} = log \frac{P(x,y)}{P(x)P(y)} \quad (2)$$

$$\text{NPMI} = \frac{PMI}{\log P(x,y)} \quad (3)$$

$$\text{EnAPMI(m,n)} = \frac{Pmi(n,m)}{p(m)(n)} + y - \eta \quad (4)$$

$$y = \frac{1+\log[p(m,n)]}{p(n)\log[p(m)] - p(m)\log[p(n)]} \quad (5)$$

$$\eta = \frac{\log[p(m).P(n)]}{\log(p(m,n))} \quad (6)$$

$$\text{H(X)} = \sum_{i=1}^{n} P(x_i)\log P(x_i) \quad (7)$$

EnAPMI is an Enriched Adaptive Pointwise Mutual Information measure (EnAPMI), a novel semantic similarity, which enriches the Adaptive Pointwise Mutual Information (APMI) which is a model based on PMI for the measurement of semantic similarities based on the likelihood of the event and terms were suggested to be co-occurring with an adaptive coefficient. The EnAPMI is as shown in the Eq. (4) between a pair of terms m and n. The EnAPMI measure which belongs to a class of the variants of the PMI models and is associated with an adaptivity coefficient y and a drift indicator η which has been employed to estimate the semantic drift between a pair of terms. The EnAPMI measure is derived from the APMI measure by eliminating the drift indicator from the APMI measure and adding the adaptivity co-efficient to the existing APMI model. Eq. (5) includes the adaptivity co-efficient of a pair of terms m and n, which is coupled with a logarithmic quotient of the probability of co-occurrence of the pair of terms m and n in its numerator.

The product of probability of occurrence of a term with the logarithm of the standalone probability of individual occurrence with its pair term is computed and their difference is included in the denominator of the adaptivity co-efficient. The reason for coupling the adaptive co-efficient with the variant of the PMI is primarily because the semantic relatedness between a term pair can be computed with greater efficacy when the word pair co-occurrence probability, the probability of word occurrence, and when the measure of self-information has been taken into consideration. The drift indicator η as depicted in Eq. (6) can be described as the ratio of the logarithm between the individual occurrence probability of the term pairs to that of the ratio of the probability co-occurrence ratio of the term pair. The semantic gap between a pair of terms is quantified and measured by the drift indicator which is computed between the pair of terms and is eliminated from the EnAPMI measure. The EnAPMI is derived from the APMI measure and the EnAPMI model acts as a better performing semantic similarity model between the pair of terms when compared to the other PMI based conventional models in a highly cohesive semantic environment. Eq. (7) depicts the information entropy H(X) which depicts the average quantity of information, inherent at the interval of the potential outcomes of the variable. X is the discrete variable, with possible outcomes {x_1, x_2, ..., x_i} which occur with probability {{P(x_1), P(x_2)....., P(x_i)}} and this is formally defined as information entropy represented in Eq. (7). The Entropy depicted as H(X) in Eq. (7) is the product of the probability of occurrence of a term over a web corpus with that of the self-information in the term. The Information Entropy is also used as a standalone measure for computing the relevance of the entities as the degree of information associated with an individual term over a corpus, which serves as a potential indicator to estimate the extent up to which the presence of the term creates an impact in the specific corpus.

V. Implementation

The implementation was carried out using JAVA as the language of choice with Eclipse as the preferred IDE. The reason for using JAVA is the ease of integration with AgentSpeak and JADE which were used to model the agent to compute the Entropy, Semantic Similarity, and the Horn's Index. The experimentation was conducted on the RARD

II: Related Article Recommendation Dataset which can be accessed from the Mr. DLib (http://mr-dlib.org). The RARD II dataset comprises of 94m recommendations which covers an item space of 24m unique items. The unique terms in the RARD II dataset were linked with google scholar to obtain the relevant research papers in full text mode and was stored in the linked repository. The research papers which were available in full text mode only were used for experimentation via google scholar and institutional repository for full text research papers. The reason for choosing the RARD II Article Recommendation Dataset is primarily for the only reason that terms based on research topics are available in the RARD II dataset, and the topic linked infographics can be extracted from the research papers in a sequence. A Language Processing Agent is also modeled for parsing the figures and ensuring that it is an infographic or, if it is a graph, that is based on the terms which are used to label the figures and the textual description of the figures. The state of the agent is described to extract the infographic images along with the image labels from the Research Articles and creates a categorical state space comprising of infographic images, the keywords in the labels, and other associated annotations. This enables the OntoInfoG++ to yield infographic images that are being queried by the user to satisfy the information needs of the user pertaining to the topic of interest of the user.

The Crowd Sourced Ontologies are generated by picking up terms from the RARD II dataset based on the keywords from the articles and from those in the labels of infographic images, and subject them to the OntoCollab [30] framework which facilitates dynamic generation of OWL and RDF Ontologies which have been hierarchically arranged, axiomatized, and reasoned out. Apart from the Domain Ontologies which have been generated using the OntoCollab, user modeled ontologies were also included. Web Protégé was used for manual modeling of domain specific ontologies. The Crowd Sourced ontologies were also collected from various online research communities and were curated into a meaningful Ontology using OntoCollab. Moreover, the index terms from Semantic Wikis were considered, these are Crowd Sourced at a large scale, extracted from user-blogs, several research articles, and from portals where user-skills and technology are emphasized. A major portion of Crowd Sourced Ontology is automatically generated by OntoCollab by facilitating access to these user-centric sources. The auxiliary knowledge is supplied into the OntoInfoG++ framework from BibSonomy, Wikidata, DBpedia, LOD Cloud, and Crowd Sourced Domain Ontologies. The reason for using an array of knowledge stores or factual knowledge bases is because diversified entities can be included to increase the density of query relevant knowledge. SPARQL Endpoints designed using AgentSpeak are integrated into the environment of OntoInfoG++ which queries Entities from several real-world knowledge stores like BibSonomy, Wikidata, DBpedia, and LOD Cloud. The reason for combining different knowledge sources is to increase the variety and heterogeneity of entities to provide auxiliary knowledge into the proposed paradigm. Moreover, the incorporation of knowledge from varied sources increases the diversity of results. The OntoInfoG++ integrates entities that are relevant to the user topics of interest and helps in topic enrichment and solves the Serendipity problem by entity integration from distinguished knowledge sources. OntoInfoG++ individually harvests topic relevant entities from various knowledge sources and Crowd Sourced Ontologies and further integrates together to facilitate topic enrichment and accelerate diversification of results.

The entities are supplied into the OntoInfoG++ Framework as collective knowledge for enrichment of the Query Terms and the User-Clicks. Among several domains which were used for experimentations, Table I documents 12 distinct and standard domains and the number of concepts and individuals in the domain. It is indicative from Table I that the number of individuals is much higher than the number of concepts. However, these concepts which are depicted in Table I comprise of the core concepts, specialized concepts, upper ontologies, and the sub-concepts which are hierarchically arranged. The individuals are the implementations of the specialized concepts which are used in experimentation.

TABLE I. Details of 12 Domains Along With the Number of Concepts and Individuals Used for Experimentations

Domains	No. of Concepts	No. of individuals
Agriculture	1745	4452
Horticulture	1245	3845
Library Science	1223	4986
Information	1435	3754
Economics	2135	3121
Sociology	845	1259
Humanities	1121	2456
Cloud & Distributed Computing	895	3856
Robotics	969	3254
Urban Planning and Sustainability	1921	4512
Life Sciences	1895	3695
Chemical Engineering	2032	4875

Moreover, the diversification of entities would result in diversified and yet relevant infographics increasing the spectrum of visibility of the diversified infographics under the purview of the topic without major deviation. In order to create a benchmark query-set for the RARD II dataset, the metadata was harvested. 124 users were given broad area topics from the RARD II dataset and were asked to use Google Scholar, other research paper search engines, and the standard knowledge stores, and were asked to formulate queries and also yield the ground truth infographic keywords of image labels for the formulated query. There were 2457 queries with the ground truth which were collected by the user participants over the period of 168 days. The manually modeled Ontology and the Dynamically generated Ontology were merged into a single Crowd Sourced Ontology and were used for experimentation. An end-to-end SPARQL agent was encompassed to obtain the metadata from the individual knowledge stores.

OntoInfoG++ is a knowledge centric paradigm for recommendation of infographics and is semantically compliant. The OntoInfoG++ strategy formalizes Topic of Interest knowledge graph which is a constituent of the query, the current user-click, and auxiliary knowledge from real-world knowledge bases. Also, the semantic alignment has been encompassed into the system using three distinct measures namely the EnAPMI measure, the Horn's index, and the Entropy. The reason for encompassing three distinct and yet effective measures is to increase the relevance. Moreover, EnAPMI is based on the probability of the occurrence and co-occurrence of words over the data corpus. The entropy computes the information measure, thereby the most informative entity in correlation of the environment in which it is contained is preferentially selected. The usage of an Agent for the infographics extraction and the ranking of the infographics based on the computation of the semantic similarity between the labels of the infographics and the enriched Topic of Interest ensures that the infographics are arranged in a chronological order and enables the users to deduce inference and also a small story gets created as soon as the infographics are arranged in a logical order. The OntoInfoG++ algorithm is depicted in Table II.

TABLE II. ALGORITHM I

Algorithm 1: Proposed OntoInfoG++ Algorithm for Infographics Recommendation

Input: Multi-word Query, Current User-Clicks, Access to Real World Knowledge Bases, Crowd Sourced Ontologies, Categorical Dataset S.

Output: Recommendation of Infographics in a chronological interpretable order

Begin

Step 1: The query Q input is subject to pre-processing constituting tokenization, lemmatization, and stop word removal to yield query word set Qs.

Step 2: The current user-clicks recorded based on the user navigation are also pre-processed and are merged with Qs to yield User Topics of Interest ToI.

Step 3: while (ToI.next()!=NULL)

Set Le ⟵ Load Entities from Real World Knowledge Stores like BibSonomy, DBpedia, Wikidata, LOD Cloud, and Crowd Sourced Ontologies.

end while

Step 4 : for each entity in Le

Formulate a Topic of Interest Knowledge Graph ToI_KG by computing the semantic similarity between Instance Pairs and Rearranging them By At Least Having a Single Link among each of the instances in the KG

end for

Step 5: Extract Keywords and Titles from S and formulate Ontologies of Titles and Keywords as OKnow.

Step 6 : Semantically Align the concepts in OKnow and ToI_KG using the SemantoSim, Horn's's Overlap Index, and Entropy to yield Enriched ToI_KG as EnToI_KG.

Step 7: Extract Research papers from the S, and thereby parse the Infographics using an infographic recognition agent and Load the Infographics and Infographic Keywords as InfoG_Keywords.

Step 8: for each entitiy in the EnToI_KG

Compute SemantoSim (EnToI_KG.currentEntity(), InfoG_Keywords())

if (SemantoSim.curr()>0.75)

HashMap RecInfoG ⟵ (InfoG_Keywords, SemantoSim Measure)

end

Step 9: Arrange RecInfoG in the increasing order of SemantoSim, and recommend the corresponding infographic to formulate a chronological order and tell a story relevant to the query words.

Step 10: Record the current user-clicks and formulate the ToI and continue Steps 2 to 9 until there are no further user-clicks recorded.

VI. RESULTS AND PERFORMANCE EVALUATION

The Performance of the proposed OntoInfoG++ was evaluated using the Precision, Recall, Accuracy, F-Measure, False Discovery Rate (FDR), and the Normalized Discounted Cumulative Gain (nDCG) as the potential metrics. Precision, Recall, Accuracy, and F-Measure compute the relevance of results to the query as well as the user-interests. The FDR indicates the number of false positives recommended by the system. The nDCG measures the diversity of recommendation of results to quantitatively indicate the degree of diversity in the recommended infographics. Precision, Recall, Accuracy, and F-Measure are indicated by Eq. (8), Eq. (9), Eq. (10), and Eq. (11) respectively. Eq. (12), Eq. (13), and Eq. (14) represent the FDR, nDCG, and the Discounted Cumulative Gain respectively.

$$\text{Precision}=\frac{\text{Retrieved}\cap\text{Relevant}}{\text{Retrieved}} \tag{8}$$

$$\text{Recall}=\frac{\text{Retrieved}\cap\text{Relevant}}{\text{Relevant}} \tag{9}$$

$$\text{Accuracy}=\frac{\text{Proportion Corrects qualifying ground truth test}}{\textit{Total No.of Queries}} \tag{10}$$

$$\text{F-Measure}=\frac{2(\text{Precision}\times\text{Recall})}{(\text{Precision}+\text{Recall})} \tag{11}$$

$$\text{False Discovery Rate}= 1-\text{Positive Predictive Value} \tag{12}$$

$$\text{nDCG} = \frac{\text{DCG}\alpha}{\text{IDCG}\alpha} \tag{13}$$

$$DCG = \sum_{i=1}^{\alpha}\frac{Rel_i}{log(i+1)} \tag{14}$$

From Fig. 2, it is easily inferable that OntoInfoG++ framework has yielded an overall Precision of 98.12%, an overall Recall of 96.4%, an overall Accuracy of 97.21%, and an overall F-Measure of 97.27 % . It can be easily interpreted from Fig. 3, that the proposed OntoInfoG++ framework furnishes a FDR of 0.02 with an nDCG of 0.95.

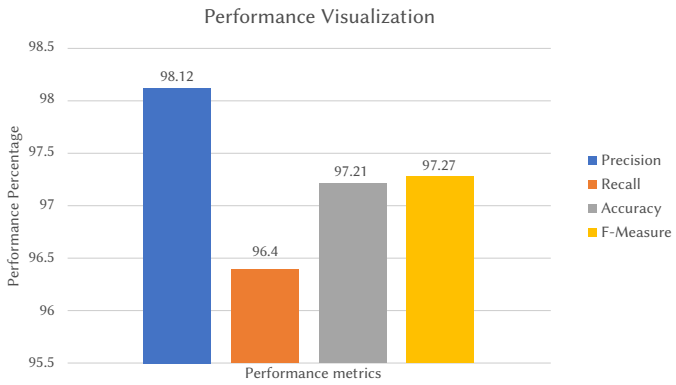


Fig. 2. Performance of the Proposed OntoInfoG++ for Infographics Recommendation.

The reason for the high values of Precision, Recall, Accuracy, and F-Measure is owing to the reason that the OntoInfoG++ is a knowledge centric paradigm for recommendation of infographics and is semantically compliant. The OntoInfoG++ strategy formalizes Topic of Interest knowledge graph which is a constituent of the query, the current user-click, and auxiliary knowledge from real-world knowledge bases. Moreover, OntoInfoG++, is a infographics recommendation approach that takes into consideration both the user-query and the current user-clicks of the user. It amalgamates topics of user interest from varied sources from which the entities are sourced. The sources of user interests include contents from BibSonomy, DBpedia, Wikidata, LOD Cloud, and Crowd Sourced Ontologies pertaining to varied domains which are being considered for experimentation. The domains are chosen in a way such that it is recurrent in the dataset. The approach also considers the elements from the dataset which include the titles from the dataset and the keywords. Furthermore, the approach specifically formulates the Ontologies inclusive of titles and keywords which are subject to semantic concept alignment from the knowledge base which was formulated initially from the topic of interest. The proposed OntoInfoG++ enhances the relevance of results predominantly considering the fact that it uses Horn's Index, EnAPMI measure to compute the semantic similarity, and the Information Entropy separately for computing the semantic relevance of results.

The use of a system of three different and yet comprehensive measures ensures that relevance of results is maintained and is non-deviated with respect to the user query, query-clicks, and the topic of interest of the user is adhered strictly.

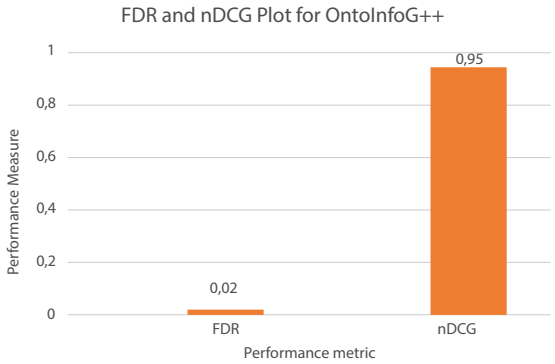


Fig.3. FDR and nDCG Plot.

The performance of the OntoInfoG++ is compared with the base line approaches as shown in Table III. Since infographics recommendation is quite new, there are not many baseline models available, except for the PCMSI. However, the other famous models in image recommendation were considered and implemented for the infographics and were taken as baseline models, namely the Collaborative Filtering, Fuzzy c-means clustering, and the CNN-K-Means Clustering, respectively. The PCMSI used pairwise comparisons and graphs for selection of infographics where the hierarchical evaluation of elements has been realized for recommendation of infographics. To evaluate the PCMSI in the proposed environment, the model has been used to recommend infographics in the exact same environment of the OntoInfoG++. PCMSI is quite fascinating in its approach, however the lack of metadata into the approach makes it linger in its performance when comparison to the OntoInfoG++. The CNN K-Means clustering when implemented in the environment of OntoInfoG++ makes it computationally expensive and learning the annotations makes it lag. The K-Means when coupled with CNN does not do wonders to the performance. The combination of Collaborative Filtering with Fuzzy c-Means clustering also could not perform well as it exhibited cold start problem and there was sparsity in the recommendation results.

The low value of the FDR is a clear indication that the proposed OntoInfoG++ performs well with a high degree of efficacy. The primary reason for the low FDR value is mainly due to the incorporation of auxiliary knowledge by aggregating entities from various knowledge bases and fact stores which are multi-faceted. Also, inclusion of entities from Wikidata, DBpedia, BibSonomy, LOD Cloud, and Crowd Sourced Domain Ontologies enhances the density of knowledge and thereby solves the serendipity problem in infographics recommendation. The inclusion of EnAPMI, Horn's Index, and Entropy together facilitates the integration of entities that are relevant in all aspects with respect to the user query and the user preferences in terms of query click. The reason why three different strategies with different perspectives are used is mainly for the reason to filter out and eliminate the false positives to a large extent such that diverse and yet highly relevant entities to the query and the user-preferences are to be retained. The Precision, Recall, Accuracy, F-Measure vs the number of recommendations are plotted in Fig. 4 (a), (b), (c), (d) respectively. From the Fig.4 (a) it is very clear that the precision of the proposed OntoInfoG++ is 10.44% higher that the PCMSI [23] when compared to the Collaborative Filtering with Fuzzy c-means clustering. The precision of OntoInfoG++ is 13.41% higher than that of Collaborative Filtering with Fuzzy c-means clustering. When compared with that of CNN-K-Means Clustering the precision is 89.43% higher. From Fig. 4 (b) it is inferable that the recall of the proposed OntoInfoG++ is 11.48%, 16.22%, 16.57 % higher than that of the PCMSI [23], Collaborative Filtering with Fuzzy c-means clustering, and CNN-K-Means clustering respectively. Seeing the plot in Fig. 4 (c) it is inferable that the accuracy of OntoInfoG++ is 10.93%, 14.83%, 15.98% higher than PCMSI [23], Collaborative Filtering with Fuzzy c-means clustering, and CNN-K-Means clustering respectively.

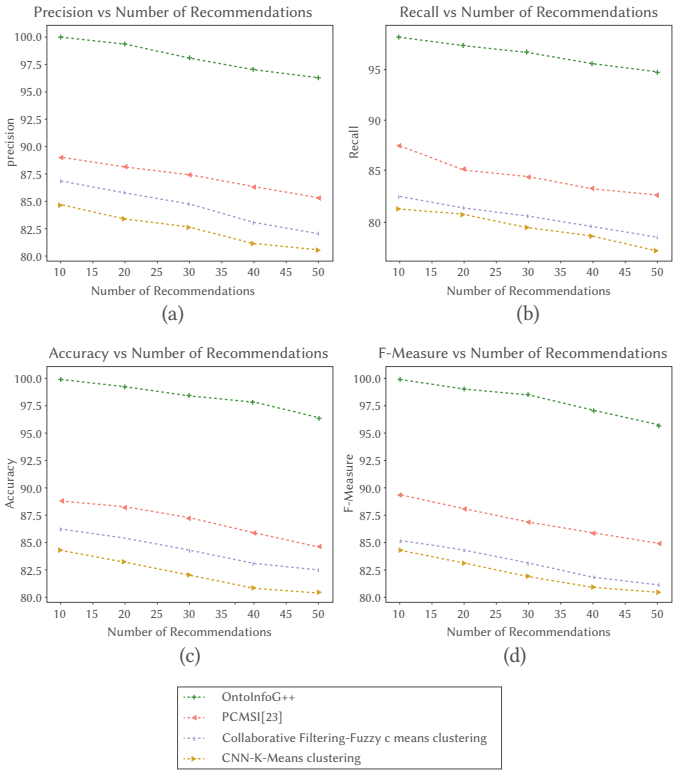


Fig.4. Performance metrics vs number of recommendations.

From the Fig. 4 (d) it is inferable that the F-Measure of the proposed OntoInfoG++ is 10.99%, 14.89%, 16.04% higher than that of the PCMSI [23], Collaborative Filtering with Fuzzy c-means Clustering, and CNN-K-Means clustering, respectively. As the number of recommendations increases, the curve for precision, recall, accuracy and F-Measure vs Number of recommendations also tend to decrease naturally. The reason why it decreases is because, as the number of recommendations increases, the irrelevance in the recommendations increases. However, the relative performance of the proposed OntoInfoG++ is higher when compared to the baseline models and benchmark approaches irrespective of the number of recommendations.

TABLE III. Comparison of Performance of the Proposed OntoInfoG++ With Other Approaches

Search Technique	Average Precision %	Average Recall %	Accuracy %	F-Measure	FDR
PCMSI [23]	87.68	84.92	85.87	86.28	0.13
Collaborative Filtering-Fuzzy c means clustering	84.71	80.18	82.45	82.38	0.16
CNN-K-Means Clustering	82.69	79.83	80.46	81.23	0.18
OntoInfoG++	98.12	96.4	97.21	97.27	0.02

The reason for this superiority of performance of the proposed approach is because the PCMSI [23] is a graphical approach that uses spectrum based comparison, which requires a proper graph to be modeled and pairwise relations need to be computed. The deviations occurring while calculating the pairwise relations results in a fair amount of un-correlated associations, and in Collaborative Filtering with Fuzzy c-means clustering approach there will be cold start problem and data sparsity problem and also, as it is using Fuzzy c-means Clustering with it, it makes the approach more computationally complex and, as the ratings can definitely differ, it is not a feasible technique. In the CNN-K-Means clustering, CNN is a learning algorithm which takes in hand crafted features, thereby increases the learning load of the model. All these three approaches do not use any form of real-World Knowledge sources to learn entities from the World Wide Web and there is a very low amount of data and very low amount of information and therefore all the three approaches lack in diversity. The proposed OntoInfoG++ is a lightweight inferential paradigm as there is dynamic computation of semantic relatedness using software agents, which make it quite efficient.

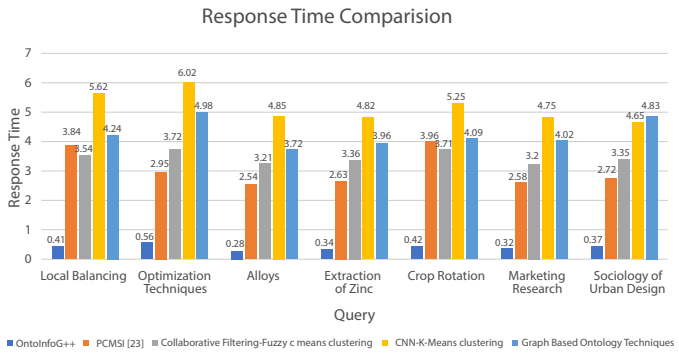


Fig. 5. Response Time Comparison of the Proposed OntoInfoG++ for Infographics Recommendation.

The comparison of response time of the OntoInfoG++ framework for a set of 7 distinct queries for recommendation of infographics is depicted in Fig. 5. Although, the overall evaluation of average response time computation is done for all the 2457 queries, the comparisons are tabulated only for 7 queries each of them from a distinct domain, by including the minimum and the maximum response time in the table. It is clear that OntoInfoG++ has a range of response time between 0.32 ms and 0.56 ms, while the PCMSI has recoded the response time between the range of 2.54 ms and 0.39 ms. However, the Collaborative Filtering with Fuzzy c-means clustering has recorded a response time in the range between 3.2 ms and 3.54 ms. The CNN with K-means clustering has recorded a response time in the range between 4.65 ms and 6.02 ms. The Graph Based Ontology Technique has recorded a response time in the range between 3.72 ms and 4.98 ms. It is quite evident and clear that from the tabulations of response time yielded by several approaches, the OntoInfoG++ has recorded the lowest average response time of 0.38 ms for 2457 queries. The PCMSI, Collaborative Filtering with Fuzzy c- means clustering has recorded an average response time of 3.05ms and 3.45 ms respectively for 2457 queries. The CNN with K-Means Clustering and Graph Based Ontology Techniques have recorded an average response time of 5.14 ms and 4.26 ms respectively for 2457 queries.

The reason why OntoInfoG++ has the lowest application response time is mainly due to the fact that it is an Agent Centered Approach and does not use a learning algorithm or a learning scheme for recommendation. Instead, the OntoInfoG++ is built on an inferential mechanism which uses three tactical approaches namely the EnAPMI Semantic Similarity Measure, the Horn's Index, and the Information Entropy. The incorporation of three distinct strategies for computing the most relevant items to the user-query ensures that the recommendation items are most relevant to the query without any deviations. The response time of OntoInfoG++ is mainly because of the Agents which are modeled for computing the EnAPMI, Horn's Index, and Information Entropy at a single step. However, the PCMSI is a graphical model where spectrum-based comparison is done by computing the pairwise relations each time the relevance between the elements in the graph has to be computed. Moreover, the PCMSI does not follow an inferential paradigm and the absence of specialized agents increases the response time of the model.

In case of Collaborative Filtering with Fuzzy c-means clustering, the two techniques are carried out in series one after the other and there is no parallel processing involved. Moreover, Collaborative Filtering is based on User-Item Matrix and the Ratings which result in complex computation and cold start problem which increases the processing time of the framework. CNN with K-Means clustering has the highest average response time primarily due to the reason that the CNN is based on Convolutions, and training the Neural Network and testing consume a lot of GPU cycles which increases the processing time of the queries. Moreover, the CNN when coupled with K-means clustering tends to make the application bulky increasing the overall response time. Finally, the Graph Based Ontology Techniques also consume a lot of CPU cycles mainly due to the fact that large Ontological Graph is traversed using BFS or DFS and path-based computations are done which tend to increase the overall complexity of the application, thereby increasing the overall response time.

The case in OntoInfoG++ is much different as the Ontologies and Auxiliary Knowledge are fed into the framework as Knowledge Embeddings representing the most distinct relations. The entities are populated and fed from several sources using SPARQL Endpoint which co-operate with the actual recommendation application framework but does not consume its burst time. Similarly, the agents modeled using JADE and AgentSpeak also co-operate together in parallel to compute the semantic relatedness and load Ontologies and Entities Dynamically. In OntoInfoG++ there is no concept of path-based traversal or learning from the dataset. Instead, the OntoInfoG++ infers from the knowledge embeddings which are fed as auxiliary knowledge from various concrete sources. The knowledge is already reasoned out, modeled, and accepted by a community and the OntoInfoG++ infers from the knowledge embeddings by inferencing through agents which makes OntoInfoG++ quite light weight in nature and has the least response time of 0.38 ms.

The qualitative evaluation results for the query "Chemistry in Everyday Life" for infographics recommendation is depicted for the proposed OntoInfoG++ and the baseline models which were used for comparison. Fig. 6 depicts the top 10 infographic recommendations collaged as a single image for the OntoInfoG++. It is quite clear that the individual infographics are quite relevant to the query and the essence of the query is clearly visible by yielding infographics, which are not only the best fit to the query but also yield infographics that are quite informative and create a chronology between the individual infographics yielded. Fig. 7 furnishes the infographics which are yielded by the PCMSI model. The infographics yielded focuses mainly on the term "Chemistry" in general and the essence of the query "Chemistry in Everyday Life" is not brought out by the top 10 infographics yielded by the PCMSI model as it is a graphical model which factors priorities among elements and a spectrum based comparison is followed where each time pairwise relations has to be computed. The infographics furnished by the Collaborative Filtering with Fuzzy c-means clustering sandwich model is depicted by Fig. 8 where again the term "Chemistry" is given more weightage than the query "Chemistry in Everyday Life". However, the CNN with K-means

clustering furnishes the results in Fig. 9 where the query is learnt, and the essence of the query "Chemistry in Everyday Life" is brought out in a few recommendations while most of the recommendations cater to the generic query term "Chemistry". It is very clear from the qualitative analysis that the proposed OntoInfoG++ furnishes results that are quite comprehensive to the query term and ensures that the essence of the query terms is preserved as a whole. The relevance of results in OntoInfoG++ is comprehended mainly because of the usage of three distinct measures for computing the semantic relatedness, namely the EnAPMI measure, Horn's Index, and the Information Entropy. Apart from this, the encompassment of Ontology Alignment, usage of several cognitive real-world Knowledge Sources, namely the BibSonomy, DBpedia, Wikidata, LOD Cloud, and Crowd-Sourced Ontologies, supplies auxiliary knowledge and populates entities which increase both the diversity and the relevance of results.

Fig. 6. Qualitative Results of OntoInfoG++ for the Query "Chemistry in Everyday Life".

Fig. 7. Qualitative Results of PCMSI for the Query "Chemistry in Everyday Life".

Fig. 8. Qualitative Results of Collaborative Filtering with Fuzzy c-means clustering for the Query "Chemistry in Everyday Life".

Fig. 9. Qualitative Results of CNN-K-Means Clustering for the query "Chemistry in Everyday Life".

VII. Conclusion

Infographics is a very effective tool to represent lots of information in a single picture which can be easily understood and memorized and recommending the relevant and appealing infographics for the query will make it efficient and will enable the users in learning. A semantically driven knowledge fusion approach, OntoInfoG++ has been proposed to recommend infographics based upon the user queries and user clicks. The OntoInfoG++ achieves topic enrichment by integrating entities from real-world knowledge sources like the BibSonomy, DBpedia, Wikidata, LOD Cloud, and Crowd Sourced ontologies. The semantic alignment is achieved by computing the semantic similarity between the knowledge graph formulated from the enriched topic of interest and the Ontology formulated from the titles and keywords of research papers from the dataset. The semantic similarity computation has been realized with three distinct measures namely the EnAPMI, Horns's index, and the information entropy amalgamated through an agent. The OntoInfoG++ has achieved an overall accuracy of 97.21% with a very low FDR of 0.02 with a very low response time of 0.39 ms for the experimentations conducted on RARD II dataset which makes OntoInfoG++, the best in class approach for recommendation of infographics from research papers. The high value of nDCG furnished by the proposed OntoInfoG++ indicates that OntoInfoG++ has solved the serendipity problem by improving the diversification of recommended results.

Acknowledgment

The authors thank the Ministry of Human Resources Development, India and the National Institute of Technology, Tiruchirappalli for funding this research by timely release of HTRA Research Fellowship. The authors thank God the Eternal Father and Lord Jesus Christ for providing the required knowledge and insights for carrying our this work.

References

[1] Siricharoen, Waralak, "Infographics: The New Communication Tools in Digital Age," *Proceedings of The International Conference on E-Technologies and Business on the Web, Bangkok, Thailand* , pp.169-174, 2013.

[2] Sujia Zhu, Guodao Sun, Qi Jiang, Meng Zha, Ronghua Liang, "A Survey on Automatic Infographics and Visualization Recommendations," *Visual Informatics*, vol. 4, no. 3, pp. 24-40, 2020.

[3] Wilkinson JL, Strickling K, Payne HE, Jensen KC, West JH, "Evaluation of Diet-Related Infographics on Pinterest for Use of Behavior Change Theories: A Content Analysis," *JMIR Mhealth Uhealth*, vol.4, no.4, pp.1-11, 2016.

[4] Featherstone, Robin, "Visual Research Data: an Infographics Primer," *Journal of the Canadian Health Library Association*, vol. 35, no.4, pp. 147-150, 2014.

[5] Mohd Noh, Mohd Amin & Shamsudin, Wan Nur & Amin Nudin, Anith & Narimah, Nik & Harun, Mohd, "The Use of Infographics as a Tool for Facilitating Learning," *Proceedings of The International Colloquium of Art and Design Education Research, Malaysia, Springer*, pp.559-567, 2014.

[6] Siricharoen, Waralak & Siricharoen, Nattanun, "How Infographic Should be Evaluated?," *Proceedings of The 7th International Conference on Information Technology, Amman, Jordan*, pp.558-564, 2015.

[7] Murray, Iain & Murray, A. & Wordie, Sarah & Oliver, Chris & Murray, A. & Simpson, Hamish, "Maximising the Impact of Your Work Using Infographics," *Bone and Joint Journal*, vol.6, no.11, pp.619-620, 2017.

[8] Cifçi, Taner, "Effects of Infographics on Students Achievement and Attitude towards Geography Lessons," *Journal of Education and Learning*, vol.5, no.1, pp.154-166, 2016.

[9] Nuhoglu Kibar, Pınar & Sullivan, Kevin & Akkoyunlu, Buket, "Creatıng Infographics Based on the Bridge21 Model for Team-based and Technology-mediated learning," *Journal of Information Technology*

Education: Innovations in Practice, vol.18, pp.87-111, 2019.

[10] Chen, Z., Wang, Y., Wang, Q., Wang, Y., & Qu, H, "Towards Automated Infographic Design: Deep Learning-based Auto-Extraction of Extensible Timeline," *IEEE Transactions on Visualization and Computer Graphics*, vol.26, no.1, pp.917-926, 2020.

[11] Cui, W., Zhang, X., Wang, Y., Huang, H., Chen, B., Fang, L., Zhang, H., Lou, J., & Zhang, D, "Text-to-Viz: Automatic Generation of Infographics from Proportion-Related Natural Language Statements," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no.1, pp. 906-916, 2020.

[12] Jock Mackinlay, "Automating the Design of Graphical Presentations of Relational Information," *ACM Transactions on Graphics*. vol.5, no.2, pp.110–141, 1986.

[13] Deepak, Gerard, and A. Santhanavijayan, "OntoBestFit: A Best-Fit Occurrence Estimation strategy for RDF driven faceted semantic search," *Computer Communications, Elsevier*, vol.160, pp.284-298, 2020.

[14] Middleton S.E., De Roure D, "Shadbolt N.R. Ontology-based Recommender Systems," *Staab S., Studer R. (eds) Handbook on Ontologies, International Handbooks on Information Systems. Springer, Berlin, Heidelberg*, pp.477-498, 2004.

[15] Werner, David & Cruz, Christophe & Nicolle, Christophe, "Ontology-based Recommender System of Economic," *ArXiv*, vol.1, pp.1-4, 2012.

[16] Peis, E. & Morales-del-Castillo, José & Delgado-López, J, "Semantic Recommender Systems. Analysis of the state of the topic," *Hipertext.net; Edició en anglès*, vol.6, pp.1-10, 2008.

[17] Pazahr, Ali & Samper Zapater, J. Javier & Garcia-Sanchez, Francisco & Botella, Carmen & Martínez, Rafael, "Semantically-enhanced Advertisement Recommender Systems in Social Networks," *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services, Singapore*, pp.179-189, 2016.

[18] Bafna, Prafulla & Shirwaikar, Shailaja & Pramod, Dhanya, "Task recommender system using semantic clustering to identify the right personnel," *VINE Journal of Information and Knowledge Management Systems*, vol.49, no.4, pp.1-28, 2019.

[19] Huijsduijnen, Lies & Hoogmoed, Thom & Keulers, Geertje & Langendoen, Edmar & Langendoen, Sanne & Vos, Tim & Hogenboom, Frederik & Frasincar, Flavius & Robal, Tarmo, "Bing-CSF-IDF+: A Semantics-Driven Recommender System for News," *New Trends in Databases and Information Systems*, 2020.

[20] O. Tymchenko, S. Vasiuta, O. Khamula, O. Sosnovska and M. Dudzik, "Using the method of pairwise comparisons for the multifactor selection of infographics design alternatives," *20th International Conference on Research and Education in Mechatronics, Wels, Austria*, pp. 1-6, 2019.

[21] Deepak, Gerard, and Dheera Kasaraneni, "OntoCommerce: An Ontology Focused Semantic Framework for Personalised Product Recommendation for User Targeted E-commerce," *International Journal of Computer Aided Engineering and Technology*, vol.11, no. 4-5, pp.449-466, 2019.

[22] Deepak, G., Teja, V., & Santhanavijayan, A, "A Novel Firefly Driven Scheme for Resume Parsing and Matching Based on Entity Linking Paradigm," *Journal of Discrete Mathematical Sciences and Cryptography*, vol.23, no.1, pp.157-165, 2020.

[23] 1S. Haribabu, P. S. Sai Kumar, S. Padhy, G. Deepak, A. Santhanavijayan and N. Kumar D., "A Novel Approach for Ontology Focused Inter-Domain Personalized Search based on Semantic Set Expansion," *Fifteenth International Conference on Information Processing, Bengaluru, India, IEEE*, pp. 1-5, 2019.

[24] Deepak, Gerard, Naresh Kumar, G. VSN Sai Yashaswea Bharadwaj, and A. Santhanavijayan, "OntoQuest: An Ontological Strategy for Automatic Question Generation for e-assessment using Static and Dynamic Knowledge," *Proceedings of Fifteenth International Conference on Information Processing, Bengaluru, India, IEEE*, pp.1-6, 2019.

[25] Deepak, G., & Priyadarshini, J. S. "Personalized and Enhanced Hybridized Semantic Algorithm for Web Image Retrieval Incorporating Ontology Classification, Strategic Query Expansion, and Content-Based Analysis," *Computers & Electrical Engineering, Elsevier,* vol.72, pp.14-25,2018.

[26] Berkani, Lamia & Belkacem, Sami & Ouafi, Mounira & Guessoum, Ahmed, "Recommendation of Users in Social Networks: A Semantic and Social Based Classification Approach," *Expert Systems*, vol.38, no.2, 2020.

[27] Javed, Umair & Shaukat Dar, Kamran & Hameed, Ibrahim & Iqbal, Farhat & Mahboob Alam, Talha & Luo, Suhuai, "A Review of Content-Based and Context-Based Recommendation Systems," *International Journal of Emerging Technologies in Learning (iJET).* vol.16, no.3, pp.274-306, 2021.

[28] Houari, N. S., & Taghezout, N, "An Effective Tool for the Experts' Recommendation Based on PROMETHEE II and Negotiation: Application to the Industrial Maintenance," *International Journal of Interactive Multimedia and Artificial Intelligence*, no.6, pp. 67-77, 2021.

[29] Bobadilla, J., Lara-Cabrera, R., González-Prieto, Á., & Ortega, F, "DeepFair: Deep Learning for Improving Fairness in Recommender Systems," *International Journal of Interactive Multimedia and Artificial Intelligence*, no.6, pp.86-94, 2021.

[30] Pushpa, C. N., Gerard Deepak, J. Thriveni, and K. R. Venugopal, "Onto Collab: Strategic Review Oriented Collaborative Knowledge Modeling using Ontologies," *Proceedings of The Seventh International Conference on Advanced Computing, Chennai, India, IEEE*, pp. 1-7, 2015.

Gerard Deepak

Gerard Deepak holds a Masters in Engineering degree in Computer Science and Engineering from UVCE, Bangalore University. He is a University Level First Rank Holder for his masters and has qualified the KSET examination. Currently he is pursuing ahis PhD from National Institute of Technology, Tiruchirappalli. He has a h-index of 17 and has received the Budding researcher award from NITT twice successively. He also has received the best paper award 7 times in several international conferences in India and Abroad. He has 50 articles to his credit to date in both Journals and Conferences of repute. His areas of interests include Semantic Web, eXplainable AI, Semantic Web Mining, and Ontology Engineering.

V. Adithya

V. Adithya is a student pursuing his final year Undergraduate degree in Computer Science and Engineering. He is an AI and Machine Learning enthusiast who has also published and presented papers in many international conferences and has also received one best paper award. His other areas of interest include Software Engineering, Deep learning and eXplainable AI.

Dr. Santhanavijayan A

Dr. Santhanavijayan A holds a PhD from National Institute of Technology Tiruchirappalli and a Masters in Computer Engineering from Anna University, Chennai. He has received the Best Performing Faculty Award under his cadre from National Institute of Technology Tiruchirappalli. His research interests are Semantic Web, Natural Language Processing, Deep Learning, and Data Science.

Adaptation of Applications to Compare Development Frameworks in Deep Learning for Decentralized Android Applications

Beatriz Sainz-de-Abajo[1] *, Sergio Laso[2], Jose Garcia-Alonso[3], Javier Berrocal[3]

[1] Universidad de Valladolid, Valladolid (Spain)
[2] Global Process and Product Improvement S.L., Cáceres (Spain)
[3] Universidad de Extremadura, Cáceres (Spain)

Received 24 April 2022 | Accepted 14 March 2023 | Published 19 April 2023

Abstract

Not all frameworks used in machine learning and deep learning integrate with Android, which requires some prerequisites. The primary objective of this paper is to present the results of the analysis and a comparison of deep learning development frameworks, which can be adapted into fully decentralized Android apps from a cloud server. As a work methodology, we develop and/or modify the test applications that these frameworks offer us a priori in such a way that it allows an equitable comparison of the analysed characteristics of interest. These parameters are related to attributes that a user would consider, such as (1) percentage of success; (2) battery consumption; and (3) power consumption of the processor. After analysing numerical results, the proposed framework that best behaves in relation to the analysed characteristics for the development of an Android application is TensorFlow, which obtained the best score against Caffe2 and Snapdragon NPE in the percentage of correct answers, battery consumption, and device CPU power consumption. Data consumption was not considered because we focus on decentralized cloud storage applications in this study.

Keywords

Android Applications, Decentralized, Deep Learning, Framework, Images, TensorFlow.

DOI: 10.9781/ijimai.2023.04.006

I. Introduction

THE availability of large volumes of data allows the evolution of artificial intelligence (AI) [1], [2]. For the first time in the history of humankind, systems can analyse the information generated at exponentially faster speeds.

Machine learning (ML) is the practice of using algorithms to analyse data, learn from it and make a prediction about something [3]. ML and deep learning (DL) algorithms must manage a lot of information to produce results that accurately describe reality [4], [5]. That information can draw conclusions about what we think and feel.

These models have drawn ever-increasing research interest due to their intrinsic capability to overcome the drawbacks of traditional algorithms [6]. ML, DL and IA have grown in their use given their benefits in different contexts [7]. In recent years, many studies have shown that combining ML and DL techniques is especially useful in image analysis [8], [9]. DL have proven effectiveness in object and image recognition, natural language processing, speech recognition, robot navigation systems, self-driving cars and health care. [10], [11]. This allows the precise detection of a disease, locating stolen and sold objects via the Internet, searching for missing persons, etc. Due to its great potential, this technique is applied in a large number of sectors such as security, health, finance, automotive and agriculture. However, DL takes ML to a more detailed level, reducing the margin of error and increasing the accuracy of the conclusions it reaches [12], [13]. In this case, the system goes through layers or neuronal units. While in ML, to perform a classification, it is necessary to indicate the characteristics; in DL, the algorithm will perform the classification during the training by itself.

Each layer processes the information and returns a result in the form of weighting. The second layer that analyses the image will combine the result obtained by the first layer with its own judgement. As a result, the weighting will change. The third layer will use this new modified weighted result to perform its calculations, reducing the margin of error and thus increasing the accuracy of its results. The system trains itself due to a large amount of information being considered, improving its weighting.

Data storage and preparation tasks for further processing require the most time [14] but are essential because AI algorithms develop complex processes of understanding and interpreting data and therefore need them to provide value.

Although existing ML and DL services use cloud computing and servers to run, and therefore require an Internet connection, there is a trend towards decentralization [15], [16]. [17] argues that decentralizing AI opens the door for more equitable development. Instead of connecting to data centre-based services, queried through mobile communications, AI capabilities will reside on the device itself.

* Corresponding author.
E-mail address: beasai@uva.es

ML and DL are the key technologies on which new functionalities, personalization and connectivity with other devices in the Internet of Things (IoT) will be based.

II. Tasks and Methods

This study has been structured into three tasks: (A) revision of the most well-known frameworks; (B) test application development; and (C) analysis and comparison. Finally, we show the results of this study. Fig. 1 shows the flow chart followed in this study.

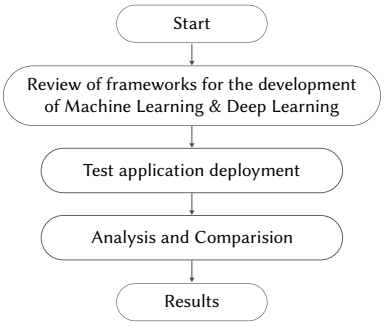


Fig. 1. Flow chart.

A. Review of Frameworks for the Development of Machine Learning & Deep Learning

The eight most commonly used frameworks globally were reviewed [18]. Table I shows links to the official websites that list the best features of each framework and indicates whether integration with Android is allowed.

TABLE I. ML & DL Frameworks

	Name	Official Web	Android integration
1	**TensorFlow** [19]-[21]	https://www.tensorflow.org/	Yes
2	**Caffe** [22]	https://caffe.berkeleyvision.org/	No
3	**Caffe2** [a] [23]	https://pytorch.org/	Yes
4	**Amazon Machine Learning** [24]	https://aws.amazon.com/es/machine-learning/	No
5	**CNTK** [25]	https://docs.microsoft.com/en-us/cognitive-toolkit/	No
6	**Torch** [26]	http://torch.ch/	No
7	**Snapdragon NPE**	https://developer.qualcomm.com/software/qualcomm-neural-processing-sdk	Yes
8	**DeepLearning4J** [27]	https://deeplearning4j.konduit.ai/	Yes

[a] Caffe2 and PyTorch projects are merging now [23].

After a first filter, we analyse the frameworks that are exportable to Android. Table II summarizes the requirements to be executed on a smartphone and its uses.

B. Test Application Deployment

We define the functional requirements (FR) and non-functional requirements (NFR) of applications to be developed or adapted.

All applications have the same requirements so that they behave in a similar way.

- FR-1. Through the trained model, the included images can be recognized.
- FR-2. The applications show the recognition result together with a percentage of success probability.
- FR-3. The recognition will be static from a photograph included in the application.
- NFR-1. The applications must achieve reasonable response times when executing the deep learning model.
- NFR-2. The applications will be functional for smartphones with an Android OS 6.0 or higher operating system.

For each framework, we implement the corresponding application. Whenever possible, we use the test applications from the official repositories because it would take a long time to implement the integration of these frameworks in Android from the beginning. If necessary, we make adjustments and developments, such as modifications to the code so that all applications have the same features.

The primary goal of this study is not to create commercial applications but rather to provide simple functionality to facilitate the objective of this study: to compare frameworks under equal conditions.

The applications include an image gallery. The user clicks on each image, and the application shows the result of the recognition and the probability of success of the clicked image.

Listed below are the changes implemented in each application and the problems found:

1. TensorFlow app.

   The TensorFlow app was retrieved from the official TensorFlow repository [28]. The most important change was to modify the primary functionality of the application. TensorFlow originally used a live camera to recognize objects, but this was changed to measure parameters correctly in the subsequent comparison. The functionality changed to a list of images where the user clicks an image, and the application returns the recognition result.

2. Caffe2 app.

   The AICamera application was obtained from the official repository [29]. The most important change was the same as that in the TensorFlow application. Caffe2 integrates with C++ to perform model recognition and execution.

3. Snapdragon NPE app.

   The SNPE Image Classifier application was obtained from the SDK, available on the official Qualcomm Developer Network repository [30]. No changes were necessary because the application provided the functions that were proposed in the requirements. Its development is only possible on Linux because it uses Snapdragon libraries that are included in the repository and only compatible on Linux.

4. Deeplearning4J app.

   The DL4JImageRecognitionDemo application was obtained from the official repository [31]. Due to its limited development to date, it has not been possible to use this application. Although it was modified, the result was unsuccessful. The application did not compile correctly, possibly due to bugs with the libraries or some type of incompatibility. Because implementing an application that integrates the framework from the beginning requires a long time, this application was discarded for the testing phase.

When defining the requirements of the model before training, the authors agreed to use a pre-trained DL model that was available in the frameworks because each framework has a different format.

The operation is the same for all three applications, which have a list of images that the user clicks on. The apps then display the recognition result along with the probability of success, which the model thinks is the clicked product. At this stage, we make a limited number of attempts.

TABLE II. Frameworks Compatible With the Android Operating System

Name	Characteristics	Requirements	Uses
TensorFlow	• Execution of neural models. • Hardware acceleration thanks to the Android Neural Networks API.	• Android API 23 (Marshmallow) or later and NDK 12b or later.	• Computer vision. • Voice and image recognition. • Medical applications. • Intelligent searches. • Intelligent answers in emails.
Caffe2*	• Execution of neural models. • Hardware acceleration thanks to the Android Neural Networks API. • Offers conversion from Torch models to Caffe2.	• Android API 21 (Lollipop) or higher.	• Computer vision. • Voice and image recognition. • Translation. • Chatbots. • IoT. • Medical applications.
Snapdragon NPE	• Execution of neural models. • Compatibility with TensorFlow, Caffe and Caffe2. • Developed on Linux.	• For GPU: Qualcomm Snapdragon 845, 820, 835, 625, 626, 650, 652, 653, 660, 630, 636, and 450. • For Adreno GPU: libOpenCL.so	• Object classification. • Face detection. • Natural language understanding. • Speech recognition. • Security/authentication. • Resource management.
DeepLearning4J	• To create & train a neural network on an Android device.	• Android API 21 (Lollipop) or higher.	• Object and speech recognition. • Natural language processing. • Data prediction.

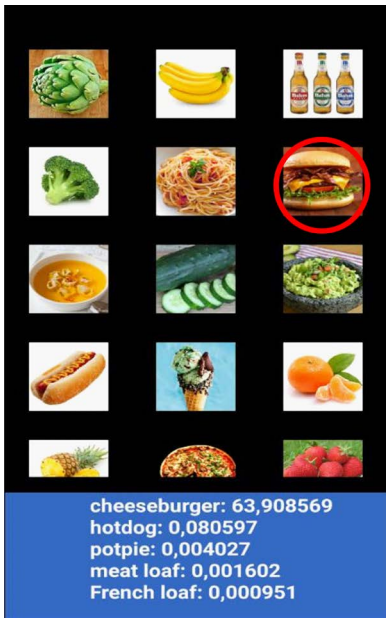


Fig. 2. Test in TensorFlow.

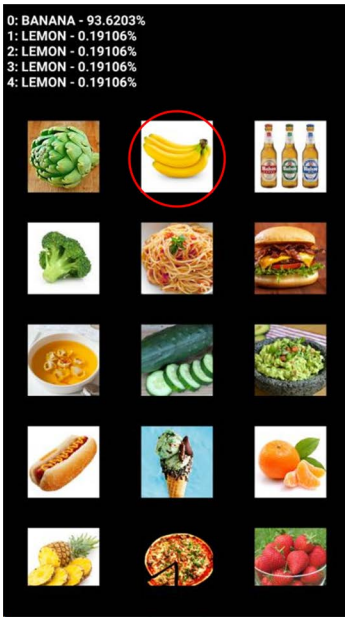


Fig. 3. Test in Caffe2.

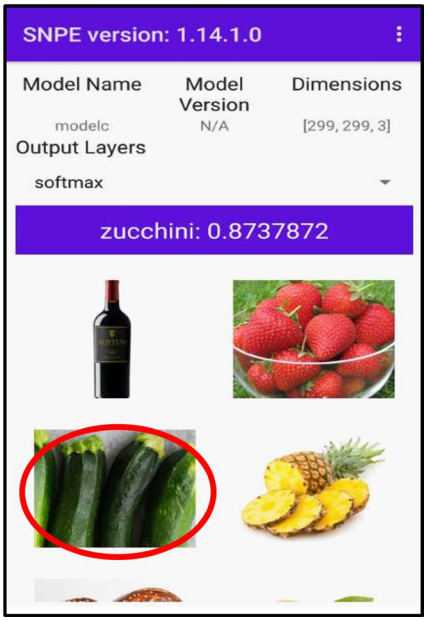


Fig. 4. Test in Snapdragon NPE.

Fig. 2, Fig. 3 and Fig. 4 show the image recognition of each application. We highlight with a red circle the image pressed during the preliminary evaluation tests.

Fig. 2 shows the test performed with TensorFlow when clicking on the photo of a cheeseburger. The percentage of success is 63.91%, and those of the other images were less than 10%.

In the test with Caffe2, the percentage is 93.62% after choosing the photo of bananas, with the other options near 0% (see Fig. 3)

In Fig. 4, after testing on the Snapdragon NPE, we also obtained a high success rate (87.38%) when selecting zucchini.

C. Analysis and Comparison

As basic applications, there are no differences in the aspects related to code optimization. The proposed analysis parameters were (1) success rate; (2) battery consumption; and (3) processor power consumption.

1. *Success rate:* The result is determined by the success or failure of each food. When all frameworks obtain the same successes or failures, by of the highest percentage of probability, i.e., the one with the highest probability of success for each detection. For this purpose, the same gallery with 20 food images was included in each application. After this, a round was performed in which each image was clicked on, and the result of the prediction was saved in a table. We checked the results in all cases along with the percentage of the probability of success (see Table III).
2. *Battery consumption:* Nowadays, we are always on the lookout for our smartphone's battery to reach the end of the day without running out, because with the multimedia content we watch and the hours of use we give it, few smartphones have large battery capacities. For this reason, it is essential to choose the framework that consumes the least so that the user notices it as little as possible. In order to evaluate this parameter, the new versions of Android,

TABLE III. PERCENTAGE OF CORRECT ANSWERS

Food	Image	TensorFlow	Caffe2	Snapdragon NPE (CPU)	Snapdragon NPE (GPU)
Artichoke		Artichoke 83.24%	Artichoke 99.99%	Artichoke 99.74%	Artichoke 99.70%
Banana		Banana 80.10%	Banana 93.91%	Banana 51.02%	Banana 51.66%
Beer Bottle		Beer Bottle 67.65%	Scale 55.14%	Beer Bottle 54.38%	Beer Bottle 54.68%
Broccoli		Broccoli 83.18%	Broccoli 99.74%	Broccoli 99.56%	Broccoli 99.51%
Burrito		Burrito 82.31%	Burrito 99.93%	Pinwheel 19.58%	Pinwheel 15.36%
Carbonara		Carbonara 83.17%	Carbonara 99.07%	Swab 41.59%	Swab 36.49%
Cheeseburger		Cheeseburger 83.17%	Hot Dog 47.40%	Cheeseburger 72.18%	Cheeseburger 70.55%
Consommé		Consommé 78.59%	Consommé 98.19%	Washbasin 60.95%	Washbasin 69.87%
Cucumber		Cucumber 83.20%	Cucumber 98.95%	Cucumber 99.85%	Cucumber 99.80%
Guacamole		Guacamole 82.98%	Guacamole 98.86%	Mortar 37.96%	Mortar 40.62%
Hotdog		Hot Dog 82.86%	Hot Dog 99.87%	Jellyfish 3.62%	Jellyfish 3.32%
Ice Cream		Ice Cream 24.05%	Honeycomb 33.80%	Pedestal 12.36%	Pedestal 15.29%
Meat Loaf		Meat Loaf 83.13%	Meat Loaf 67.83%	Ice Lolly 22.66%	Ice Lolly 28.56%
Orange		Orange 77.62%	Orange 77.56%	Orange 43.36%	Orange 41.82%
Pineapple		Pineapple 77.76%	Spaghetti squash 51.51%	Necklace 23.38%	Necklace 20.78%
Pizza		Pizza 80.11%	Pizza 99.19%	Wall clock 13.53%	Wall clock 13.64%
Pretzel		Pretzel 83.27%	Pretzel 99.40%	Pretzel 90.92%	Pretzel 89.79%
Strawberry		Strawberry 83.22%	Strawberry 94.68%	Golf Ball 82.70%	Golf Ball 82.47%
Wine bottle		Red Wine 69.69%	Whistle 30.39%	Wine Bottle 82.07%	Wine Bottle 80.71%
Zucchini		Zucchini 79.95%	Cucumber 78.24%	Zucchini 88.22%	Zucchini 87.30%
Correct		**20**	**14**	**11**	**11**
Wrong		**0**	**6**	**9**	**9**
Percentage of correct answers		**100%**	**70%**	**55%**	**55%**

Google offers a web service called Battery Historian [32], in which we enter a log obtained from Batterystats. Batterystats.bin is a file that works as a registry, where Android saves the consumption data of the mobile device either via hardware or software services. The operating system uses this file to monitor consumption and battery level, and to display consumption statistics. The operating system is programmed to reset the file when the battery is fully recharged. When the battery is discharged, we record new data about battery use and charging. With the tool, we will evaluate (1) device estimated power use; (2) device estimated power use due to CPU usage; and (3) CPU user time.

3. *Processor power consumption:* One of the factors affecting the battery is CPU consumption. For this reason, we will measure what percentage is consumed each time an image recognition is performed. Another factor is that the lower the CPU power consumed, the faster and smoother the app will experience on lower-end devices with a more moderate processor, so the app will cover more of the market. To measure the power consumed, the Android Profiler tool from Android Studio was used [33]. This tool provides real-time data related to the CPU, memory and network activity of an application. You can perform sample-based method tracing for time code execution, capture dumps, view memory allocation, and inspect information from files transmitted over the network. In this study, we focus on the CPU Profiler, which shows the power that the CPU is consuming on any interaction that we make in the system or after selecting a specific application in real time.

One of the premises of this study is that the application is integrated into the device and, therefore, decentralized with respect to any server. Thus, it is not necessary to measure data traffic.

D. Resources Used

The hardware resources used in this study included the following:

- PC Asus X54HR
- Smartphone Xiaomi Mi3
  - To test the test applications.
  - To measure the parameters to compare the test applications.
- Smartphone Xiaomi Redmi Note 4
  - To test the analysis with the GPU offered by the Snapdragon Neural Processing Engine (Snapdragon NPE), because its processor and GPU are compatible with the framework for that function.

The tools for the implementation and development of the applications are as follows:

- Android Studio.
- Inception V3 model.
  - Trained by ImageNet content (https://image-net.org/) with data from 2012.
  - This model is composed of more than 1000 different classes: objects, animals, food, etc.

III. Results

The framework that stands out in a single parameter is not the best but the one that is more balanced considering all parameters. Next, we show the results for each framework.

A. Success Rate

To perform tests whose results are comparable, every time the test on each application was performed, all system applications were closed; having applications open in the background may influence the measurements. Regarding the Snapdragon NPE application, a smartphone compatible with GPU analysis was used to verify the differences between frameworks. Table III shows the results obtained.

Tensorflow has a 100% success rate. It also shows a stable behaviour, i.e., it obtains a high probability percentage in the cases it gets right, with a small exception. In the detection of the Ice Cream image, the prediction is correct but the probability percentage is low (24.05%).

Caffe2 is in second place with a success rate of 70%. Despite not achieving a 100% hit rate, the behaviour is also stable. In the cases it succeeds it obtains high probability percentages, and in the cases it fails it gets low probability percentages no higher than 55%. Therefore, we can easily detect whether a detection is wrong based on the probability percentage. Regarding the wrong predictions, in most cases the result is not similar or interpretable with the original. For example, with Pineapple the framework has detected Spaghetti squash, or with Wine bottle the detection obtained has been Whistle.

Snapdragon NPE is in third and last place with a success rate of 55%. Its results have been fair/poor, even with the GPU analysis that supposedly increases performance. The behaviour has not been as stable as in the previous frameworks. In certain successful results it gets a low percentage of probability, for example Banana (51.02%) or Orange (43.36%). The opposite also occurs. With Strawberry, it detects Golf Ball with 82.70%. Regarding the wrong predictions, some of the wrong results obtained, if they can resemble with the original, e.g. Guacamole and Mortar or Strawberry and Golf Ball.

Although the three applications use the same image recognition model (Inception V3), the model has to be adapted to each framework, so performance may change [34]. It can also affect the software optimization of each framework in the operating system (OS). In this case, Tensorflow is developed by Google, the same developer as the Android OS, so it could be better optimized and therefore get better results [35].

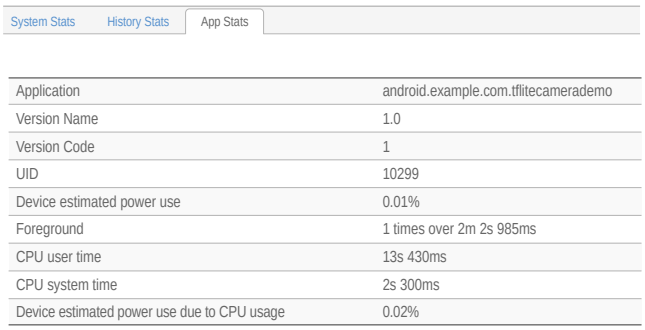
B. Battery Consumption

A 2-minute execution test was performed in which each image was analysed 2 times.

Before starting the test and running the applications, we reset the device's consumption log file and its history using the "adb shell dumpsys batterystats" command. Fig. 5, Fig. 6 and Fig. 7 show the captures made by the battery historian tool.

The frameworks that consume less power are Caffe2 and TensorFlow. Snapdragon NPE performs the worst, with a difference of 10% (estimated battery consumption) compared to TensorFlow.

The "CPU user time" in TensorFlow and Caffe2 is 13.430 s and 34.320 s, respectively, while in Snapdragon NPE, it is 112.530 s, indicating much higher consumption with Snapdragon NPE. While TensorFlow and Caffe2 seem to use the CPU only when they parse the image or update the display, Snapdragon NPE constantly consumes resources.


System Stats | History Stats | App Stats

Application	android.example.com.tflitecamerademo
Version Name	1.0
Version Code	1
UID	10299
Device estimated power use	0.01%
Foreground	1 times over 2m 2s 985ms
CPU user time	13s 430ms
CPU system time	2s 300ms
Device estimated power use due to CPU usage	0.02%

Fig. 5. Tersorflow battery test.

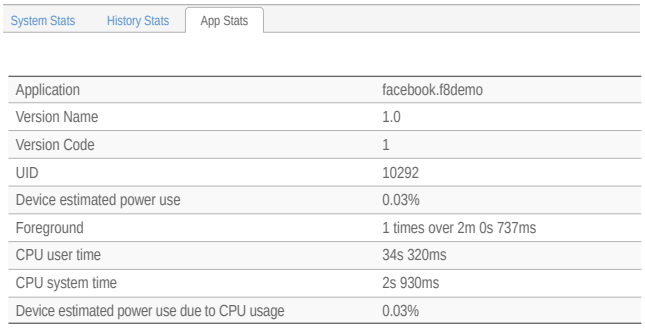


Fig. 6. Caffe2 battery test.

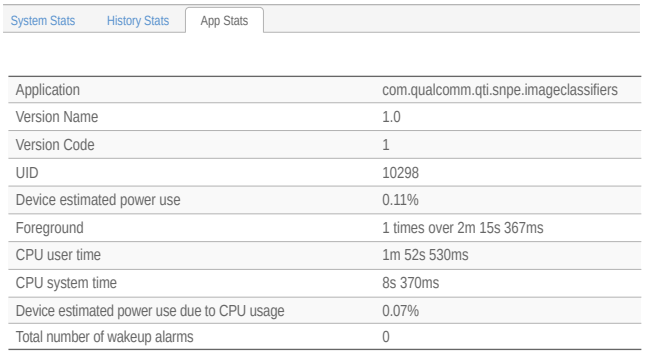


Fig. 7. Snapdragon NPE battery test.

TABLE IV. Battery Consumption

	Device estimated power use	Device estimated power use due to CPU usage	CPU user time
TensorFlow	0.01%	0.02%	13s 430ms
Caffe2	0.03%	0.03%	34s 320ms
Snapdragon NPE	0.11%	0.07%	1min 52s 530ms

C. CPU Power Consumption

In this test, we connect the mobile device to a PC to use an Android CPU Profiler. With this tool, we can measure the power consumed by the application when it analyses an image.

We report the average calculation in a given timeframe in which the application analyses ten images. After the calculations, we obtain an average consumption for the analysis of a single image (see Table V).

TABLE V. CPU Power Consumption

	Average consumption per image analysis
TensorFlow	38%
Caffe2	35%
Snapdragon NPE	43%

D. Summary

Although there are no noticeable differences, Snapdragon NPE consumes more CPU time to analyse an image than Caffe2 and TensorFlow.

After the tests, a summary of the results of the comparison is shown in Table VI.

TABLE VI. Comparison Summary

	Percentage of correct answers	Battery consumption	CPU power consumption
TensorFlow	100%	0.01%	38%
Caffe2	70%	0.03%	35%
Snapdragon NPE	55%	0.11%	43%

IV. Conclusion and Future Lines of Work

There are more and more developers in the application market and, therefore, more competition. For this reason it is necessary to choose the framework with the best results, so that the user does not feel disappointed.

TensorFlow and Caffe2 produce have much better results than Snapdragon NPE, which also exhibited highest battery consumption and a fair to poor response success rates.

The battery and CPU consumptions are similar for TensorFlow and Caffe2, but the response rate is better with TensorFlow. Additionally, TensorFlow is a Google framework, has a large community on both Github and Stack Overflow, and is well documented with questions, reviews and tutorials online. TensorFlow also keeps a close eye on these user communities to improve their platform. Thus, based on this study's results, TensorFlow is the most recommended for the implementation of an Android application.

There are different areas where image recognition is applied that could benefit from this type of development. The most interesting are the areas of health and wellness. First, through the diagnosis of diseases, after analysing the alterations in the X-rays. In the second, improving the management of food purchases. With a simple application on the mobile, the user could check the lack of products in the pantry.

Having selected this framework, we plan to develop an assistance application for food that will allow a user to take one or more photos with a smartphone camera and recognize food using the trained model. Next, we plan to transfer the list of products from a kitchen pantry to a smartphone. Using the same recognition function, we plan to generate a shopping list by checking which products are missing in the pantry and which we should buy.

Acknowledgment

This study was supported in part by the DIN2020-011586 Grant, funded by MCIN/AEI/10.13039/501100011033 and the European Union "NextGenerationEU/PRTR," and by the Interreg V-A España-Portugal 2014-2020, under Project 0786_CAP4ie_4_.

It was also supported by the "movilidad investigadores e investigadoras UVa-BANCO SANTANDER 2022" grant.

References

[1] Y. Duan, J. S. Edwards, and Y. K. Dwivedi, "Artificial intelligence for decision making in the era of big data – evolution, challenges and research agenda," *International Journal of Information Management*, vol. 48, pp. 63–71, 2019, doi: 10.1016/j.ijinfomgt.2019.01.021.

[2] L. Spector, "Evolution of artificial intelligence," *Artificial Intelligence*, vol. 170, no. 18. pp. 1251–1253, Dec. 2006, doi: 10.1016/j.artint.2006.10.009.

[3] M. S. Mahdavinejad, M. Rezvan, M. Barekatain, P. Adibi, P. Barnaghi, and A. P. Sheth, "Machine learning for internet of things data analysis: a survey," *Digital Communications and Networks*, vol. 4, no. 3. pp. 161–175, 2018, doi: 10.1016/j.dcan.2017.10.002.

[4] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," *IEEE Access*, vol. 7. pp. 53040–53065, 2019, doi: 10.1109/

ACCESS.2019.2912200.
[5] W. Samek and K. R. Müller, "Towards explainable artificial intelligence," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11700 LNCS, 2019, pp. 5–22.
[6] N. Bouchra, A. Aouatif, N. Mohammed, and H. Nabil, "Deep belief network and auto-encoder for face classification," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 5, p. 22, 2019, doi: 10.9781/ijimai.2018.06.004.
[7] F. J. García-Peñalvo *et al.*, "Application of artificial intelligence algorithms within the medical context for non-specialized users: The cartier-ia platform," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, 2021, doi: 10.9781/ijimai.2021.05.005.
[8] S. H. Chen, C. W. Wang, I. H. Tai, K. P. Weng, Y. H. Chen, and K. S. Hsieh, "Modified yolov4-densenet algorithm for detection of ventricular septal defects in ultrasound images," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 7, 2021, doi: 10.9781/ijimai.2021.06.001.
[9] M. I. Khattak, M. Al-Hasan, A. Jan, N. Saleem, E. Verdú, and N. Khurshid, "Automated detection of covid-19 using chest x-ray images and ct scans through multilayer-spatial convolutional neural networks," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, 2021, doi: 10.9781/ijimai.2021.04.002.
[10] A. Venkat, T. Rusira, R. Barik, M. Hall, and L. Truong, "SWIRL: high-performance many-core CPU code generation for deep neural networks," *International Journal of High Performance Computing Applications*, vol. 33, no. 6, 2019, doi: 10.1177/1094342019866247.
[11] S. S. Nisha, M. M. Sathik, and M. N. Meeral, "Application, algorithm, tools directly related to deep learning," in *Handbook of Deep Learning in Biomedical Engineering: Techniques and Applications*, 2020.
[12] Y. Xin *et al.*, "Machine learning and deep learning methods for cybersecurity," *IEEE Access*, vol. 6, pp. 35365–35381, 2018, doi: 10.1109/ACCESS.2018.2836950.
[13] S. Dargan, M. Kumar, M. R. Ayyagari, and G. Kumar, "A survey of deep learning and its applications: a new paradigm to machine learning," *Archives of Computational Methods in Engineering*, vol. 27, no. 4, pp. 1071–1092, 2020, doi: 10.1007/s11831-019-09344-w.
[14] N. El Aboudi and L. Benhlima, "Big data management for healthcare systems: architecture, requirements, and implementation," *Advances in Bioinformatics*, vol. 2018. 2018, doi: 10.1155/2018/4059018.
[15] F. L. Koch, "Decentralized network management using distributed artificial intelligence," *Journal of Network and Systems Management*, vol. 9, no. 4, pp. 375–388, 2001, doi: 10.1023/A:1012976206591.
[16] I. Gupta, "Decentralization of artificial intelligence: analyzing developments in decentralized learning and distributed AI networks," *Researchgate.Net*, no. May, 2020, doi: 10.13140/RG.2.2.17018.93124.
[17] G. A. Montes and B. Goertzel, "Distributed, decentralized, and democratized artificial intelligence," *Technological Forecasting and Social Change*, vol. 141. pp. 354–358, 2019, doi: 10.1016/j.techfore.2018.11.010.
[18] Z. Wang, K. Liu, J. Li, Y. Zhu, and Y. Zhang, "Various frameworks and libraries of machine learning and deep learning: a survey," *Archives of Computational Methods in Engineering*, 2019, doi: 10.1007/s11831-018-09312-w.
[19] Google, "TensorFlow Lite guide," *TensorFlow*, 2020.
[20] Tutorials Point, "TensorFlow tutorial," *Tutorials Point Pvt. Ltd.*, p. 90, 2019.
[21] M. Abadi *et al.*, "TensorFlow: a system for large-scale machine learning," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016,* Savannah, GA, USA, 2016, pp. 265–283, doi: 10.5555/3026877.3026899.
[22] Y. Jia *et al.*, "Caffe: convolutional architecture for fast feature embedding," in *MM 2014 - Proceedings of the 2014 ACM Conference on Multimedia*, Nov. 2014, pp. 675–678, doi: 10.1145/2647868.2654889.
[23] Facebook, "Caffe2 and PyTorch join forces to create a research + production platform PyTorch 1.0," *Caffe2 Documentation*, 2018.
[24] A. Mishra, "Amazon machine learning," in *Machine Learning in the AWS Cloud*, 2019, pp. 317–351.
[25] F. Seide and A. Agarwal, "CNTK : Microsoft's open-source deep-learning toolkit," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 2135-2135, doi: 10.1145/2939672.2945397.
[26] R. Collobert, C. Farabet, and K. Kavukcuoğlu, "Torch | Scientific computing for LuaJIT.," *NIPS Workshop on Machine Learning Open Source Software*, 2008.
[27] S. Lang, F. Bravo-Marquez, C. Beckham, M. Hall, and E. Frank, "WekaDeeplearning4j: a deep learning package for Weka based on deeplearning4j," *Knowledge-Based Systems*, vol. 178, pp. 48–50, Aug. 2019, doi: 10.1016/j.knosys.2019.04.013.
[28] M. Abadi *et al.*, "TensorFlow, Large-scale machine learning on heterogeneous systems." 2015, doi: 10.5281/zenodo.4724125.
[29] Facebook, "AICamera application," 2017. https://github.com/facebookarchive/AICamera (accessed Jul. 28, 2022).
[30] ©2022 Qualcomm Technologies Inc. and/or its affiliated companies, "Qualcomm neural processing SDK for AI," 2022. https://developer.qualcomm.com/software/qualcomm-neural-processing-sdk (accessed Jul. 28, 2022).
[31] Eclipse Foundation, "deeplearning4j," *github.com*, 2019. https://github.com/eclipse/deeplearning4j (accessed Jul. 28, 2022).
[32] Google Developers, "Analyze power use with battery historian," 2022. https://developer.android.com/topic/performance/power/battery-historian (accessed Jul. 28, 2022).
[33] Google Developers, "The Android profiler," 2022. https://developer.android.com/studio/profile/android-profiler (accessed Jul. 28, 2022).
[34] A. Chowanda and R. Sutoyo, "Convolutional neural network for face recognition in mobile phones," *ICIC Express Letters*, vol. 13, no. 7, pp. 569–574, 2019, doi: 10.24507/icicel.13.07.569.
[35] H. C. Takawale and A. Thakur, "Talos App: on-device machine learning using TensorFlow to detect Android malware," in *2018 Fifth International Conference on Internet of Things: Systems, Management and Security*, Oct. 2018, pp. 250–255, doi: 10.1109/IoTSMS.2018.8554572.

Beatriz Sainz-de-Abajo

She is currently an Associate Professor in Telecommunications Engineering at the University of Valladolid in Spain. She received the Ph.D. degree (summa cum laude) in the University of Cordoba in 2009 and has a master's degree in Data Networks and Transportation Networks from Lucent Technologies. Her fields of action are the development and evaluation of e-Health systems, m-Health, medicine 2.0., cloud computing, etc., focuses on topics related to electronic services for the information society. She belongs to the GTe Research Group, integrated within the UVa Recognized Research Group "Information Society". Actually, she also collaborates with the research "Quercus Software Engineering Group" of the University of Extremadura, Spain. Among the lines of research, the group works to develop innovative solutions in the field of health that help patients improve their quality of life and facilitate the work of health professionals.

Sergio Laso-Mangas

He received the Industrial Ph.D. degree in computer science from the University of Extremadura, Spain, in 2023. He is currently a researcher at the company Global Process and Product Improvement, Cáceres, Spain. His research interests include mobile computing, pervasive systems, the Cloud-to-thing continuum, Quality of Service and the Internet of Things

José Manuel García-Alonso

He is an Associate Professor at the University of Extremadura, Spain and co-founder of Gloin, a software-consulting company and Health and Aging Tech an eHealth company. He got his PhD on software engineering at the University of Extremadura in 2014. He is currently working in the department of Computer and Telematics Systems Engineering, in the area of Languages and Computer Systems. He is part of the research "Quercus Software Engineering Group" and his research interests include quantum software engineering, mobile computing, pervasive computing, eHealth, gerontechnology.

Javier Berrocal

He received the Ph.D. degree in computer science from the University of Extremadura, Spain, in 2014. In 2016, he obtained an Associate position at the University of Extremadura. He is currently working in the department of Computer and Telematics Systems Engineering in the University of Extremadura. He is part of the research “Quercus Software Engineering Group”, and his main research interests are mobile computing, context awareness, pervasive systems, crowd sensing, the Internet of Things, and fog computing. He is a cofounder of the company Gloin, which is a software-consulting company, and Health and Aging Tech an eHealth company.

On the Importance of UX Quality Aspects for Different Product Categories

Martin Schrepp[1]*, Jessica Kollmorgen[2], Anna-Lena Meiners[2], Andreas Hinderks[3], Dominique Winter[4], Harry B. Santoso[5], Jörg Thomaschewski[2]

[1] SAP SE, Walldorf (Germany)
[2] University of Applied Sciences Emden/Leer, Emden (Germany)
[3] University of Seville, Seville (Spain)
[4] University of Siegen, Siegen (Germany)
[5] University of Indonesia, Depok (Indonesia)

Received 29 January 2022 | Accepted 29 August 2022 | Published 13 March 2023

Abstract

User experience (UX) is a holistic concept. We conceptualize UX as a set of semantically distinct quality aspects. These quality aspects relate subjectively perceived properties of the user interaction with a product to the psychological needs of users. Not all possible UX quality aspects are equally important for all products. The main use case of a product can determine the relative importance of UX aspects for the overall impression of the UX. In this paper, the authors present several studies that investigate this dependency between the product category and the importance of several well-known UX aspects. A method to measure the importance of such UX aspects is presented. In addition, the authors show that the observed importance ratings are stable, i.e., reproducible, and hardly influenced by demographic factors or cultural background. Thus, the ratings reported in our studies can be reused by UX professionals to find out which aspects of UX they should concentrate on in product design and evaluation.

Keywords

Questionnaires, User Experience, UX Quality Aspects, UX Measurement, UX Scales.

DOI: 10.9781/ijimai.2023.03.001

I. Introduction

Due to the availability of highly efficient development environments and agile deployment processes, it is increasingly difficult to position a product based on its superior functionality alone. Competitors can catch up fast concerning functions and features and it is easy for customers to switch from one cloud-based product to another. Thus, to be successful in the long run and build a loyal customer base, an interactive product needs to provide a high level of UX quality. Otherwise, the user base will decrease fast. This creates the need to measure the UX quality of a product continuously and compare the results to the UX of competing products [1].

But what do we exactly mean by the term UX? A UX definition that supports design decisions must relate the psychological needs [2]–[4] and goals of users to concrete properties of the user interface of a product [5]. We need a clear conceptualization of UX to enable different persons and roles (UX designers, software developers or product owners) in the development process of a product to contribute to efficient design discussions. If it stays unclear what UX means and which UX aspects are important for a product, there will always be misunderstandings that may cause long and useless discussions about the impact of specific design decisions on overall UX quality [6].

* Corresponding author.
E-mail address: martin.schrepp@sap.com

Since UX is undoubtedly a highly subjective impression, meaning that crucial aspects of the experience can only be ascertained during the first-hand perception by one (or more) user(s) and can consequently only be described by them, it is required to ask users about their experience of using a product [5]. In this sense, there is no objective method to measure the UX of a product. Thus, researchers need a conceptualization of UX that can be communicated easily to users.

Of course, we cannot simply ask "*How do you judge the user experience of this product?*", because this term will be interpreted inconsistently by end users. But concrete UX aspects can be easily transformed into questions or items in a UX questionnaire, so a clear understanding of relevant UX aspects is also a basis for UX measurement by using questionnaires.

Certainly, UX is a heterogeneous concept. If we review the research literature and existing UX questionnaires we will find aspects such as *Efficiency*, *Ease of learning*, *Dependability*, *Adaptability*, *Fun of use*, *Aesthetics*, *Loyalty*, etc. Which of these aspects are important for a specific product depends firstly on the user group (personal preferences and experience) and secondly on the product type [7]–[10].

This paper describes a set of UX quality aspects that relate properties of a product with the needs and expectations of users. Such a list of semantically distinct aspects should help UX designers and researchers to develop a better understanding of UX. It also helps to streamline design discussions by providing a common understanding of the relevant UX quality aspects [6] and can even be helpful when the final design is evaluated, for example with a questionnaire [11], [12].

In this paper, we present 5 independent studies with a total of 361 participants who have submitted nearly 70,000 ratings during the years 2015–2021. We clarify how important the described UX quality aspects are for different types of products. This provides some guidance for UX practitioners and researchers on which aspects of a design they should focus on during the design phase of new products and when evaluating interactive products.

II. Concepts of UX and Usability

In this section, we elaborate on some basic concepts that are important for the general reasoning in this paper.

One important point is the distinction between usability and UX. A well-known definition of usability is provided by ISO 9241-110 [13]. Usability is defined here as "*the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use*".

In contrast, ISO 9241-210 [14] defines UX as "*a person's perceptions and responses that result from the use or anticipated use of a product, system or service*".

There are some fundamental differences between these two concepts. Firstly, usability is focused on tasks or goals of a user. Properties like efficiency, error tolerance, and dependability are commonly known criteria for usability. In contrast, UX is not restricted to achieving specified goals with the help of a product. It covers many more properties, such as beauty of the design, fun of use, or novelty of the design.

Secondly, usability only includes experiences related to the actual use of a system. In contrast, UX also encompasses impressions before (anticipated use) and after using the system (episodic and remembered use) [15].

Thirdly, the usability definition does not state if usability is an objective or subjective concept, but there is indication that usability and UX are differentiable in terms of subjectivity. For example, objective measures, such as the number of failed tasks in a usability test [16] or the average time needed to finish a task (sometimes such averages are also estimated from models, for example GOMS [17]), are commonly used to measure the usability of a product. These metrics are highly product-dependent, i.e., are not necessarily user-specific. In contrast, UX is a purely subjective concept, meaning that UX can only be explored as the individual, unique perspective of one (or more) user(s). Furthermore, as with any experience, the meaning and thus an assessment of UX is conceived in intersubjective exchange and co-understanding between individuals [18].

Let us illustrate this with an example. Assume that we have a new product, and that this product is evaluated in a usability test which covers all the main tasks that can be performed with the product. Now, also assume that we find 80% of the users fail to finish these tasks without assistance. We will most likely conclude that the product has poor usability. Assume now that we hand over a UX questionnaire after the test and the results indicate that the users judge the UX of the product quite high (for example, because most of them attribute their failure not to the poor product design, but to a lack of skills or knowledge). In this case, we would conclude that the product has a good UX, that is, the subjective impressions of the users are good, even though most of them have failed to use the product successfully. Of course, this is a hypothetical example, since in practice, high failure rates concerning the tasks will nearly always result in bad UX ratings in a UX questionnaire.

There are two other conceptualizations of UX which need to be mentioned here. In [19] the product qualities that are related to the UX perception of users are distinguished into two categories. *Usability goals* are related to product qualities with respect to the tasks users must perform with the product. They correspond to the classical dialog principles of the ISO 9241-110 [13]. *User experience goals* relate to the impression of users towards the interaction with the product that are not directly related to the tasks, for example the aesthetic impression of the product.

A nearly identical distinction by using a different terminology is introduced in [20]. *Pragmatic qualities* describe how well users can perform the tasks necessary to reach their goals with the product. *Hedonic qualities* are all other product qualities that are not directly related to tasks.

This distinction into hedonic and pragmatic qualities (respectively *usability goals* and *user experience goals*) has some inherent problems.

The pragmatic qualities share some underlying concept: They are aspects of an interaction related to tasks. The hedonic qualities do not share such a common concept, they are simply defined as all non-task related UX aspects. This immediately raises the question whether they can be further split into some meaningful sub-groups. In addition, it is not so easy for some natural qualities to be defined as pragmatic or hedonic. For example, the quality of content of a web page can be classified as pragmatic (if the user searches for a detailed information on the page a good content quality helps him or her to reach the goal efficiently) or hedonic (if a user just browses the web and finds an interesting page by chance). Thus, for some qualities of interactive products it depends on the concrete usage situation if they are hedonic or pragmatic.

In this paper, we follow the approach to define UX by a set of semantically distinct quality aspects. It has some clear advantages.

As stated above, we understand UX as a purely subjective concept. Thus, we need to ask users when we want to measure their impression of the UX of a product. Concrete UX quality aspects relate the psychological needs of users to properties of the interaction of a user with a product. Therefore, they can be used to formulate questions that are detailed enough to be answered by users. In addition, designers need clearly described qualities to guide them during the design phase and to evaluate their prototypes.

Another advantage is that a clear understanding of the important quality aspects helps UX researchers to decide which UX questionnaire should be used to evaluate a product. Scales of a questionnaire typically map such quality aspects [5]. Another potential application is the definition of new UX questionnaires tailored to the specific needs of a project.

III. UX Quality Aspects

In the following, we describe the UX quality aspects used in this paper. But first we need to define how we conceptualize the term *UX quality aspect*. Here, we follow a definition given in [5]: "*A UX quality aspect describes the subjective impression of users towards a semantically clearly described aspect of product usage or product design.*"

Let us look at some examples to clarify this definition. Learnability (how easy it is to learn how to use a product) is clearly a UX quality aspect. Product design elements, for example, using an easy-to-understand terminology on the user interface, can of course influence this impression. The same is true, for example, for the efficiency of the interaction, dependability of the interaction, fun of use or beauty of the design. Satisfaction with the product price or with the service provided are not considered as UX quality aspects. They are part of the more general concept of customer experience or short CX. These impressions of a user are not influenced by the design of the product itself.

The UEQ+ [11] is a modular framework that allows one to combine predefined UX scales to create a concrete UX questionnaire. Currently, the framework contains 20 UX scales, but they can be extended as needed. In [21], the construction of the clarity scale can be read as an example. It is built on the ideas described in this paper. The descriptions of the following UX quality aspects are oriented towards the descriptions of the corresponding UEQ+ scales [11]. See ueqplus.ueq-research.org or the UEQ+ handbook [12] for more information. A more detailed description of different UX quality aspects, including some UX aspects that are not contained in the list below or in the UEQ+, can be found in [5]. In this paper, we limit the description to UX quality aspects to those aspects that are used in our studies.

The UX quality aspects described below have been extracted by an analysis of existing UX questionnaires and a detailed investigation of UX research literature. Some of the UX quality aspects appear under different labels in research literature. In these cases, alternative names are shown in brackets.

- **Perspicuity (Learnability):** Is it easy to get familiar with the product and to learn how to use it?
- **Efficiency:** Can users solve their tasks without unnecessary effort? Does the product react fast?
- **Dependability (Controllability):** Does the user feel in control of the interaction? Does the product react predictably and consistently to user commands?
- **Usefulness:** Does using the product bring advantages to the user? Does using the product save time and effort?
- **Intuitive use:** Can the product be used immediately without any training or help?
- **Adaptability (Customization):** Can the product be adapted to personal preferences or personal working styles?
- **Novelty (Originality):** Is the design of the product creative? Does it catch the interest of users?
- **Stimulation (Fun of use):** Is it exciting and motivating to use the product? Is it fun to use?
- **Clarity:** Does the user interface of the product look ordered, tidy, and clear?
- **Quality of Content:** Is the information provided by the product always actual and of good quality?
- **Immersion:** Does the user forget time and sink completely into the interaction with the product?
- **Aesthetics (Beauty):** Does the product look beautiful and appealing?
- **Identity:** Does the product help the user to socialize and to present themselves positively to other people?
- **Loyalty:** Do people stick with the product even if there are alternative products for the same task?
- **Trust:** Do users think that their data is in safe hands and not misused to harm them?
- **Value:** Does the product design look professional and of high quality?

Of course, this is not an exhaustive list of all possible UX quality aspects. Nor will there ever be such a list. New products introduce new interaction paradigms and, therefore, new UX quality aspects become important [5]. For example, voice interaction introduces new UX quality aspects that need to be measurable, for example *response behavior* (Does a voice assistant behave respectfully, politely, and in a trustworthy manner?) or *comprehensibility* (Does a voice assistant correctly understand the user's instructions and questions using natural language?) [22].

For most of the aspects mentioned above, corresponding scales are available in the UEQ+ framework [11]. However, the framework also contains several other scales that are not used in the context of our studies and are, therefore, not mentioned in the above list, for example: the scales for voice interaction as well as haptics, and acoustics. See ueqplus.ueq-research.org or the UEQ+ handbook [12] for further and actual information.

It is easy to see that not every UX quality aspect mentioned is relevant in all situations. It depends on demographic factors or simply personal preferences how a specific user judges the importance of a UX factor for a given product. The mean importance rating over all users will also vary considerably between products, that is, the same user will find some of these aspects important for one product, but unimportant for another.

For some of the quality aspects mentioned above, it is also obvious that they are only important for certain products. Of course, *clarity* makes sense for products with a graphical user interface but is pointless for voice interaction. *Identity* describes the user's perception that using a product helps them to create a positive impression and increase their reputation. This only comes into full effect if the user makes the decision to use a product or can significantly contribute to it. In the case of business software, which is used professionally and usually procured by the company, this aspect plays a subordinate role. For a smartphone or usage of certain social platforms *identity* is quite important.

In these examples it is intuitively clear that some UX quality aspects are not important for certain types of products. But this is not always the case. The goal of this paper is to investigate the dependency between product and the importance of certain UX quality aspects.

IV. Research Questions

We address the following research questions in this paper:

- RQ 1: How important are the UX quality aspects described in the previous section for different types of products?
- RQ 2: How can we measure the importance of UX quality aspects for product types in a replicable and stable form?
- RQ 3: How big is the impact of demographic variables and the cultural background on the importance rating of UX quality aspects?
- RQ 4: How accurate is the prediction of the importance rating of a UX quality aspect for a concrete product from the rating of the corresponding product type?

To answer these questions, we summarize the results from some published studies and enrich them with some unpublished new results. In the following we present 5 studies with a total of 361 participants who have submitted nearly 70,000 ratings during the years 2015–2021.

In study 1, we define and test a method to measure the importance of UX aspects for given products or product categories. Study 2 captures the relative importance of a set of well-known UX quality aspects for a larger number of important product categories. Study 3 is a replication of study 2. The goal of this replication is to investigate how stable the results are. Study 4 answers the question if we can infer the relative importance of UX aspects for a concrete product from the corresponding product category. Study 5 replicates study 2 with participants from a different cultural background. The goal of this study is to check if the results are replicable in different countries or cultures.

V. Study 1: How to Measure the Importance of UX Aspects?

A. Introduction

The main goal of this first study was to define a technique to evaluate how important different UX quality aspects were for different types of products.

B. Participants

Participants were recruited over a distribution list. Fifty-one persons took part in the study (35 males, 15 females, 1 gender unknown). The average age was 35 years (minimum 20, maximum 55 years).

C. Method

Each participant was asked to evaluate a browser (Safari, Firefox, Internet Explorer, Chrome, Opera), a text-processing software (Word, Pages, Writer) and a communication tool (WhatsApp, Skype, Facebook, Me) in an online survey. Examples for each type were given according to the products in parentheses. Participants were instructed to skip a product type if they did not use it on a regular basis. Thus, some of the participants evaluated only one or two product categories.

The survey first asked for age, gender, and job title. After this, all the UX aspects described above were presented. Each UX aspect was described by a short text, and the participant should rate on a 7-point scale if the aspect was present in the product and how important this aspect was. Hence, each UX aspect was presented in a block containing two questions.

The following example shows the English translation of the two-items block for the UX aspect *Efficiency* and the product *WhatsApp*. The original study was done in German.

Efficiency	**Fully disagree**						**Fully agree**
I can finish my tasks with WhatsApp with minimal effort. No unnecessary steps are required.	○	○	○	○	○	○	○
Efficiency is important to me for products like WhatsApp	○	○	○	○	○	○	○

The 15 blocks for the UX quality aspects (*Quality of Content* was left out, since it does not make sense for the product types used in the study) were presented below each other.

D. Results

The mean importance ratings over all participants are shown in Fig. 1. To focus mainly on the important UX quality aspects, the corresponding bars are highlighted in color. Accordingly, the irrelevant factors per product category are greyed out.

We also checked the impact of the demographic variables such as age and gender, on the importance ratings. Age seems to have no real impact. The correlations between the age of a participant and the importance rating for the UX aspects are all very small. They are all in the range from -0.20 to +0.17.

Concerning gender, we only found a significant impact (t-test, two-tailed, $p<.05$) for *Novelty* and *Aesthetics*. In both cases the importance ratings from females were a bit higher than those from males. But in general, we can conclude that gender does not have a big impact on perceived importance of UX quality aspects.

The results demonstrate that the ratings for the UX aspects show clear differences. Hence, the method can uncover different levels of importance for the different UX aspects. The impact of the product type is not so clearly visible in this study. Especially, the importance ratings for the pragmatic UX aspects (for example, *Usefulness*, *Dependability*, *Intuitive use*, *Perspicuity* and *Efficiency*) do not differ too much between the three investigated product types.

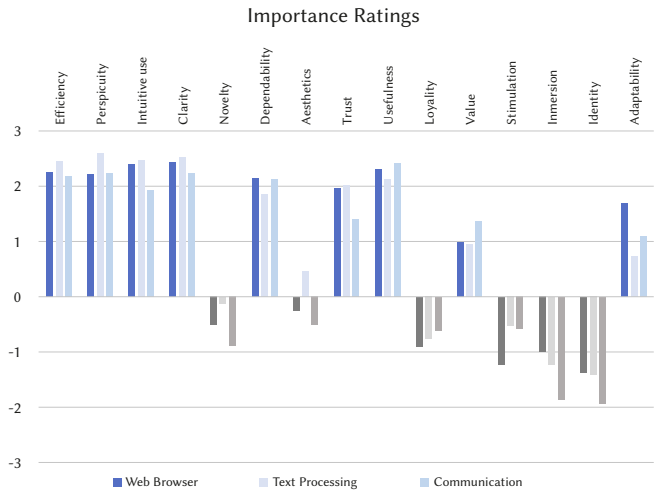


Fig. 1. Mean importance ratings for the UX quality aspects per product category. Scale ranges from -3 to +3.

E. Limitations

The number of participants was small. In addition, since one participant rated several products, it was possible that these ratings were not fully independent. Also, different concrete products were summarized under a category which might have impacted the results.

VI. Study 2: Product Type and Importance of UX Quality Aspects

A. Introduction

The goal of study 2 was to investigate the dependency of product type and the importance of UX quality aspects in a broader context. To avoid the possibility that different products were rated to form a rating for a particular product type, we used abstract product types explained by several examples in this study.

B. Participants

Fifty-eight students from the University of Applied Sciences Emden/Leer were recruited for this study. They received some course credits for participating.

C. Method

The study was conducted in German. The participants were asked to provide their judgements regarding the importance of the UX aspects described above for several software product categories. The UX quality aspects were described by short texts as shown in section III. Each product type was described by a name and several examples:

- *Word processing:* Microsoft Word, Microsoft Power Point, Latex, LibreOffice Writer
- *Spreadsheet:* Microsoft Excel, OpenOffice Calc
- *Messenger:* WhatsApp, Facebook Messenger, SnapChat
- *Social Network:* Facebook, Xing, LinkedIn
- *Video Conferencing:* Skype, Facebook Video Call
- *Web Shops:* Amazon, Conrad, Redcoon, eBay
- *News Portals:* Spiegel.de, Zeit.de, Sueddeutsche.de
- *Booking Systems:* Bahn.de, Lufthansa.de, Booking.com, Hrs.de
- *Info web pages:* Club web site, web site of hometown
- *Learning platforms:* Moodle, Open Elms

- *Programming tools:* Microsoft Visual Studio, Eclipse
- *Image Processing:* Photoshop, Gimp
- *Online Banking:* Online portal of own bank, Starmoney
- *Video Portals:* YouTube, Netflix, Amazon Prime Video
- *Games:* World of Warcraft, Minecraft

Students received a Microsoft Excel list which contained the UX aspects as rows and the product categories as column headers. They were asked to fill out the Excel list and send it back within one week.

Each cell could be filled with the following answer options: *Extremely unimportant* (-3), *Somewhat unimportant* (-2), *Slightly unimportant* (-1), *Neutral* (0), *Slightly important* (1), *Somewhat important* (2), and *Extremely important* (3). Also, the option *Meaningless* could be selected if the UX aspect did not make sense for a product category.

Overall, the Microsoft Excel list consisted of 16 UX aspects and 15 software product categories. The participants had to provide their judgements by filling 240 cells, i.e. making 240 decisions.

D. Results

Let us first look at the mean importance ratings for the UX quality aspects per product type. These data are shown in Fig. 2. The numerical mean values can be found in Table A in the Appendix (German data set, upper values).

As we can see in Fig. 2, the different product types differ clearly in terms of the participants' assessments of the importance of the different UX quality aspects.

Similar to Fig. 1, Fig. 2 is intended to highlight the relevant UX quality aspects. To make cross-category similarities more recognizable, the irrelevant aspects have been lightened.

For related product types (e.g., word processing and spreadsheets, see Fig. 2 first and second row) there are also very similar patterns of importance ratings.

Regarding the inter-individual differences, it can be stated that the observed standard deviations are between 0.46 and 1.99. The average of all standard deviations is 1.32. This means that there are large differences in the assessment by the test participants, and the size of the standard deviation also depends on the product type and UX quality aspect.

E. Limitations

The number of participants is relatively small. They are all German students and, therefore, quite homogeneous as far as demographic factors are concerned. It is therefore questionable if the results can be generalized to other user groups. In addition, only abstract product categories are used. Thus, it must be clarified if results of a particular product category can be used to predict the ratings for concrete products from this category.

VII. Study 3: Checking Stability of the Results of Study 2

A. Introduction

The goal of study 3 was to check how stable the results of study 2 were. So, the study 2 was replicated one year later with a new cohort of students. To keep the effort for the students somewhat lower, only 10 of the former 15 product categories were used. A more detailed description of the study is given in [10].

B. Participants

Sixty-three students from a German university were recruited for the study. They received some course credits for taking part in the study.

C. Method

The study was also conducted in German. The method was completely identical to the method used in the second study.

The product categories *Spreadsheet*, *Video Conferencing*, *Image Processing*, *Info Web Pages*, and *Games* were not used in this study. Thus, we only had 10 product types but the full list of UX aspects.

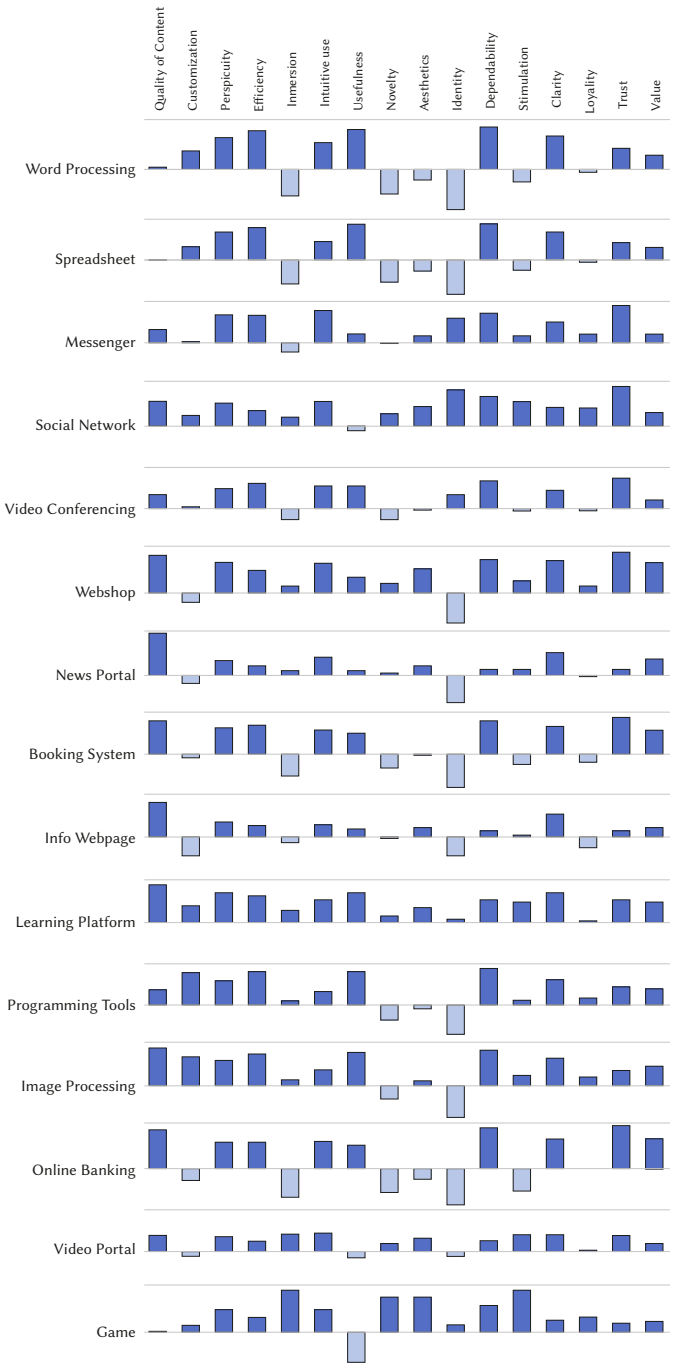


Fig. 2. Means of the importance ratings for the UX quality aspects per product category. Scale ranges from -3 to +3.

D. Results

We were able to compare the 160 mean importance ratings (10 product categories and 16 UX aspects) from both studies. 160 t-Tests (equal variances assumed, $p < .05$) were performed to compare the observed rating in study 2 with the corresponding rating in study

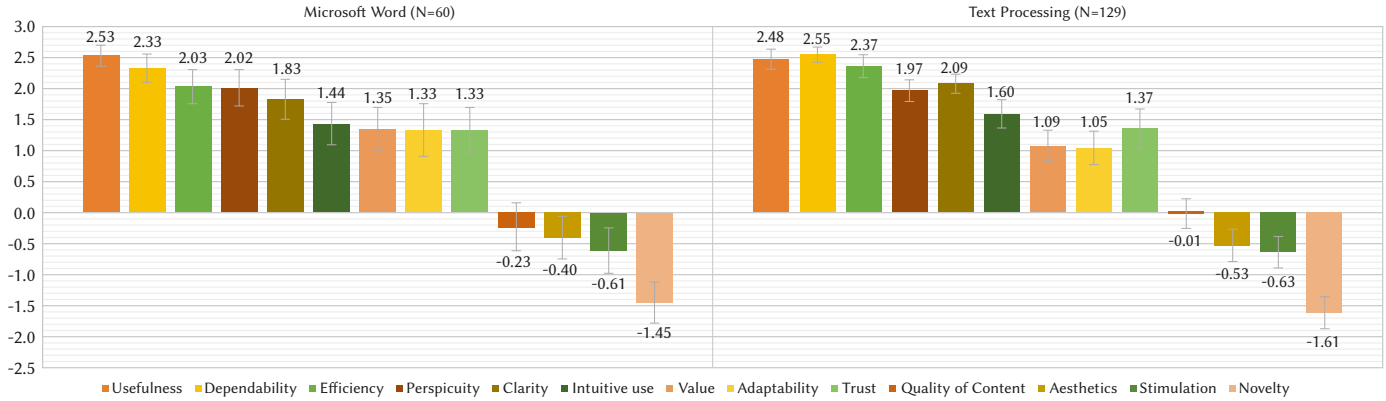


Fig. 3. Importance ratings for the product MS Word and the product type Word Processing.

3. For 154 of these 160 tests the null hypothesis that there were no differences could not be rejected. Only in 6 cases could a statistically significant difference be detected which could be expected with 160 tests and an error probability of .05. Thus, the ratings obtained in study 2 could be reproduced very well.

In addition, we calculated the rank correlations per product category between the two studies. We first transferred the mean importance ratings per product category to ranks from 1 (highest) to 16 (lowest). Then we correlated the ranks for the two studies per product category. The results are shown in Table I.

TABLE I. Rank Correlations Between the Importance Ratings From Study 2 and Study 3

Product category	Rank correlation
Text Processing	0.99
Messengers	0.94
Social Networks	0.88
Webshops	0.97
News Portals	0.94
Booking Systems	0.97
Learning Platforms	0.98
Development Tools	0.98
Online banking	0.99
Video Portals	0.92

The correlations are all extremely high. Thus, the relative importance of a UX aspect for a product category seems to be nearly identical for both data sets. More practically speaking, if we choose, for example, the five most important UX aspects for a product category, it will not make much of a difference if we use the results from study 2 or study 3 for our selection.

In conclusion, the results from study 2 could be reproduced very well. The importance ratings of the UX aspects seemed to be very stable, which provided a positive answer to our research question 2.

E. Limitations

The study showed that the results from study 2 could be reproduced very well. The replication study was conducted with the same target group: German students. Thus, this study does not allow one to conclude that the results also apply to groups with different demographic parameters or from different cultural backgrounds.

VIII. Study 4: Importance of UX Aspects For Specific Products

A. Introduction

In studies 2 and 3, the participants judged the importance of UX quality aspects for product categories. These categories were described by a category name and several examples. In practical projects, however, we are interested to use this knowledge for specific products. Thus, it is very important for us to infer the importance of UX aspects for such a concrete product from the corresponding product category. Study 4 investigated if this is possible. A more in-depth discussion of the study can be found in [9].

B. Participants

Sixty-two master's students in the course User Experience at University of Applied Sciences Emden/Leer participated in this survey. 36 indicated that they were female, 21 identified as male, and 5 did not specify. The average age was 29.6 years, with the youngest participant reporting their age as 23 and the oldest as 48.

C. Method

The method was nearly identical to the previous studies. However, only a subset of the UX aspects was used: *Quality of Content, Adaptability, Perspicuity, Efficiency, Intuitive Use, Usefulness, Novelty, Visual Aesthetics, Dependability, Stimulation, Clarity, Trust,* and *Value.*

Instead of product categories the participants rated the importance of these aspects for specific products (*Google Maps*, *Microsoft Word*, *WhatsApp*, *Instagram*, *Microsoft Teams*, *Discord*, *Trello*, *Zalando.de*, *Tagesschau.de*, *Netflix*, *Spotify*, and *YouTube*). Participants were instructed to rate only products they used and, therefore, not every participant rated every product.

TABLE II. Assignment of Products and Product Categories

Software Product	Product Category
Discord	Video Conference Tools & Messengers
Microsoft Teams	Video Conference Tools & Messengers
WhatsApp	Messengers
Netflix	Video Portals
YouTube	Video Portals
Instagram	Social Networks
Microsoft Word	Text Processing
Tagesschau.de	News Portals
Zalando.de	Web Shops

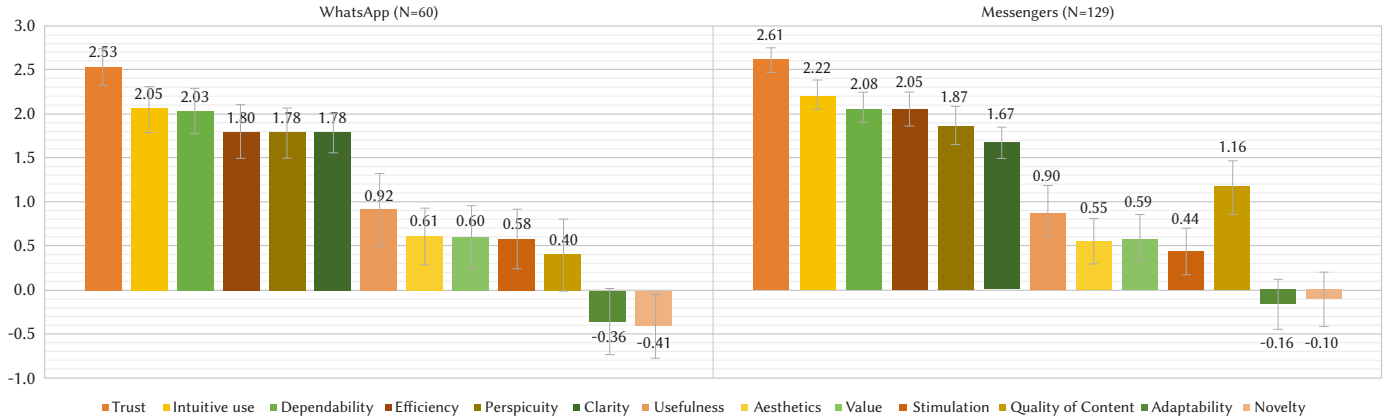


Fig. 4. Importance ratings for the product WhatsApp and the product type Messengers.

Products and product categories were assigned as mentioned in Table II. Notice that both *Discord* and *Microsoft Teams* were assigned to the categories V*ideo Conference Tools* and *Messengers*, which were assumed to cover their main use cases.

D. Results

An overview of all the importance ratings used in the analysis is given in Table C in the Appendix. A detailed description of all the results from this study can be found in two research reports that are available online [23], [24] and in-depth analysis is given in [9]. Exemplary insights into rankings and rank correlations are given in the following.

The importance ratings for the products correspond in most cases quite well to the ratings of the corresponding categories (see examples in Fig. 3 and Fig. 4). The latter are calculated by combining the results of study 2 and study 3; so, they come from independent studies. Hence, we can answer our research question RQ 4 positively: It is possible to use the product category to infer the importance of UX aspects for a concrete product. In the following two examples, we show the comparison of two products to the corresponding categories.

The correspondence between the ratings of the products and the corresponding ratings for the product categories becomes even more evident when we look at the rank orders of the importance ratings for product categories and products.

Table III shows the rank correlations for the importance ratings for the products and the corresponding categories. These correlations are all extremely high. Thus, if a UX aspect is judged as important for a product category, this is also the case for products of this category. Therefore, we can infer the importance of a UX aspect for a single product from the values obtained for the corresponding category.

TABLE III. Rank Correlations for the Importance Ratings of Products and Corresponding Product Categories

Software Product	Product Category	Rank Corr.
Discord	Video Conference Tools	0.88
Discord	Messengers	0.75
Microsoft Teams	Video Conference Tools	0.94
Microsoft Teams	Messengers	0.78
WhatsApp	Messengers	0.93
Netflix	Video Portals	0.86
YouTube	Video Portals	0.83
Instagram	Social Networks	0.64
Microsoft Word	Text Processing	0.97
Tagesschau.de	News Portals	0.96
Zalando.de	Web Shops	0.60

E. Limitations

In this study, only a subset of the product categories and UX aspects from the previous studies was used. Of course, it makes sense to repeat this study for all the categories and UX aspects that are not considered in this study. In addition, if we look at the correlations in Table 3, we see that they vary between 0.97 for Microsoft Word & Text Processing and 0.6 for Zalando.de & Web Shops. Although 0.6 is a very high correlation, it would be an interesting follow-up study to find out why the correspondence is higher for some combinations of product categories and concrete products than for others.

IX. Study 5: A Replication in a Different Culture

A. Introduction

In the previous studies, we investigated how the product type influenced the importance rating of different UX quality aspects. The participants of these studies were German students. This immediately raises the question if the results can be replicated in other countries, that is, with participants that have a different cultural background.

Several papers show a cultural influence on the concrete elements of a user interface design. In [25] the cross-cultural use of computing metaphors is investigated, and they are often deeply rooted in culture. Such metaphors are, for example, the basis for icon design; and clearly, icons based on a metaphor not known in the culture of the users are very difficult for them to understand. Other papers deal with the cultural use of colors [26].

Design teams are usually small and often quite homogeneous in terms of cultural background. Is the design created from such a team acceptable in all cultures? Several research papers deal with this question, which is of high practical relevance, for example [27], [28], [29].

On a more abstract level, the impact of culture on UX was investigated in a number of papers. [30] showed that users performed better if the user interface was designed to match their cultural profile (in the sense of Hofstede's model [31]). [32] found that users' cultural profile impacted their acceptance of specific technologies, and [33] demonstrated that the perceived usability of a web site was higher if it was originally designed in the users' native language.

But there are also papers that doubt the influence of culture on UX. In [34] it was argued that the goals of users when they used a product or web site were the main influence on UX. Due to the increasing globalization people get used to products designed by designers from a different cultural background. Thus, typical interaction patterns become more and more important, and the impact of cultural background should decrease over time [35].

B. How to Define Culture?

Intuitively, we all believe to understand what culture, or the cultural background of a person means. However, it is not easy to give culture a clear and scientific definition. Several theories try to explain differences between cultural groups over sets of cultural dimensions [36]–[40]. The most popular of these theories is the model of cultural dimensions by Hofstede [31], which is based on extensive empirical data. In this paper, we rely on Hofstede's model.

Hofstede assumes that culture is a set of learned traits. These traits make certain behaviors or reactions towards specific situations occur more often in some cultures than in others. This model contains six distinct cultural dimensions (adapted from [31]):

- *Power distance:* Level of acceptance of an unequal power distribution in a country.
- *Individualism vs. Collectivism:* Extent to which members of a culture prioritize their individual goals over the goals of the group.
- *Masculine vs. Feminine:* A masculine culture is mainly driven by competition, while in a feminine culture cooperation and caring for others are the more important values.
- *Uncertainty avoidance:* Desire to accept or avoid uncertain situations.
- *Long-term orientation:* If planning and action are based more on long- or short-term goals.
- *Indulgence vs. Restraint:* Extent to which people try to control their desires and impulses.

The strength of Hofstede's model is that it provides concrete scores for these dimensions in several cultures. According to these scores, Germany and Indonesia are quite different [41]. Indonesia is described as strongly collectivistic and shows a high level of power distance. In contrast, Germany is described as highly individualistic and shows a relatively low level of power distance. For the dimension *Indulgence vs. Restraint*, both countries show nearly the same value. The other dimensions scores for Germany are moderately higher than those for Indonesia. Due to these differences Germany and Indonesia are good candidates for investigation if different cultures cause a different importance rating for UX quality aspects.

Of course, there are also other models of culture [42]–[44] and there is also some critique that the Hofstede model is too stereotypical [45]. But this model is clearly the best investigated cultural model with respect to usability and UX. See, for example, the studies on the connection of the Hofstede dimensions to user interface design elements of web sites [28], [45]. In [39] it was shown that the cultural dimensions defined by Hofstede [31] had dramatic correlations to the development of e-government in countries.

C. Participants

The study was conducted at a large Indonesian university. The 114 participants (average age 21.34 years, 64 males, 50 females) enrolled in a human-computer interaction course and got some credit points for their participation.

D. Method

The study was planned as an exact replication of study 2. Thus, the experimental procedure was completely identical. Of course, all texts were translated carefully into Indonesian language. Some of the examples for the product types were unfamiliar to the Indonesian participants and, therefore, had to be changed. Otherwise, the procedure was exactly as described in study 2.

E. Results

Fig. 5 shows the mean importance ratings for the Indonesian (green bars) and the German sample (blue bars). Just as in Fig. 2, the irrelevant UX quality aspects have been lightened to put the focus on cross-category similarities. The exact values can be found in Table A in the Appendix.

Again, each product type has a typical pattern of importance ratings. In addition, similar product types (see Fig. 2, for example the productivity-oriented tools such as *Word Processing*, *Spreadsheet*, *Programming Tools* and *Image Processing* or the communication tools like *Messenger* and *Video Conferencing*) show quite similar patterns.

If we compare the importance ratings, we see some differences in their values. Especially, the hedonic UX quality aspects are rated a bit higher in the Indonesian sample. If we compare the mean importance ratings per product type and the UX quality aspect, we see that in 156 out of 240 cases the ratings differ significantly (t-test, $p < .05$, two-tailed).

But we cannot simply infer cultural differences from this simple comparison of mean importance ratings. There is an overall answer tendency that must be considered. The average rating over all UX quality aspects and product types is 4.85 for the German and 5.51 for the Indonesian sample. Thus, Indonesian participants in general use higher ratings than German participants do.

If we look only at the relative importance of the UX quality aspects, we see that the judgements in both samples are quite similar. This is also confirmed by the very high correlations between importance ratings for both groups (see last column in Table A in the Appendix).

To show this in greater detail, we transformed the mean importance ratings to ranks. Rank 1 is assigned to the UX quality aspect with highest mean importance rating for a product type, rank 2 to the UX quality aspect with the second highest mean importance rating, and so on. Table B in the Appendix shows these ranks. If we compare the ranks for both samples per product type, we see that they are quite similar. This is also confirmed by the correlations in the last column of Table B.

How big is the impact of culture compared to the impact of individual differences between persons in one cultural group? To answer our research question RQ 3, an analysis of variance was performed. For each combination of product type and UX quality aspects, the total variance VAR over the complete data set that included the German and Indonesian participants was calculated. The variance explained by the two cultural groups $VAR(G,I)$ was then calculated by the formula:

$$VAR(G,I) = \left(n_G * \sum_{1}^{n_G} (\bar{x}_G - \bar{x})^2 + n_I * \sum_{1}^{n_I} (\bar{x}_I - \bar{x})^2 \right) \Big/ (n_G + n_I)$$

where n_G, n_I are the sample sizes for the German and Indonesian groups of students, and $\bar{x}_G$, $\bar{x}_{GI}$, and $\bar{x}$ are the mean values in both samples and the complete data set. This is the variance we would expect if all persons in one cultural group show the same importance rating, that is, if the importance ratings are completely dependent on culture.

The value $\frac{VAR(G,I)}{VAR}$ can be interpreted as the relative amount of variance explained by the two groups compared to the total variance. The results show that that the proportion of variance explained by culture is very small compared to the impact of individual preferences (for detailed results, see Table V in [46]). Thus, the cultural background of the users only seems to play a minor role concerning the importance of UX aspects for the overall UX impression for products. The product type and the main usage scenarios for a product are the main factors here.

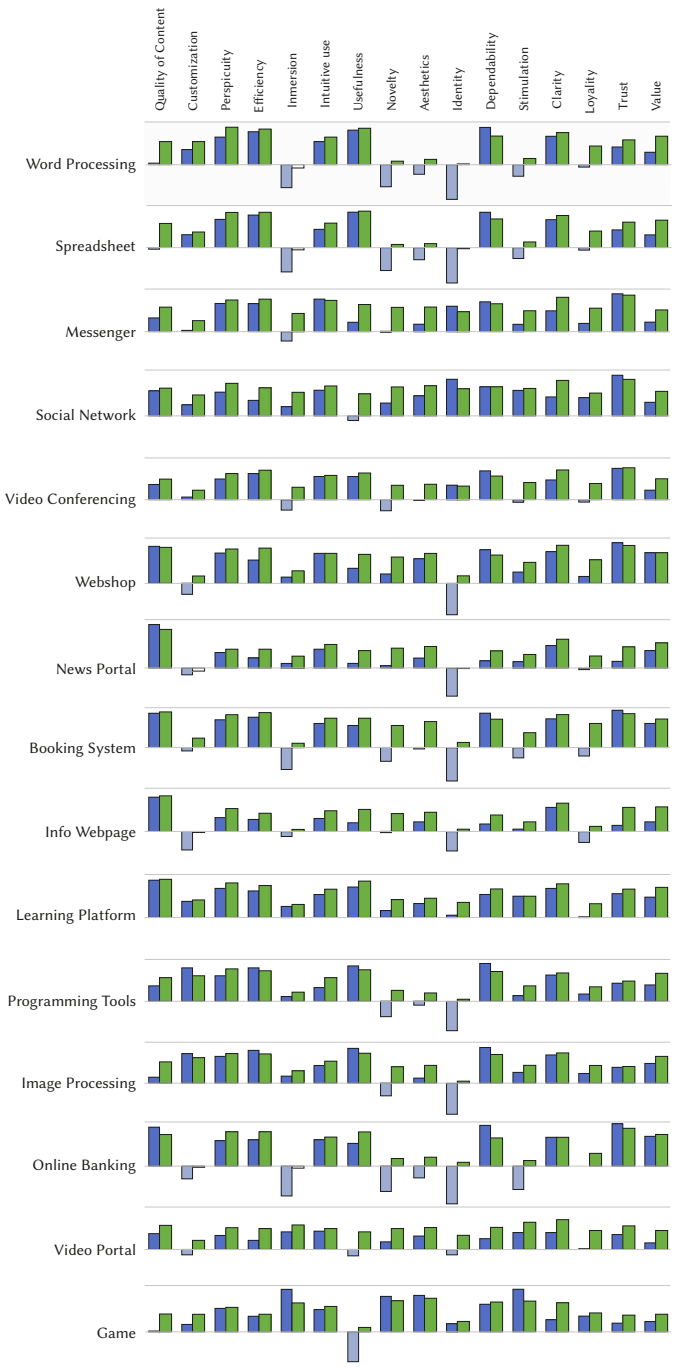


Fig. 5. Means of the importance ratings for the UX quality aspects per product category. The values for the German sample are shown as blue bars and those for the Indonesian sample as green bars.

F. Limitations

Our results indicate that the impact of the cultural background of a person on the importance rating of UX quality aspects is relatively small. Of course, we compared data from only two different countries. German and Indonesian cultural values differ quite substantially according to an established model of cultural differences. If there are no substantial differences between such different cultures, then it is very likely that this also applies to other cases. But similar studies in other countries are required to check if the findings can really be generalized.

X. What Are the Most Important UX Aspects for a Project?

In our studies it has shown that the importance of UX aspects for the overall UX impression of users differs with the product category. We have also presented a simple rating method that can be used to get valid data which are replicable and do not depend too much on demographic data or cultural background of users.

To summarize the illustrations and tables that have been provided, especially the complete overviews in Tables A and B in the Appendix, important UX quality aspects have clearly been assigned to certain product types.

Accordingly, the following most important UX aspects should be considered when launching or evaluating the corresponding product types:

- **Word Processing, Spreadsheet, Programming Tool, and Image Processing:** The most important aspects, as shown in all studies, are *Usefulness* and *Dependability*. In addition, UX quality aspects such as *Efficiency*, *Perspicuity* and *Clarity* are also important for these categories.
- **Booking System and Online Banking**: For these two categories, the aspects *Trust* and *Dependability* are most important. As can be seen in Table B in the Appendix, the aspects *Quality of Content* and *Clarity* are not to be neglected either.
- **WebShop, Messenger, Social Network, and Video Conferencing Tool**: While *Trust* is most important for these four categories, *Dependability* and *Intuitive use* should also be considered.
- **Learning Platform, News Portal, Info Web Page**: Clearly, *Quality of Content* is the most important in these three categories. But the aspects *Clarity*, *Intuitive use* and *Perspicuity* should also be considered.
- **Video Portal**: Only in this category, the most important aspect is *Intuitive use*. But *Immersion*, *Stimulation* and *Clarity* are also scoring highly.
- **Game**: The UX quality aspects *Immersion* and *Stimulation* should be considered most important for this category.

The presented data can be used to determine what the most important aspects are for the design of a new product or evaluating an existing one. Table B shows the importance ranks for the UX aspects and can be used to determine which UX aspects are likely to be the most important ones for the overall UX judgement for a product from this category.

A quite simple method to pick which aspects to focus on for a product would be to order the corresponding UX aspects by their ratings and then to consider the top 5 (or any other number). However, in practice things are typically not so easy. Our data shows what is most likely important for the users of a product. But in real design projects other interests must be considered as well, that is, what is most important for the users of a product may not completely cover the design goals of a project.

Let us illustrate this with a simple example. Assume that a new user interface for a programming environment should be designed and implemented. If we look at Table A in the Appendix, we see that the most important aspects are *Dependability*, *Usefulness*, *Efficiency*, *Customization* and *Clarity* (it is not surprising that these are the tools used in a working context). *Visual aesthetics* or *Beauty* of the design does not play a role for users (this aspect is ranked on position 12 of 16). However, aesthetics of the design may be highly relevant to

the marketing of the new user interface. Hence, it may be a highly relevant UX aspect, even though users report not caring about it.

Thus, our results will help designers and product owners to identify which UX aspects are important for the real users, but it is not wise to rely solely on this data when it comes to executing real projects.

XI. Summary of the Results

In this section we will summarize the results of our studies and answer the research questions described above.

A. RQ 1: Importance of UX Aspects

One of the goals of the research was to find out how important the different UX quality aspects were for different product types.

As we described in the last section, our results could help designers and product owners to get an idea about which UX aspects they should focus on during designing and for the later evaluation of the finished product. Here, the obtained ratings from our studies can be used to determine the most relevant aspects for a product from a given product category.

B. RQ 2: Measurement Method

We showed in several studies how the importance of UX quality aspects for product types or products could be measured by a simple rating mechanism. The results proved to be replicable and stable. The method can be used to get ratings for product categories and for concrete products.

C. RQ 3: Impact of Demographic Variables or Cultural Background

For the practical applicability of our findings concerning the importance of UX aspects for the overall UX impression, it was important to clarify if they were greatly influenced by the cultural background of the users. This is especially essential for products that are developed for international markets.

In study 5 we compared the importance ratings obtained from persons living in countries with a quite different culture according to the cultural dimension of Hofstede. The results indicate that the impact of culture on the importance of UX aspects for the different product categories is small. The type of the product defines which UX quality aspects are important for a concrete product and the cultural background only seems to have a limited impact here.

D. RQ 4: Prediction From the Product Category

Another important question for the practical applicability of our findings was if it is possible to predict the importance rating of a UX quality aspect for a concrete product from the rating of this aspect in the corresponding product type.

As the results of study 4 show, this seems to be the case. The ratings obtained for concrete products are very similar to those obtained for the corresponding product categories. The rank correlations between the importance ratings for the product categories and specific products from these categories are extremely high. Hence, if a UX aspect is important for a product category it is also important for concrete products from that category and vice versa.

XII. Practical Implications

In this section, we will focus on the practical implications of our work. How can our results be applied by UX designers and UX researchers?

The design process of a product is a complex series of detailed design decisions. For a designer, it is beneficial to know which aspects of a design are important for the potential users of the product and which UX quality aspects are less important or completely irrelevant. In the design process, this helps to concentrate on those decisions that have an impact on the important UX quality aspects. Typically, a larger team of designers, developers or product owners collaborate in a design process. To streamline the discussion, it is crucial that all members of the team share the same understanding about the UX quality aspects that are important for the users.

Changes to an existing design or the introduction of new features to an existing product often have positive and negative effects on various UX quality aspects. If it is not clear how important these UX quality aspects are, this can lead to endless discussions.

For example, assume that a new feature should be added to a product that makes data entry much more efficient but adds conceptual complexity and therefore makes it more difficult to understand how to operate the product. If the product is used often during a workday, efficiency is of great importance to users. For such tools, typically, some short learning phase will be required and therefore, perspicuity will not be equally important. Thus, introducing the new feature is in this case a good idea. If, on the other hand, the product is a rarely used self-service (for example for requesting holidays), then things are different. Due to the rare product usage, efficiency will be not very important, but the tool must be usable intuitively since the user may not remember how to operate it. Therefore, in this case the introduction of the new feature would be a bad idea.

Of course, things are trivial in this example. For other UX quality aspects it is ultimately not clear how important they are for a particular product. Our results provide guidance to UX designers about the importance of UX quality aspects in different situations. If the new product belongs to one of our investigated product types, the results shown above can be used to determine the UX quality aspects that are most likely important for the product.

Our results are also helpful when evaluating already existing products: Questionnaires are a popular method to do so. They allow collecting data of larger target groups with little effort. In addition, standard questionnaires allow one to compare different products or product versions by the measured scale values. But which questionnaire should be used for such an evaluation? A large number of UX standard questionnaires are available [2]. All of them offer a different combination of UX scales that represent different UX quality aspects. Being interested in evaluating how our users like the UX with the product, we should measure exactly the UX quality aspects that users consider important to the overall quality of the UX. The results reported in this paper can help UX researchers to select the best questionnaire (or a combination of questionnaires) for their product evaluation.

This knowledge about the importance of UX quality aspects for a certain product is also required to select the right scales in modular approaches, for example the UEQ+ [11]. The use of UEQ+ requires the researcher to choose the most relevant scales out of a catalogue of the currently 20 available scales in the framework. To keep the length of the questionnaire within a reasonable limit it is recommended to use not more than six scales in a single product evaluation. Thus, the knowledge of how important the underlying UX quality aspects are for the overall UX impression is crucial to making a good selection.

Appendix

TABLE A. Means of the Importance Ratings for the UX Quality Aspects for the 15 Product Categories and the German and Indonesian Samples. The Last Column Contains the Correlation Between the Importance Ratings for the German and Indonesian Samples. The Scale Ranges From-3 to+3

Category	Sample	Perspicuity	Efficiency	Dependability	Intuitive use	Usefulness	Adaptability	Clarity	Novelty	Aesthetics	Identity	Stimulation	Immersion	Value	Loyalty	Trust	Quality of Content	Correlation
Word Processing	**German**	1.91	2.31	2.60	1.60	2.45	1.07	2.00	-1.54	-0.67	-2.48	-0.76	-1.63	0.84	-0.16	1.26	0.10	**0.94**
	Indonesian	2.62	2.48	2.03	1.91	2.55	1.24	2.27	0.22	0.35	0.02	0.42	-0.25	1.98	1.30	1.75	1.64	
Spreadsheet	**German**	2.00	2.33	2.53	1.29	2.53	0.91	1.98	-1.61	-0.81	-2.52	-0.76	-1.75	0.88	-0.16	1.23	-0.05	**0.95**
	Indonesian	2.51	2.54	2.02	1.75	2.58	1.12	2.27	0.21	0.28	-0.05	0.39	-0.13	1.96	1.15	1.80	1.67	
Programming Tool	**German**	1.83	2.41	2.69	1.00	2.52	2.38	1.86	-1.09	-0.26	-2.11	0.38	0.35	1.21	0.50	1.31	1.11	**0.92**
	Indonesian	2.34	2.23	2.14	1.66	2.25	1.82	2.02	0.79	0.58	0.15	1.13	0.65	1.99	1.04	1.44	1.70	
Image Processing	**German**	1.79	2.22	2.43	1.14	2.34	2.00	1.88	-0.89	0.31	-2.22	0.70	0.41	1.33	0.63	1.02	0.36	**0.91**
	Indonesian	1.96	1.98	1.93	1.46	2.03	1.72	2.06	1.06	1.19	0.09	1.14	0.79	1.79	1.19	1.12	1.45	
Booking System	**German**	1.97	2.14	2.47	1.74	1.55	-0.24	2.03	-1.00	-0.07	-2.44	-0.78	-1.62	1.71	-0.60	2.70	2.44	**0.89**
	Indonesian	2.35	2.51	2.06	2.13	2.12	0.67	2.33	1.61	1.84	0.33	1.06	0.28	2.04	1.75	2.40	2.53	
Online Banking	**German**	1.79	1.84	2.83	1.83	1.59	-0.86	2.02	-1.71	-0.79	-2.62	-1.61	-2.02	2.10	0.00	2.98	2.68	**0.94**
	Indonesian	2.36	2.38	1.98	2.03	2.38	-0.07	2.03	0.50	0.61	0.26	0.36	-0.13	2.22	0.88	2.65	2.24	
Web-Shop	**German**	1.97	1.47	2.17	1.93	0.96	-0.69	2.07	0.60	1.59	-2.00	0.70	0.40	1.98	0.45	2.62	2.42	**0.91**
	Indonesian	2.24	2.29	1.82	1.92	1.86	0.47	2.44	1.72	1.94	0.50	1.33	0.86	2.00	1.56	2.46	2.36	
Messenger	**German**	1.95	1.98	2.09	2.28	0.60	0.03	1.45	-0.02	0.50	1.75	0.51	-0.65	0.60	0.58	2.67	0.93	**0.76**
	Indonesian	2.22	2.25	1.98	2.19	1.88	0.77	2.37	1.71	1.71	1.38	1.46	1.26	1.51	1.61	2.58	1.67	
Social Network	**German**	1.55	1.00	1.91	1.66	-0.27	0.69	1.26	0.84	1.31	2.39	1.65	0.58	0.90	1.19	2.66	1.67	**0.65**
	Indonesian	2.14	1.83	1.89	1.95	1.39	1.38	2.28	1.91	1.98	1.78	1.77	1.57	1.61	1.48	2.39	1.84	
Video Conferencing	**German**	1.48	1.86	2.07	1.67	1.64	0.16	1.36	-0.78	-0.05	1.04	-0.14	-0.76	0.66	-0.13	2.24	1.05	**0.83**
	Indonesian	1.83	2.12	1.68	1.74	1.91	0.65	2.06	1.03	1.10	0.92	1.19	0.88	1.50	1.11	2.19	1.44	
Learning Platforms	**German**	1.95	1.78	1.53	1.55	2.05	1.05	1.97	0.38	0.90	0.08	1.43	0.70	1.36	-0.04	1.56	2.53	**0.84**
	Indonesian	2.35	2.12	1.91	1.89	2.44	1.12	2.26	1.13	1.25	0.99	1.39	0.86	2.04	0.90	1.91	2.56	
Video Portal	**German**	1.00	0.65	0.72	1.28	-0.43	-0.33	1.16	0.53	0.93	-0.33	1.14	1.22	0.50	0.02	1.07	1.12	**0.84**
	Indonesian	1.57	1.46	1.54	1.47	1.22	0.63	2.11	1.50	1.54	0.99	1.92	1.74	1.38	1.33	1.68	1.75	
News Portal	**German**	0.98	0.66	0.43	1.23	0.30	-0.46	1.50	0.12	0.66	-1.85	0.42	0.27	1.14	-0.07	0.39	2.88	**0.91**
	Indonesian	1.26	1.26	1.09	1.51	1.14	-0.22	1.90	1.29	1.39	-0.01	0.85	0.74	1.68	0.76	1.38	2.53	
Info Web Page	**German**	0.97	0.79	0.41	0.86	0.56	-1.33	1.60	-0.07	0.64	-1.37	0.11	-0.36	0.60	-0.75	0.35	2.41	**0.93**
	Indonesian	1.57	1.26	1.15	1.45	1.50	-0.08	1.96	1.23	1.31	0.12	0.61	0.08	1.71	0.31	1.61	2.45	
Game	**German**	1.53	1.02	1.82	1.48	-1.94	0.47	0.81	2.33	2.38	0.50	2.84	2.84	0.67	1.06	0.56	0.04	**0.89**
	Indonesian	1.64	1.14	1.96	1.67	0.22	1.17	1.95	2.03	2.22	0.70	2.03	1.92	1.15	1.30	1.08	1.15	

TABLE B. Ranks of the Importance Ratings for the UX Quality Aspects for the 15 Product Categories and the German and Indonesian Samples. The Last Column Contains the Rank Correlation for the German and Indonesian Samples. The Five Most Important UX Aspects Are Shown in Bold Font

Category	Sample	Perspicuity	Efficiency	Dependability	Intuitive use	Usefulness	Adaptability	Clarity	Novelty	Aesthetics	Identity	Stimulation	Immersion	Value	Loyalty	Trust	Quality of Content	Correlation
Word Processing	**German**	**5**	**3**	**1**	6	**2**	8	**4**	14	12	16	13	15	9	11	7	10	**0,91**
	Indonesian	**1**	**3**	**5**	7	**2**	11	**4**	14	13	15	12	16	6	10	8	9	
Spreadsheet	**German**	**4**	**3**	**1**	6	**1**	8	**5**	14	13	16	12	15	9	11	7	10	**0,94**
	Indonesian	**3**	**2**	**5**	8	**1**	11	**4**	14	13	15	12	16	6	10	7	9	
Programming Tool	**German**	6	**3**	**1**	10	**2**	**4**	**5**	15	14	16	12	13	8	11	7	9	**0,90**
	Indonesian	**1**	**3**	**4**	9	**2**	7	**5**	13	15	16	11	14	6	12	10	8	
Image Processing	**German**	6	**3**	**1**	8	**2**	**4**	**5**	15	14	16	10	12	7	11	9	13	**0,85**
	Indonesian	**4**	**3**	**5**	8	**2**	7	**1**	14	11	16	12	15	6	10	13	9	
Booking System	**German**	6	**4**	**2**	7	9	11	**5**	14	10	16	13	15	8	12	**1**	**3**	**0,89**
	Indonesian	**4**	**2**	8	6	7	14	**5**	12	10	15	13	16	9	11	**3**	**1**	
Online Banking	**German**	8	6	**2**	7	9	12	**5**	14	11	16	13	15	**4**	10	**1**	**3**	**0,76**
	Indonesian	**4**	**2**	9	7	**2**	15	8	12	11	14	13	16	6	10	**1**	**5**	
Webshop	**German**	6	9	**3**	7	10	15	**4**	12	8	16	11	14	**5**	13	**1**	**2**	**0,86**
	Indonesian	**5**	**4**	10	8	9	16	**2**	11	7	15	13	14	6	12	**1**	**3**	
Messenger	**German**	**5**	**4**	**3**	**2**	10	14	7	15	13	6	12	16	9	11	**1**	8	**0,71**
	Indonesian	**4**	**3**	6	**5**	7	16	**2**	9	8	14	13	15	12	11	**1**	10	
Social Network	**German**	7	11	**3**	**5**	16	14	9	13	8	**2**	6	15	12	10	**1**	**4**	**0,59**
	Indonesian	**3**	9	7	**5**	15	16	**2**	6	**4**	10	11	13	12	14	**1**	8	
Video Conferencing	**German**	6	**3**	**2**	**4**	**5**	11	7	16	12	9	14	15	10	13	**1**	8	**0,81**
	Indonesian	**5**	**2**	7	6	**4**	16	**3**	13	12	14	10	15	8	11	**1**	9	
Learning Platform	**German**	**4**	**5**	8	7	**2**	11	**3**	14	12	15	9	13	10	16	6	**1**	**0,93**
	Indonesian	**3**	**5**	7	9	**2**	13	**4**	12	11	14	10	16	6	15	7	**1**	
Video Portal	**German**	7	10	9	**1**	16	14	**3**	11	8	15	**4**	**2**	12	13	6	5	**0,83**
	Indonesian	6	11	7	10	14	16	**1**	9	7	15	**2**	**4**	12	13	5	3	
News Portal	**German**	**5**	6	8	**3**	11	15	**2**	13	6	16	9	12	**4**	14	10	**1**	**0,85**
	Indonesian	9	8	11	**4**	10	16	**2**	7	**5**	15	12	14	**3**	13	6	**1**	
Info Web Page	**German**	**3**	**5**	9	**4**	8	15	**2**	12	6	16	11	13	7	14	10	**1**	**0,84**
	Indonesian	**5**	9	11	7	6	16	**2**	10	8	14	12	15	**3**	13	**4**	**1**	
Game	**German**	6	9	**5**	7	16	14	10	**4**	**3**	13	**1**	**1**	11	8	12	15	**0,82**
	Indonesian	8	13	**4**	7	16	10	**5**	**2**	**1**	15	**3**	6	12	9	14	11	

TABLE C. Average Importance Ratings of UX Aspects for Products and Categories. Scale Ranges From -3 to +3.

Product Category	Quality of Content	Adaptability	Perspicuity	Efficiency	Intuitive use	Usefulness	Novelty	Aesthetics	Dependability	Stimulation	Clarity	Trust	Value
Video Conf. Tool	1.07	-0.06	1.39	1.89	1.58	1.61	-0.78	-0.06	2.05	-0.05	1.32	2.17	0.68
Discord	0.36	1.09	1.57	2.04	1.43	1.57	0.26	0.78	1.91	0.57	1.48	2.17	1.09
Microsoft Teams	1.08	1.03	1.83	2.07	1.43	2.10	-0.77	0.50	1.87	-0.14	1.53	2.20	1.20
Messenger	1.16	-0.16	1.87	2.05	2.22	0.90	-0.10	0.55	2.08	0.44	1.67	2.61	0.59
WhatsApp	0.40	-0.36	1.78	1.80	2.05	0.92	-0.41	0.61	2.03	0.58	1.78	2.53	0.60
Video Portal	1.10	-0.04	1.07	0.64	1.51	-0.60	0.52	1.09	1.01	1.48	1.27	1.00	0.58
Netflix	1.75	1.15	1.58	1.07	1.82	-1.47	0.95	1.38	1.22	2.07	1.27	0.69	0.82
YouTube	0.93	0.70	1.39	1.07	1.87	-0.21	0.32	0.79	1.26	2.10	1.31	1.07	0.46
Social Network	1.69	0.56	1.46	0.98	1.73	-0.25	0.63	1.11	1.80	1.46	1.40	2.51	0.94
Instagram	0.62	0.34	1.02	0.86	1.45	-1.12	1.23	2.23	1.20	2.25	1.23	1.68	1.14
Text Processing	-0.01	1.05	1.97	2.37	1.60	2.48	-1.61	-0.53	2.55	-0.63	2.09	1.37	1.09
Microsoft Word	-0.23	1.33	2.02	2.03	1.44	2.53	-1.45	-0.40	2.33	-0.61	1.83	1.33	1.35
News Portal	2.87	-0.58	0.86	0.64	1.16	0.37	-0.18	0.56	0.62	0.37	1.50	0.31	1.21
Tagesschau.de	2.84	-1.00	1.00	0.82	1.19	0.43	-0.54	0.35	0.41	0.38	1.39	0.04	2.11
Web Shop	2.42	-0.76	1.92	1.60	2.07	1.05	0.27	1.33	2.26	0.81	2.12	2.70	1.96
Zalando.de	1.54	0.09	1.43	1.35	2.00	-0.42	0.71	1.74	0.94	1.60	1.86	2.03	1.91

References

[1] M. Schrepp, A. Hinderks, J. Thomaschewski, "Applying the User Experience Questionnaire (UEQ) in Different Evaluation Scenarios." In: *Marcus, A. (Ed.): Design, User Experience, and Usability. Theories, Methods, and Tools for Designing the User Experience. Lecture Notes in Computer Science*, Volume 8517, pp. 383-392, Springer International Publishing, 2014, DOI: 10.1007/978-3-319-07668-3_37.

[2] R. M. Ryan, "Psychological needs and the facilitation of integrative processes. Journal of personality", 63(3), pp. 397-427, 1995.

[3] K. M. Sheldon, A. J. Elliot, Y. Kim, T. Kasser, "What is satisfying about satisfying events? Testing 10 candidate psychological needs", in *Journal of personality and social psychology*, 80(2), 325, 2001.

[4] C. Lallemand, V. Koenig, G. Gronier, "How relevant is an expert evaluation of user experience based on a psychological needs-driven approach?", in *Proceedings of the 8th Nordic conference on human-computer interaction: Fun, fast, foundational* (pp. 11-20), 2014.

[5] M. Schrepp, User Experience Questionnaires: How to use questionnaires to measure the user experience of your products?, Kindle Direct Publishing, 2021, ISBN-13: 979-8736459766.

[6] A. Hinderks, F. J. Domínguez-Mayo, A.-L. Meiners, J. Thomaschewski, "Applying Importance-Performance Analysis (IPA) to Interpret the Results of the User Experience Questionnaire (UEQ)", in *Journal of Web Engineering*, 19(2), pp. 243-266, 2020.

[7] D. Winter, A. Hinderks, M. Schrepp, J. Thomaschewski, "Welche UX Faktoren sind für mein Produkt wichtig?", In: *Hess, S. & Fischer, H. (Hrsg.), Mensch und Computer 2017 - Usability Professionals.* Regensburg: Gesellschaft für Informatik e.V., pp. 191 – 200, 2017.

[8] D. Winter, M. Schrepp, J. Thomaschewski, "Faktoren der User Experience - Systematische Übersicht über produktrelevante UX-Qualitätsaspekte", In: *Endmann, A.; Fischer, H. & Krökel, M. (Eds.), Mensch und Computer 2015 – Usability Professionals*, pp. 33-41, DE GRUYTER, 2015,DOI: 10.1515/9783110443882-005.

[9] A.-L. Meiners, J. Kollmorgen, M. Schrepp, and J. Thomaschewski, "Which UX aspects are important for a software product? Importance ratings of UX aspects for software products for measurement with the UEQ+", in *Proceedings of Mensch und Computer 2021 (MuC '21)*, Association for Computing Machinery, New York, NY, USA, 2021, pp. 136–139.

[10] J. Kollmorgen, A.-L. Meiners, M. Schrepp, and J. Thomaschewski, „Ermittlung relevanter UX-Faktoren je Produktkategorie für den UEQ+", in *Mensch und Computer 2021 Workshopband*, Association for Computing Machinery, New York, NY, USA, 2021.

[11] M. Schrepp, J. Thomaschewski, "Design and Validation of a Framework for the Creation of User Experience Questionnaires", in *International Journal of Interactive Multimedia and Artificial Intelligence*, 5(7), 2019, DOI:10.9781/ijimai.2019.06.006

[12] M. Schrepp, J. Thomaschewski, "Handbook for the modular extension of the User Experience Questionnaire. - All you need to know to apply the UEQ+ to create your own UX questionnaire", 2019, DOI: 10.13140/RG.2.2.15485.20966.

[13] DIN EN ISO 9241-110: Ergonomic requirements for office work with visual display terminals (VDTs) - Part 110: Guidance on usability. *International Organization for Standardization.*

[14] DIN EN ISO 9241-210: Ergonomics of human-system interaction - Part 210: Human-centred design for interactive systems. *International Organization for Standardization.*

[15] V. Roto, E. Law, A. Vermeeren, and J. Hoonhout, (Eds), "User Experience White Paper. Outcome of the Dagstuhl Seminar on Demarcating User Experience", Germany, 2011. Accessed: Sep. 09, 2022.. https://drops.dagstuhl.de/opus/volltexte/2011/2949/pdf/10373_AbstractsCollection.2949.pdf

[16] J. R. Lewis, "Usability testing." in *Handbook of human factors and ergonomics*, 12, e30, 2006.

[17] B. E. John, D. E. Kieras, "The GOMS family of user interface analysis techniques: Comparison and contrast", in *ACM Transactions on Computer-Human Interaction (TOCHI)*, 3(4), pp. 320-351, 1996.

[18] K. Battarbee, and I. Koskinen, "Co-Experience: Product experience as social interaction", in *Product Experience*, H. N. J. Schifferstein, and P. Hekkert, Eds., Amsterdam, Elsevier, 2008, pp. 461-476.

[19] J. Preece, Y. Rogers, H. Sharpe, Interaction design: Beyond human-computer interaction, Wiley, New York, 2002.

[20] M. Hassenzahl, "The effect of perceived hedonic quality on product appealingness", in *International Journal of Human-Computer Interaction*, 13(4), pp. 481-499, 2001.

[21] M. Schrepp, R. Otten, K. Blum, and J. Thomaschewski, "What Causes the Dependency between Perceived Aesthetics and Perceived Usability?", *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, pp. 78-85, 2021, doi: 10.9781/ijimai.2020.12.005.

[22] A. Klein, A. Hinderks, M. Schrepp, J. Thomaschewski, "Construction of UEQ+ Scales for Voice Quality - Measuring User Experience Quality of Voice Interaction", In: *Proceedings of the Conference on Mensch und Computer (MuC '20)*. Association for Computing Machinery, New York, NY, USA, pp. 1–5, 2020, DOI:https://doi.org/10.1145/3404983.3410003.

[23] J. Kollmorgen, A.-L. Meiners, M. Schrepp, J. Thomaschewski, "Protokoll zur Ermittlung relevanter UX-Faktoren je Produktkategorie für den UEQ+", 2021, DOI: 10.13140/RG.2.2.16623.76960.

[24] A.-L. Meiners, J. Kollmorgen, M. Schrepp, J. Thomaschewski, "Research Protocol: Ranking of important UEQ+ factors for established products", 2021, DOI: 10.13140/RG.2.2.34986.95688.

[25] E. Duncker, "Cross-cultural usability of the library metaphor", In: *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries*. ACM, pp. 223–230, 2002.

[26] J. Noiwan, A. F. Norcio, "Cultural differences on attention and perceived usability: investigating color combinations of animated graphics", in *Int. J. Hum. Comput. Stud.* 64 (2), pp. 103–122, 2006.

[27] J. Nielsen, "Usability Engineering", Elsevier, 1994.

[28] A. Marcus, E. W. Gould, "Cultural dimensions and global web design: what? So What? Now what?", In: *Proceedings of the 6th Conference on Human Factors in the Web*, Austin Texas, 2001.

[29] R. Heimgärtner, Interkulturelles User Interface Design - Von der Idee zum erfolgreichen Produkt, *Springer*, Berlin Heidelberg, 2017.

[30] D. Forer, G. Ford, "User performance and user interface design: usability heuristics versus cultural dimensions", In: *Proceedings of the South African Computer Lecturer's Association 2003 Conference*. Johannesburg, South Africa, 2003.

[31] G. Hofstede, "Culture's Consequences – Comparing Values, Behaviors, Institutions and Organizations across Nations", 2. Edition, Thousand Oaks, London, Neu Delhi, 2001.

[32] D. Straub, M. Keil, W. Brenner, "Testing the technology acceptance model across cultures: a three country study", in *Inf. Manag.* 33 (1), pp. 1–11, 1997.

[33] J. Nantel, E. Glaser, "The impact of language and culture on perceived website usability", in *J. Eng. Technol. Manag.* 25 (1-2), pp. 112–122, 2008.

[34] W. Fitzgerald, "Models for Cross-Cultural Communications for Cross-Cultural Website Design", in *National Research Council Canada, Institute for Information Technology*, 2004.

[35] N. B. Bryan, E. R. McLean, S. J. Smith, J. Burn, "The structure of work perceptions among Hong Kong and United States IS professionals: a multidimensional scaling test of the Hofstede cultural paradigm", In: *Duane, A., Finnegan, P. (Eds.), Proceedings of the 1994 Computer Personnel Research Conference on Reinventing IS: Managing Information Technology in Changing Organizations*. ACM Press, New York, NY, USA, pp. 219–230, 1994.

[36] M. Aparicio, F. Bacao, T. Oliveira, "Cultural impacts on e-learning systems' success", in *Internet High Educ.* 31, pp. 58–70, 2016.

[37] J. C. Cronje, "Using Hofstede's cultural dimensions to interpret cross-cultural blended teaching and learning", in *Comput. Educ.* 56, pp. 596–603, 2010.

[38] A. Tarhini, K. Hone, X. Liu, "User acceptance towards web-based learning systems: investigating the role of social, organizational and individual factors in European higher education", In: *The 2013 International Conference on Information Technology and Quantitative Management*. Elsevier, pp. 189–197 2013.

[39] F. Zhao, K. N. Shen, A. Collier, "Effects of national culture on e-government diffusion—a global study of 55 countries", in *Inf. Manag.* 51, pp. 1005–1016, 2014.

[40] D. H. Zhu, Z. Q. Ye, Y. P. Chang, "Understanding the textual content of online customer reviews in B2C websites: a cross-cultural comparison between the U.S. and China", in *Comput. Hum. Behav.* 76, pp. 483–493, 2017.

[41] H. B. Santoso, M. Schrepp, A. Hinderks, J. Thomaschewski, "Cultural differences in the perception of user experience", in: *Burghardt, M., Wimmer, R., Wolff, C., Womser-Hacker, C., Hrsg (Eds.), Mensch und Computer 2017 - Tagungsband. Gesellschaft für Informatik e.V.*, Regensburg, pp. 267–272, 2017.

[42] F. Trompenars, "Riding the Waves of culture. Understanding Cultural Diversity in Business", Brealey, London, 1995.

[43] D. A. Victor, "International Business Communications", Prentice Hall, New York, 1997.

[44] E. T. Hall, "Beyond Culture", Doubleday, New York, 1989.

[45] A. Marcus, V. J. Baumgartner, "Mapping user-interface design components vs. culture dimensions in corporate websites", Visible Lang. J. MIT Press 38 (1), pp. 1–65, 2004.

[46] H. B. Santoso, M. Schrepp, "The impact of culture and product on the subjective importance of user experience aspects", Heliyon, 5(9), e02434, 2019.

Martin Schrepp

Martin Schrepp has been working as a user interface designer and researcher for SAP SE since 1994. He finished his diploma in mathematics in 1990 at the University of Heidelberg (Germany). In 1993 he received a PhD in Psychology (also from the University of Heidelberg). His research interests are the application of psychological theories to improve the design of software interfaces, the application of *Design for All* principles to increase accessibility of business software, measurement of usability and user experience, and the development of general data analysis methods. He has published several papers concerning these research fields.

Jessica Kollmorgen

Jessica Kollmorgen received a dual Bachelor of Science degree in Business Informatics from Wismar University of Applied Sciences, Technology, Business and Design. She is currently completing her Master's degree in Media Informatics with a specialization in Mobile Computing and Security at the University of Applied Sciences Emden/Leer, and is a part-time working student in the area of User Experience Research. As a part of a research group, her research interests lie in user experience factors and statistics as well as Agile UX.

Anna-Lena Meiners

Anna-Lena Meiners received a Bachelor's degree in Theatre Studies, Philosophy and Dutch Language and Literature from Freie Universität Berlin and a Bachelor's degree in Computer Science and Digital Media from University of Applied Sciences Emden/Leer. Currently, she is completing her Master's degree in Computer Science and Digital Media with a focus on Human-Computer Interaction at University of Applied Sciences Emden/Leer and working as a product owner in web development. Her main research interest is with different quality aspects of UX factors and metrics.

Andreas Hinderks

Andreas Hinderks holds a PhD in Computer Science by University of Sevilla. He has worked in various management roles as a Business Analyst and a programmer from 2001 to 2016. His focus lay on developing user-friendly business software. Currently, he is a freelancing Product Owner, Business Analyst and Senior UX Architect. He is involved in research activities dealing with UX questionnaires, measuring user experience and User Experience Management since 2011.

Dominique Winter

Dominique Winter received his M.Sc. in Computer Science from the University of Applied Sciences Emden/Leer (Germany) and his M.A. in Organizational Development from the TU Kaiserslautern (Germany). His academic research at the University of Siegen brings together the topics of agile product development, user experience design and the intentional development of organizational capabilities. Among other aspects, he examines aspects such as interactions, development opportunities and operational techniques at the levels of individuals, groups and the whole organization. Always with the goal of enabling organizations to develop better products. He combines his scientific research with many years of practical experience as a product owner and product manager in various companies.

Harry B. Santoso

Harry B. Santoso is an associate professor at the Faculty of Computer Science, Universitas Indonesia (UI). He received his B.Sc. and M.Sc. in Computer Science from UI, and his PhD in Engineering Education from the Department of Engineering Education at Utah State University, USA. He is also currently the Head of Digital Library and Distance Learning lab at the Faculty of Computer Science Universitas Indonesia. His research interest includes learning personalization, engineering and computer science education, metacognition and self-regulated learning, HCI, UX, and online distance learning.

Jörg Thomaschewski

Jörg Thomaschewski received a PhD in physics from the University of Bremen (Germany) in 1996. He became Full Professor at the University of Applied Sciences Emden/Leer (Germany) in September 2000. His research interests are human-computer interaction, agile software development, and e-learning. Dr. Thomaschewski is the founder of the research group “Agile Software Development and User Experience”.

Development of a Shared UX Vision Based on UX Factors Ascertained Through Attribution

Dominique Winter[1]*, Carolin Hausmann[2], Andreas Hinderks[3], Jörg Thomaschewski[4]

[1] University of Siegen (Germany)
[2] Fujitsu Technology Solutions GmbH (Germany)
[3] University of Seville (Spain)
[4] University of Applied Science Emden/Leer (Germany)

Received 19 January 2022 | Accepted 13 January 2023 | Published 11 April 2023

Abstract

User experience (UX) is an important quality in differentiating products. For a product team, it is a challenge to develop a good positive user experience. A common UX vision for the product team supports the team in making goal-oriented decisions regarding the user experience. This paper presents an approach to developing a shared UX vision. This UX vision is developed by the product team while a collaborative session. To validate our approach, we conducted a first validation study. In this study, we conducted a collaborative session with two groups and a total of 37 participants. The group of participants comprised product managers, UX designers and comparable professional profiles. At the end of the collaborative session, participants had to fill out a questionnaire. Through questions and observations, we identified ten good practices and four bad practices in the application of our approach to developing a UX vision. The top 3 good practices mentioned by the participants include the definition of decision-making procedures (G1), determining the UX vision with the team (G2), and using general factors of the UX as a basis (G3). The top 3 bad practices are: providing too little time for the development of the UX vision (B1), not providing clear cluster designations (B2) and working without user data (B3). The results show that the present approach for developing a UX vision helps to promote a shared understanding of the intended UX in a quickly and simply way.

Keywords

Collaborative Development, Product Vision, User Experience, User Experience Factors, UX Vision.

DOI: 10.9781/ijimai.2023.04.001

I. Introduction

As part of the human-centred design process [1], user requirements must be specified prior to embarking on creating a design. User experience (UX) is an important non-functional requirement that can be decisive in the acceptance and success of a product [2]. In other words, users can decide to accept or reject a product depending on its UX. As such, organisations are encouraged to consider UX as a design aspect during product development.

UX can be defined as 'a person's perceptions and responses resulting from the use and/or anticipated use of a product, system or service' [1]. However, the concept of UX is multi-layered and can comprise various factors [3]. These various factors create the possibility that the individuals involved in product development may generate different understandings of the product's intended UX, with each person having their own interpretation of what UX should be. These various interpretations create the risk of divergent design decisions for different product components, such that choices are not coordinated to serve a shared goal. For example, investing in a visually clear user interface during design has a positive impact on both usability and aesthetic ratings [4].

* Corresponding author.
E-mail address: dominique.winter@designik.de

To develop a shared understanding of what the outcome of product development should be, the product team must have a shared vision of the intended UX. A shared vision is generated through the process of developing the vision together, as joint decisions are reached through discussions on individual aspects. During the process of creating a shared vision, the product developers' shared pretence — or shared mental model — shapes the UX vision, which can be consciously reiterated in an explicit form (e.g., formulated as a statement). In this explicit form, the UX vision can guide those involved in product development to make or justify design decisions [5]. It can thus also serve as a basis for the design of other artefacts, such as storyboards [6], scenarios [7] or pastiche scenarios [8], and it reduces the scope of possible experiences to the agreed-upon aspects. Thus, with the UX vision, designers envision the experience a user should have when engaging with a product [5],[9].

In this paper, we present an approach for developing a UX vision. Product teams clarify which terms users ideally use to describe the UX. The terms are then clustered to identify the underlying UX factors. We reviewed the use of our approach as a tool to generate a shared understanding via a study on teams developing fictitious products versus teams developing real products. To this end, we pose the following research question:

RQ1: How does the development of a UX vision support an understanding of the intended UX?

To answer this research question, we applied our approach with 37 participants (product managers, UX designers, etc.). At the end of the exercise, we distributed a questionnaire to participants to discover how the development of the UX vision promoted shared understanding from their perspective.

We also wanted to uncover practices supporting the application of the presented approach. This goal led us to the second research question:

RQ2: What are the good and bad practices in the application of the approach?

To answer the second research question, we asked the participants in the same questionnaire as for RQ1 to describe which practices helped in the development of the UX vision and which should be avoided. In addition, the authors observed the behaviour of the participants during the exercise.

This paper is structured as follows. Section II reviews relevant research on the development of a UX vision. Section III presents our approach to developing a UX vision. The results of the evaluation are presented in Section IV and discussed in Section V. In Section VI, we draw conclusions in summary.

II. Background and Related Work

Visions are powerful tools that can orient and structure actions and behaviours [10]. A widely used form of vision in product development is the product vision. Product visions can be expressed through techniques applying, for example, a positioning statement [11] or a product box [12]. Structuring tools such as the product vision board [13] aggregate different artefacts such as personas [14] into a more comprehensive vision. A product vision does not have specific details or features; instead, it focuses on qualities important to users [15] and conceives the intended image of the product, its performance and its fit with the company's competencies and customers [16]. A good product vision keeps product teams focused on the customer, serves a common understanding of what they want to accomplish, inspires people, provides meaningful work, and gives clarity [17]. Even without a uniform definition of a product vision [18], visions help to improve the performance of teams [19]. Performance is enhanced by the possibility of making better decisions more quickly and filtering the noise, data, questions, and assumptions [20].

The concept of product vision can be valuable in understanding UX, which should be implemented by the design efforts of a team. Holtzblatt, Wendell, and Wood [21] define a vision in the context of experience-driven product development as a graphic representation describing an overarching story of the client's usage of a new product. The vision describes both the environment and how the interaction with the new product will work from the user's perspective. Szóstek [22] describes the UX vision as primarily an idea or a conception of future results, as something that experienced UX professionals use to help others follow a shared agenda.

According to Weichert, Quint and Bartel [23], a UX vision can be viewed from the perspectives of the product, the company and the user. Whereas the company perspective pictures the ideal image of the organisation (so that products can be developed in an experience-oriented way), the product perspective centres on how a product should be developed. The user perspective focuses on how users should feel or behave after or while using the product. In addition, a UX vision can also be seen as a vision for UX teams [22],[24]. This vision describes the impact of UX professionals and their integration into the organisation rather than the intended UX of a product. In this paper, we focus on the user perspective of the UX vision.

In addition to the formulation of a UX vision, there is also the practice of defining the UX goals on which the intentional product design should focus. UX goals describe individual objectives for interaction design, formalise the expected UX, and identify metrics that can be used as indicators for the achievement of UX goals [26]. Product-specific experience goals are ideally derived from company goals [27]. A UX vision can thus be seen as the totality of the UX goals because it summarises individual sub-goals. Defining goals can help product people to form a team and decide whether they share a vision and want to join the team [22], committing themselves to these goals.

In contrast to the product vision, there are no widely used structuring tools for the UX vision. The UX vision and its importance are mentioned, but specific methods for its representation or structured development are often lacking (e.g. [5],[9],[22],[28]). However, Weichert, Quint and Bartel [23] get more specific and present empathy maps or future journey maps as a picture of the users' intended interaction journey. The extent to which these methods generate a shared understanding in the product team remains open, however, as no review was performed. Specific metrics for evaluating important UX factors are not offered by previous approaches to developing a UX vision.

Developing a shared UX vision can increase the UX competence of an organisation involved in product development [28]. In a product team, the absence of a UX vision can diminish the understanding of the big picture and of the UX they are trying to achieve. It can also lead to poor integration of UX professionals into the product development process [29].

III. Research Approach

To determine how the development of a UX vision supports shared understanding of the desired UX (RQ1) and which practices support or hinder this process (RQ2), we conducted an evaluation study with 37 participants (product managers, UX designers, etc.). All Participants had to be professionally involved in the development of interactive products and making design decisions on a regular basis. It is assumed that the participants' statements on the practical use of the approach are more accurate if they can relate to real work experiences. Participants were divided into two groups (Group A and Group B). The first group (Group A) applied the approach to a fictitious product within different training sessions, while the second group (Group B) applied the approach to their actual products.

Group A consisted of 27 participants. This group applied the approach to developing a UX vision during training sessions with a fictitious product. A training exercise was developed and carried out with voluntary participants. This training concept was implemented in two German-language sessions (Group A.1 and Group A.2) using a fictitious product example (an app for analysing till receipts). The first session (Group A.1) involved 12 participants and occurred in January 2021 as part of a commercial product owner training session with a focus on UX. The training was promoted via LinkedIn, Xing and Twitter. The participants registered separately for the training and came from different companies. The training was held online. In real life, 10 participants worked as product owners and two worked as UX professionals. The second session (Group A.2) was held as a free online training event in February 2021 and was attended by 15 people. The event was promoted via LinkedIn and Twitter. The participants described themselves as product owners, product managers or UX professionals.

Group B applied the approach to a real product that participants were currently developing. This group consisted of 10 participants from three different German companies (Group B.1, Group B.2 and Group B.3). All three sessions were held in January 2021. Group B.1

consisted of a product owner and a UX professional who carried out the development of the UX vision in person. The underlying product was an online application for data management. Group B.2 consisted of four UX professionals and their team leader, who were in the process of developing a digital consultancy service. The UX vision was developed online. Group B.3 consisted of a managing director, a product manager and a UX professional. They were developing a platform for online communities in the field of sports. This application of the approach to develop the UX vision was done online.

All participants from Group A (27 participants) and Group B (10 participants) were asked about the development of the UX vision using a questionnaire. The questionnaire was sent to all participants by email shortly after their session and was answered online.

A. Procedure for the Development of a UX Vision

The general approach to developing a UX vision is divided into clarifying or recalling a product idea, collecting positive attributions, forming of clusters from the collected attributions, determining the relevant UX factors and integrating the selected UX factors into the existing product vision (see Fig. 1). These five steps were used with Group A (training sessions) and Group B (in-house product teams).

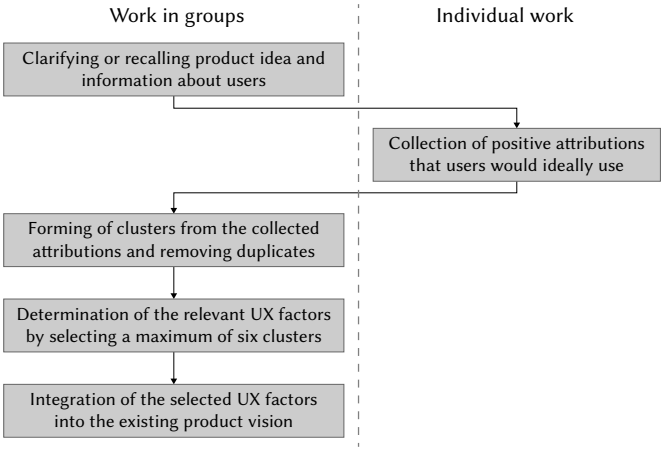


Fig. 1. Approach to developing a UX vision.

Before developing the UX vision, it was ensured that all participants understood the nature of the product they wanted to develop. For participants in an established product team, previous ideas about users and their experiences (e.g., through experience reports) were recalled. For training session participants, the idea of the product was presented in the form of a position statement (see Fig. 2; [11]). This statement was supplemented by the persona of a fictitious and realistic user [14]. These components aimed to provide the participants with the necessary contextual information to be able to develop a UX-related vision [9].

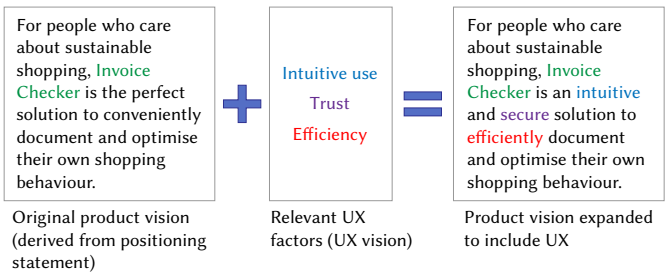


Fig. 2. Example of a product vision (derived from [11]) and its UX enrichment.

At the beginning of UX vision development, participants were introduced to how the overall approach worked. They then worked individually to identify attributions (especially adjectives) that intended users would ideally use to describe the UX. In UX vision development, the user's perspective must be considered, along with how users feel when using the product [30]. The attributions were written on individual cards with one attribution noted on each card. Participants were advised that only positive attributions should be used, as it can be assumed that UX should be positive (see Fig. 3).

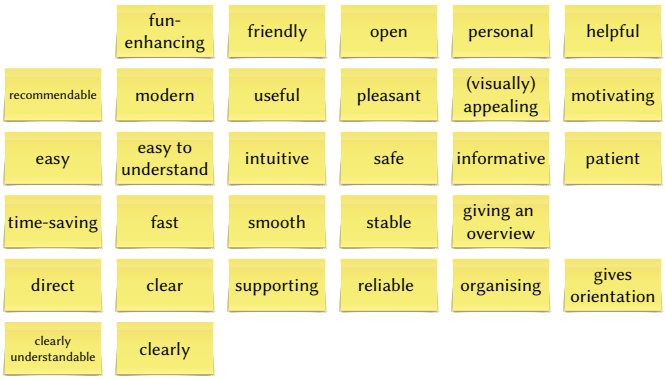


Fig. 3. Collected attributions of Group B.1.

After the participants had written their individual attributions, these attributions were further elaborated in groups. Training group participants (Group A) were divided into smaller groups with a maximum of five participants. Participants in Group B continued to work in their product teams.

The subgroups then clustered the individual cards (see Fig. 4). Duplicates were removed and unclear attributions were clarified within the group. The clusters were to be chosen such that they described a characteristic of the UX (e.g., perceived attractiveness, usefulness or controllability). Participants were asked to name the clusters themselves to check whether the added descriptions still fit the content of the cluster.

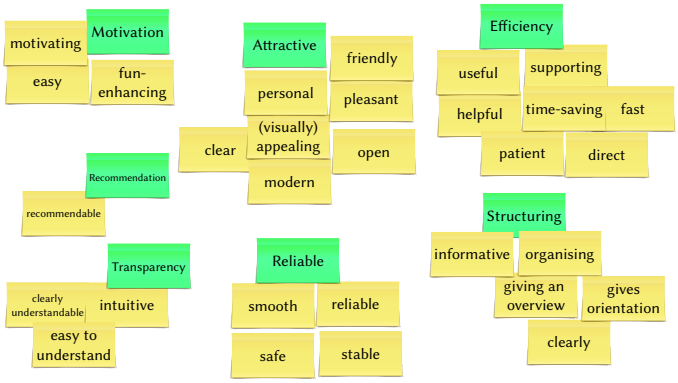


Fig. 4. Clustering of the attributions by Group B.1.

These attribution clusters represented the relevant factors of UX according to the participants' assessment. After all factors were compiled and named, the participants were asked to prioritise the identified factors. The aim was to reduce the number of possible factors to a manageable level for the product development process. Participants were asked to reduce the list of relevant factors to a maximum of six factors to be given special consideration. Up to six factors were then incorporated by the training participants (Group A) into the positioning statement presented earlier (see Fig. 2) and by the product teams (Group B) into their product vision.

B. Evaluation

After the development of the UX vision, the training participants (Group A) and the product teams (Group B) were sent an email inviting them to participate in a survey of their experience of applying the approach. We derived the questions from the research questions and structured them accordingly. As other approaches to developing a UX vision (or product vision) have not, to our knowledge, been reviewed in a structured way so far, no comparative questions could be used. The questions derived from the research questions were tested in advance with UX professionals. They were asked to verbalise what they understood by the wording. The questions were then reworked and tested again. After two iterations, the questions were finalised in their current form.

The questions were grouped into three sections: 'usefulness of the approach' (RQ1), 'good and bad practices' (RQ2) and 'personal feedback' (see Table I).

TABLE I. Item Overview

Nr	Short	Item	Type
		Section 1	
1	Better understanding (individual)	The development of the UX vision helped me to better understand the intended UX.	Likert
2	Better understanding (group)	The development of the UX vision helped us as a group to better understand the intended UX.	Likert
3	Easy procedure	The approach for developing the UX vision was easy for me to understand.	Likert
4	Produces results quickly	The approach for developing the UX vision enabled us to achieve results quickly.	Likert
5	Helpfulness for decisions	The developed UX vision will help us to make design decisions for the product.	Likert
		Section 2	
6	Good practices	What worked well in developing the experience goals?	Open question
7	Bad practices	What did not work well when developing the experience goals?	Open question
8	Helpful advice	If you were to recommend the development of experiential goals to others, what advice would you give?	Open question
		Section 3	
9	Other	What else would you like to share with us?	Open question

In the first section of the questionnaire, we asked the participants about how the presented approach supported the development of a UX vision (RQ1). Participants were asked to assess improvement in their understanding of the intended UX as a result of the approach. They were asked to state the extent to which the development of the UX vision supported them personally and the group as a whole in understanding the intended UX. In addition, the participants were asked to rate how comprehensible they found the approach for developing a UX vision and how efficient they felt it was. In a final question of the first section, they were asked to rate how much the developed UX vision would help them further develop the product. The items of the first section of the questionnaire were rated using a 5-point Likert scale [31] ranging from 'disagree' to 'agree' (see Fig. 5).

The second section of the questionnaire regarded good and bad practices in the application of the presented approach (RQ2), about which participants were asked to describe what worked well and what did not work. Participants were then asked to state what advice they would give to others who were using the approach to develop a UX vision. The second section posed open questions.

The third section of the questionnaire allowed participants to give further feedback. They were asked to indicate (via an open question) what additional information they would like to share on any aspect of the study.

IV. Results

In the following section, we present the results of the survey according to the research questions and offer additional observations regarding the development of the UX vision in the study.

A. RQ1: How Does the Development of a UX Vision Support a shared Understanding of the Intended UX?

Participants in the training sessions (Group A, N = 27) indicated that developing the UX vision helped them to understand the intended UX (Item 1, mean of 4.296, standard deviation of 0.597, confidence interval [95%] of ±0.225; see Fig. 6). Product team participants (Group B, N = 10) who developed a UX vision for their own products agreed with this statement (Item 1, mean 4.300, standard deviation of 0.640, confidence interval [95%] of ±0.397; see Fig. 6).

The approach to developing the UX vision was rated easy for the training participants (Group A) to understand (Item 2, mean 4.481, standard deviation of 0.630, confidence interval [95%] of ±0.238; see Fig. 6). This rating was also confirmed by participants in Group B in the context of applying the approach to the development of their own products (Item 2, mean 4.300, standard deviation of 0.781, confidence interval [95%] of ±0.484; see Fig. 6).

Training participants (Group A) agreed on the usefulness of the approach to support a shared understanding of the UX vision (Item 3, mean value 4.259, standard deviation of 0.745, confidence interval [95%] of ±0.283; see Fig. 6). This was confirmed by the participants in Group B (mean value 4.400, standard deviation of 0.800, confidence interval [95%] of ±0.496; see Fig. 6).

The training participants (Group A) agreed that the approach for developing the UX vision quickly yielded results (Item 4, mean value 4.333, standard deviation of 0.816, confidence interval [95%] of ±0.308; see Fig. 6). Applied to their own products, Group B respondents also agreed (to a lesser extent) that the approach quickly yielded results (mean 3.900, standard deviation of 0.831, confidence interval [95%] of ±0.515; see Fig. 6).

Regarding to how far the developed UX vision aided design decisions (Item 5), Group A participants expressed high levels of agreement (Item 5, mean 4.259, standard deviation of 0.699, confidence interval [95%] of ±0.264; see Fig. 6), while Group B expressed slightly lower levels of agreement (mean 4.000, standard deviation of 0.632, confidence interval [95%] of ±0.392; see Fig. 6).

	Disagree	**Somewhat disagree**	**Neither agree nor disagree**	**Somewhat agree**	**Agree**
The development of the UX vision helped me to better understand the intended UX.	☐	☐	☐	☐	☐

Fig. 5. Example of one item of the questionnaire using a Likert scale.

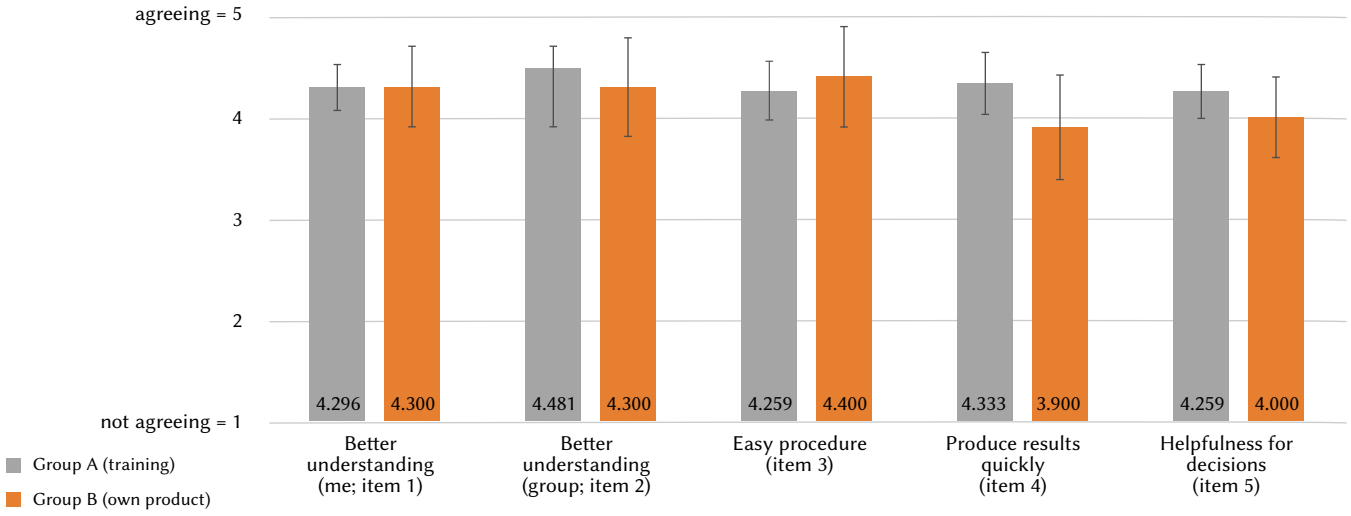


Fig. 6. Results of quantitative items (mean values [range 1–5] and confidence intervals).

B. RQ2: What Are the Good and Bad Practices in the Application of the Approach?

The evaluation of the open-text questions (Items 6–9) based on the answers given by Group A and Group B revealed seven good practices (G1–G7) and four bad practices (B1–B4) to develop a shared UX vision with the proposed approach (see Table II). The open-text questions were independently evaluated by two of the authors. In addition, three other good practices (G8–G10) were identified through observations made by one of the authors and another facilitator.

TABLE II. Good and Bad Practices Identified by Participants (the Number of Individual Participants Who Mentioned Each Concept Appears in Brackets)

Good Practices	Bad Practices
G1: Definition of decision-making procedures (12) G2: Determining the UX vision with the product team (12) G3: Using general factors of the UX as a basis (6) G4: Explicitly allowing different perspectives (5) G5: Naming the purpose of the UX vision beforehand (3) G6: Use of an external moderator (3) G7: Reflecting on customer feedback and behaviour in advance (3) Discovered by Observation G8: Repeatedly referring to the user perspective G9: Explaining the approach beforehand G10: Providing an example of a UX vision	B1: Providing too little time for development of the UX vision (8) B2: Not providing clear cluster designation (3) B3: Working without user data (2) B4: Not making the development approach transparent in advance (1)

1. Good Practices

Good practices mentioned by the participants include the *definition of decision-making procedures (G1)*, *determining the UX vision with the product team (G2)*, *using general factors of the UX as a basis (G3)*, *explicitly allowing different perspectives (G4)*, *naming the purpose of the UX vision beforehand (G5)*, the *use of an external moderator (G6)*, and *reflecting on customer feedback and behaviour in advance (G7)*. In addition, several other good practices were noted through observation, including *repeatedly referring to the user perspective (G8)*, *explaining the approach beforehand (G9)* and *providing an example of a UX vision (G10)*.

To ensure the development of a UX vision progresses quickly, it is advisable to use given *decision-making procedures (G1)*. In the observed constructions of the UX vision, the participants used dot-voting for this purpose so that relevant UX factors could be efficiently determined. If factors are put in an absolute order and discussed individually, the time required increases. However, relevant factors must be reduced to a manageable number, which can be achieved by simplifying voting procedures.

For *determining the UX vision with the product team (G2)*, it is advisable to involve several members of the product team in the development of the UX vision; this involvement can increase the acceptance of the developed UX vision among team members. In addition, different perspectives allow for a more comprehensive view of users' needs, which can lead to a more appropriate UX vision.

Using general factors of the UX as a basis (G3) means using a predefined list of possible factors (e.g., [3], [32]) to form clusters as concrete as possible from the individual attributions. However, care must be taken to ensure that other factors are also perceived as possible, beyond the predefined list.

Explicitly allowing different perspectives (G4) supports the adequate description of the UX vision for the intended UX because participants can bring their different professional perspectives to the discussion. Therefore, an open atmosphere can lead to a broader discussion. The UX vision can be discussed more comprehensively through different professional perspectives.

Naming the purpose of the UX vision beforehand (G5) helps all participants by allowing them to prepare for development and to reflect on important information in advance. Furthermore, they can align their decisions with the intended goal during the design.

Although members of product teams can also moderate the construction of the UX vision, the *use of an external moderator (G6)* is recommended. This moderator can focus exclusively on the approach. Due to their external standing, they can also critique the group discussions and contribute to solutions as mediators.

To focus discussions and decisions on user needs during the construction of the UX vision, it is advisable to *reflect on customer feedback and observations in advance (G7)*. This reflection allows participants to refer back to information they have already gathered and to keep users at the centre of the discussion.

Repeatedly referring to the user perspective (G8) helps vision developers to keep sight of the user perspective. During the exercise, it was observed that—contrary to the instructions—participants

often adopted the perspective of product designers and did not use the language of users. The facilitator then had to point out that they should formulate positive attributions from the user perspective; this was difficult for some participants but could be solved through the collaboration with other participants.

It helps product teams to *explain the approach beforehand (G9).* Participants know what is expected of them at each step, thereby reducing uncertainty.

When elaborating the UX vision, *providing an example of a UX vision (G10)* helps product teams. In this study, product teams used an example of a UX vision as an initial point of orientation for their discussions. This example supported their discussion and was repeatedly used as a guide in the formation of clusters.

2. Bad Practices

Bad practices include *providing too little time for the development of the UX vision* (B1), *not providing clear cluster designations* (B2), *working without user data* (B3) and *not making the development approach transparent in advance* (B4).

Providing too little time for the development of the UX vision (B1) can lead to the results being perceived as being too generic or as being unhelpful. Development of a UX vision usually occurs within the framework of structured and time-limited activities (e.g., as a workshop). For this reason, product teams should allow sufficient time for development. If necessary, they should set a follow-up date.

Not providing clear cluster designations (B2) means that groups cannot agree on the desired characteristics and, thus, that further discussions cannot be conducted effectively. This issue may become apparent only when the UX vision is further applied (e.g., when prioritising requirements); it may then lead to conflicts because the participants have different ideas. Each cluster must, therefore, be labelled with a unique term accepted by the whole group. A facilitator can provide support here.

Working without user data (B3) while creating the UX vision limits the developed content to assumptions about users and their UX. If user data (e.g., survey results, personas or user ratings) are available, product teams should use as much of this data as possible to develop the UX vision. These data allow product teams to develop a comprehensive picture of the users and to use this as a basis for selection decisions.

By a product team *not making the approach of development transparent in advance (B4),* participants may become insecure and may be prevented from understanding the results of the individual steps. If participants understand how positive attributions are processed after collection, they can more deliberately formulate these attributions. For this reason, the entire procedure should be presented at the beginning of developing a UX vision.

V. Discussion and Limitations

Based on the results of our study, it can be assumed that the development of a UX vision according to the presented approach promotes both individual and shared understanding of the UX (from the perspective of the participants). On this approach, participants structure their thoughts around the intended UX. Very similar results were found in the context of the training teams (Group A) and the teams with real products (Group B). Therefore, the presented approach seems to work both in a constructed scenario (Group A) and in practice (Group B).

Although all participants stated that the approach was easy to understand and led quickly to results, the introduction to the approach likely influenced participants' understanding substantially. It can also be assumed that the way the approach was introduced and explained influenced its success. We did not investigate how significant this influence was. We therefore recommend that the person introducing the approach should have both methodological and didactic experience.

Although the participants stated that the developed UX vision was helpful for design decisions in further product development, we did not check whether this was the case during actual product development. However, it can be assumed that consensus-building within the product team simplifies and promotes later design decisions because the developed UX vision can be referenced.

Some of the good and bad practices identified can be transferred to other ways of working. For example, limiting the available time can also create problems in other approaches, while considering different professional perspectives in the discussions of other workshops can have generally positive effect.

The repeated assumption of the user perspective may also be applicable to other group-based approaches (e.g., prioritising requirements). The reference to lack of user data and user feedback can also be generalised to other decision-making.

Bad practices should be avoided; they can be transformed into good practices by taking the opposite actions. For example, the bad practice of working without user data can be reversed and made positive by working with user data.

UX factors were selected by participants. Therefore, the factors are based on individual assessments and not on factors pertinent to real users. For this reason, the relevance of the factors must be validated in the course of product development. At the beginning of the product development process, no tangible product versions are available. Initial ideas also have to be implemented for prototypes.

In addition to the bad practices mentioned, it is to be expected that typical biases of group work can also come into effect in the presented approach. Dominant personalities could prevail more strongly in the selection of the UX factors to be considered further [33]. Compromises made in group work do not necessarily lead to better results. It is therefore important to compare the selected UX factors once again with the perception of users, even if the approach presented does not include this.

For the first product versions to be developed and validated, product teams must make initial design decisions that are aligned with a targeted type of experience through a UX vision. A review of the selected factors should be validated with real users in the later stages of product development.

Except for one implementation, all steps of the study occurred online. There is no indication that the development of a UX vision according to the presented approach would not also work in an in-person workshop. It is possible that different time investments must be made due to other group dynamic effects.

Since, to our knowledge, validations of other approaches to developing a shared UX vision are not available, our results cannot be compared. Our results show that the approach presented is perceived as effective. However, as of today, it is not possible to say whether another approach would be better or worse. Compared to empathy maps or future journey maps as a picture of the users' intended interaction journey [23], the approach presented allows a direct transfer into UX management, as UX factors are determined.

VI. Conclusion and Future Research

This paper describes an approach to developing a UX vision based on UX factors as a product team and presents the results of a study with 37 participants. In this study, training participants (Group A, N = 27) and real product teams (Group B, N = 10) were asked to develop a

UX vision. The study investigated how the development of a UX vision supports an understanding of the intended UX (RQ1) and establishes the good and bad practices involved in developing a UX vision using the presented approach (RQ2).

The first research question addressed how the development of a UX vision supports a shared understanding of the intended UX. We have shown that the examined approach promotes, from the perspective of the participants, both individual and collaborative understanding of the intended UX and the relevant UX factors, mainly by structuring discussions and related design decisions. Furthermore, the approach was perceived as accessible and efficient. The UX vision resulting from the approach was judged to be conducive to further decision-making, so it can be assumed that design decisions during product development are promoted by this vision.

The second research question aimed to identify helpful and detrimental practices in developing a UX vision. Good practices in developing a UX vision include definition of decision-making procedures, the use of an external moderator, the application of general factors of the UX as a basis, the naming of the purpose of the UX vision beforehand, the determination of the UX vision with the product team, explicitly allowances for varying perspectives, reflection on customer feedback and behaviour in advance, repeated reference to the user perspective, explaining the approach beforehand, and providing an example of a UX vision. Product teams should work with user data, make the development approach transparent in advance, and provide clear cluster designations and sufficient time for development of the UX vision.

In summary, this approach to developing a UX vision seems to help a product team to develop a shared idea of the intended UX quickly and simply. The developed UX vision can be the basis for many design decisions during product development, such as helping to prioritise the desired features by UX factors. By adopting the language of users, the desired UX can also be communicated comprehensibly within the organisation. A structured approach to developing the vision also allows the justification of the selection of relevant UX factors. Thus, the presented procedure represents a suitable basis on which to initiate product development and supports, for example, collaborative estimation methods such as UX Poker [34].

Further research should be completed to study the practical use of the developed UX vision in an actual product development process over a longer development period. While it can be assumed that a UX vision generates value when used and communicated, it remains to be proven what kind of communication and what kind of embedding in other product development artefacts is particularly conducive to achieving an intended UX.

References

[1] International Organization for Standardization, "*Ergonomics of human-system interaction - Part 210: Human-centred design for interactive systems*," Vernier, Switzerland, 2019. Available: https://www.iso.org/obp/ui/en/#iso:std:iso:9241:-210:ed-2:v1:en.

[2] S. Djamasbi, D. McAuliffe, W. Gomez, G. Kardzhaliyski, W. Liu, and F. Oglesby, "Designing for Success: Creating Business Value with Mobile User Experience (UX)," in *HCI in business: First international conference, HCIB 2014, held as part of HCI International 2014*, Editor F. F.-H. Nah, Heraklion, Greece: Springer, 2014, pp. 299–306, doi: 10.1007/978-3-319-07293-7_29.

[3] D. Winter, M. Schrepp, and J. Thomaschewski, "Faktoren der User Experience: Systematische Übersicht über produktrelevante UX-Qualitätsaspekte," in *Mensch und Computer 2015 - Usability Professionals*, A. Endmann, H. Fischer, and M. Krökel, Eds., Berlin: De Gruyter Oldenbourg, 2015, pp. 33–41.

[4] M. Schrepp, R. Otten, K. Blum, J. Thomaschewski, "What causes the dependency between perceived aesthetics and perceived usability?," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, pp. 78–85, 2021, doi: 10.9781/ijimai.2020.12.005.

[5] V. Roto, E. Kaasinen, M. Nuutinen, and M. Seppänen, "UX Expeditions in Business-to-Business Heavy Industry," in *#chi4good: CHI 2016*, San Jose, CA, USA, 2016, pp. 833–839, doi: 10.1145/2851581.2851600.

[6] S. M. Branham, S. Wahid, and D. S. McCrickard, "Channeling Creativity: Using Storyboards and Claims to Encourage Collaborative Design," *Workshop on Tools in Support of Creative Collaboration (part of Creativity & Cognition 2007)*, pp. 1–4, 2007.

[7] I. Alexander and N. Maiden, *Scenarios, stories, use cases: Through the systems development life-cycle*, Chichester, UK: Wiley, 2004.

[8] M. Blythe and P. Wright, "Pastiche scenarios: Fiction as a resource for user centred design," *Interacting with Computers*, vol. 18, no. 5, pp. 1139–1164, 2006, doi: 10.1016/j.intcom.2006.02.001.

[9] J. Kollmann, H. Sharp, and A. Blandford, "The Importance of Identity and Vision to User Experience Designers on Agile Projects," in *2009 Agile Conference*, Chicago, IL, USA, 2009, pp. 11–18, doi: 10.1109/AGILE.2009.58.

[10] P. Vergragt and H. Brown, "Sustainable mobility: from technological innovation to societal learning," *Journal of Cleaner Production*, vol. 15, 11-12, pp. 1104–1115, 2007, doi: 10.1016/j.jclepro.2006.05.020.

[11] G. A. Moore, *Crossing the chasm: Marketing and selling disruptive products to mainstream customers*, New York, USA: Harper Business Essentials, 2002.

[12] J. A. Highsmith, *Agile project management: Creating innovative products*, Boston, USA: Addison-Wesley, 2004.

[13] R. Pichler, *Strategize: Product strategy and product roadmap practices for the digital age*, Wendover, UK: Pichler Consulting, 2016.

[14] A. Cooper, R. Reimann, and D. Cronin, *About face 3: The Essentials of Interaction Design*, Indianapolis, USA: Wiley, 2007.

[15] M. Perri, *Escaping the build trap: How effective product management creates real value*, Sebastopol, CA, USA: O'Reilly, 2019.

[16] S. L. Brown and K. M. Eisenhardt, "Product Development: Past Research, Present Findings, and Future Directions," *The Academy of Management Review*, vol. 20, no. 2, p. 343-378, 1995, doi: 10.2307/258850.

[17] M. Cagan and C. Jones, *Empowered: Ordinary people, extraordinary products*, Hoboken, New Jersey, Wiley, 2021.

[18] J. L. G. Benassi, D. C. Amaral, and L. D. Ferreira, "Towards a conceptual framework for product vision," *International Journal of Operations & Production Management*, vol. 36, no. 2, pp. 200–219, 2016, doi: 10.1108/IJOPM-11-2013-0511.

[19] P. Tessarolo, "Is Integration Enough for Fast Product Development? An Empirical Investigation of the Contextual Effects of Product Vision," *Journal of Product Innovation Management*, vol. 24, no. 1, pp. 69–82, 2007, doi: 10.1111/j.1540-5885.2006.00233.x.

[20] P. Wille, *Strong product people: A complete guide to developing great product managers*, Hamburg, Germany: Petra Wille, 2020.

[21] K. Holtzblatt, J. B. Wendell, and S. Wood, *Rapid contextual design: A how-to guide to key techniques for user-centered design*. San Francisco, USA: Elsevier/Morgan Kaufmann, 2005.

[22] A. Szóstek, "A look into some practices behind microsoft UX management," in *2012 annual conference extended abstracts on Human factors in computing systems*, Austin, TX, USA, 2012, pp. 605-618, doi: 10.1145/2212776.2212833.

[23] S. Weichert, G. Quint, and T. Bartel, *Quick guide UX Management: So verankern sie usability und user experience im Unternehmen*, Wiesbaden, Germany: Springer Gabler, 2018.

[24] K. Kaplan, *How to Create a UX-Vision Statement*, Nielsen Norman Group, USA, 2022. Accessed: June 6, 2022. [Online]. Available: https://www.nngroup.com/articles/ux-vision-statements/.

[25] A. Lund, *User experience management: Essential skills for leading effective UX teams*, Amsterdam, Netherlands: Morgan Kaufmann/Elsevier, 2011.

[26] H. R. Hartson and P. S. Pyla, *The UX book: Process and guidelines for ensuring a quality user experience*, Amsterdam, Netherlands: Elsevier Morgan Kaufmann, 2012.

[27] V. Roto, Y. Lu, H. Nieminen, and E. Tutal, "Designing for User and Brand Experience via Company-wide Experience Goals," in *CHI 2015: extended abstracts publication of the 33*rd *annual CHI Conference on Human Factors in Computing Systems*, Seoul, Republic of Korea, 2015, pp. 2277–2282, doi: 10.1007/978-3-662-44811-3_17.

[28] D. Winter, A. Hinderks, and J. Thomaschewski, "Steigerung der organisationalen UX-Kompetenz," in *Mensch und Computer 2018 - Usability Professionals*, Dresden, Germany, 2018, pp.15-25.

[29] K. Kuusinen, "Improving UX Work in Scrum Development: A Three-Year Follow-Up Study in a Company," in *Human-centered software engineering: 5*th *IFIP WG 13.2 international conference*, Paderborn, Germany, 2014, pp. 259–266, doi: 10.1007/978-3-642-02806-9.

[30] P. Gómez-López, F. M. Simarro, and M. T. L. Bonal, "Analysing the UX scope through its definitions," in *XX International Conference on Human Computer Interaction*, Donostia Gipuzkoa, Spain, 2019, pp. 1–4, doi: 10.1145/3335595.3335648.

[31] R. Likert, "A technique for the measurement of attitudes," *Archives of Psychology*, no. 22, pp. 5–55, 1932.

[32] M. Schrepp and J. Thomaschewski, "Design and Validation of a Framework for the Creation of User Experience Questionnaires," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 7, p. 88-95, 2019, doi: 10.9781/ijimai.2019.06.006.

[33] C. Anderson and G.J. Kilduff. "Why do dominant personalities attain influence in face-to-face groups? The competence-signaling effects of trait dominance," *Journal of Personality and Social Psychology*, vol. 96 no. 2, pp. 491–503, 2009, doi: 10.1037/a0014201.

[34] A. Hinderks, D. Winter, F. J. D. Mayo, M. J. Escalona, J. Thomaschewski (2022), "UX Poker: Estimating the Influence of User Stories on User Experience in Early Stage of Agile Development", *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 7, pp. 97-104. doi. doi.org/10.9781/ijimai.2022.11.007

Dominique Winter

Dominique Winter holds a Master of Science in Media Informatics from the University of Applied Sciences Emden/Leer (Germany) and a Master of Arts in Organisational Development from the TU Kaiserslautern (Germany). He works for various companies across Europe as a product development coach and supports them in improving their user orientation. He is also a doctoral student at the University of Siegen and conducts research on the topic of UX competence in and of organisations.

Carolin Hausmann

Carolin Hausmann completed her Master of Arts in Management for Marketing and Sales at the Hochschule für Oekonomie und Management. She currently works as a Senior Marketing Specialist at Fujitsu Technology Solutions GmbH and is specifically responsible for the company's server and software products. In the past, she built up the UX/UI area for a software company and integrated it into the company processes. She continues to work intensively on the topics of UX and UI alongside her job.

Andreas Hinderks

Dr. Andreas Hinderks holds a PhD in Computer Science from the University of Sevilla. He has worked in various management roles as a Business Analyst and a programmer from 2001 to 2016. His focus lay on developing user-friendly business software. Currently, he is a freelancing Product Owner, Business Analyst and Senior UX Architect. He has been involved in research activities dealing with UX questionnaires, measuring user experience and User Experience Management since 2011.

Jörg Thomaschewski

Dr. Jörg Thomaschewski received a PhD in physics from the University of Bremen (Germany) in 1996. He became Full Professor at the University of Applied Sciences Emden/Leer (Germany) in September 2000. His research interests are human-computer interaction, and agile software engineering. Dr. Thomaschewski founded the research group "Agile Software Development and User Experience" at the University of Applied Sciences Emden/ Leer in 2009. He also has extensive experience in user experience training, user experience management, UX questionnaires, agile methods, IT analysis, e-learning, and consulting.

LA UNIVERSIDAD EN INTERNET

Rectorado
Avenida de la Paz, 137
26006 Logroño (La Rioja)
t (+34) 941 21 02 11

www.unir.net