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 | Early Access 28 October 2021

unir
LA UNIVERSIDAD EN INTERNET

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

Please cite this article in press as:
V. Singh, D. Jain. A Hybrid Parallel Classification Model for the Diagnosis of Chronic Kidney Disease, International Journal of Interactive Multimedia and Artificial Intelligence, (2021), http://dx.doi.org/10.9781/ijimai.2021.10.008

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.

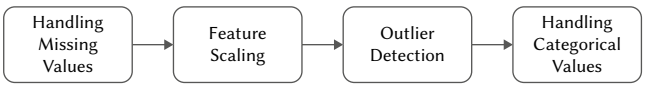

Fig. 1. Steps of Pre-processing.

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.

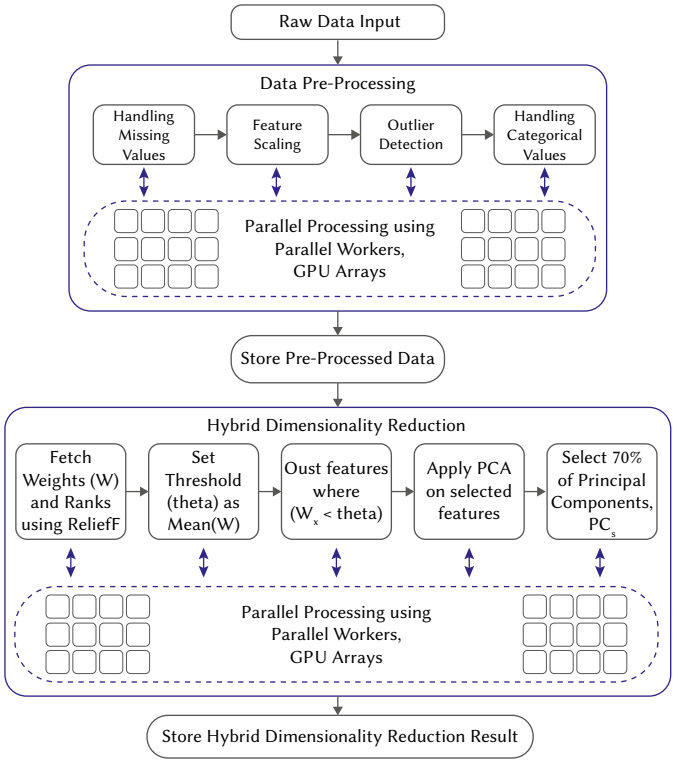

Fig. 2. Proposed Model (Phase I).

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.

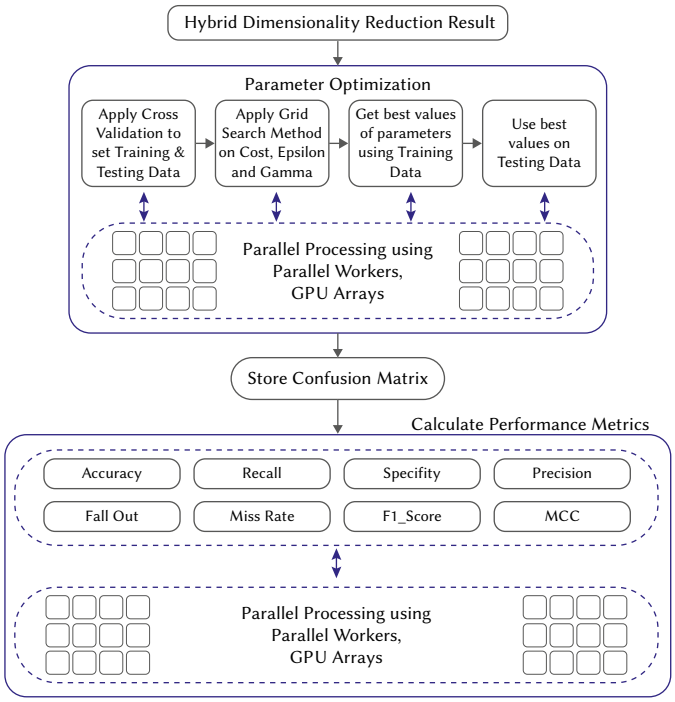

Fig. 3. Proposed Model (Phase II).

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

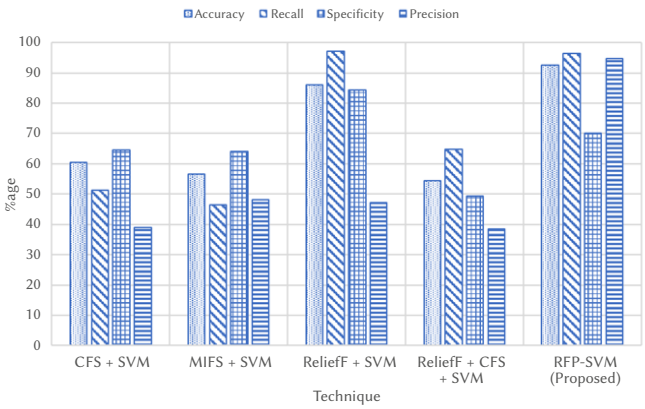

Fig. 5. Graphical representation for Clinical CKD Dataset.

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

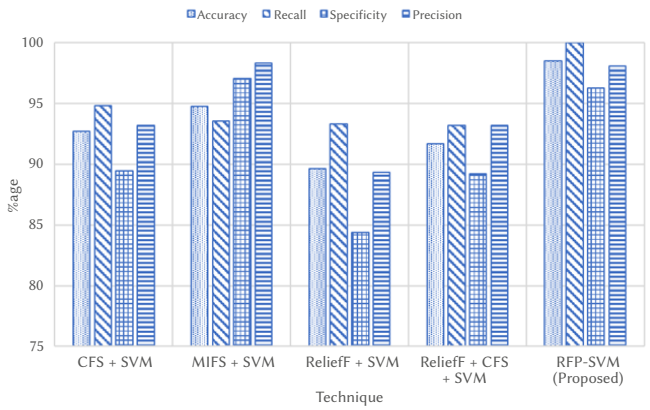

Fig. 6. Graphical representation for Repository CKD Dataset.

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.