# Deep Transfer Learning-Based Automated Identification of Bird Song

Nabanita Das[1]*, Neelamadhab Padhy[1], Nilanjan Dey[2], Sudipta Bhattacharya[3], João Manuel R.S. Tavares[4]

[1] Department of Computer Science and Engineering, GIET University, Gunupur (India)
[2] Department of Computer Science & Engineering, Techno International New Town, Kolkata (India)
[3] Department of Computer Science & Engineering, Bengal Institute of Technology, Kolkata (India)
[4] Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial, Departamento de Engenharia Mecânica, Faculdade de Engenharia, Universidade do Porto, Porto (Portugal)

Received 13 September 2022 | Accepted 14 December 2022 | Early Access 12 January 2023



## Abstract

Bird species identification is becoming increasingly crucial for avian biodiversity conservation and assisting ornithologists in quantifying the presence of birds in a given area. Convolutional Neural Networks (CNNs) are advanced deep learning algorithms that have proven to perform well in speech classification. However, developing an accurate deep learning classifier requires a large amount of data. Such a large amount of data on endemic or endangered creatures is frequently difficult to gathered. Also, in some other fields, such as bioinformatics and robotics, the high cost of data collection and expensive annotation limit their progress, so large, well-annotated data creating a set is also difficult. A transfer learning method can alleviate overfitting concerns in a deep learning model. This feature serves as the inspiration for transfer learning, which was created to deal with situations where the data are distributed across a variety of functional domains. In this study, the ability of deep transfer models such as VGG16, VGG19 and InceptionV3 to effectively extract and discriminate speech signals from different species of birds with high prediction accuracy is explored. The obtained accuracies using VGG16, VGG19 and InceptionV3 were equal to 78, 61.9 and 85%, respectively, which are very promising.

## Keywords

Bird Species Recognition, Convolution Neural Network, Data Augmentation, InceptionV3, Transfer Learning, VGG16, VGG19.

DOI: 10.9781/ijimai.2023.01.003

## I. Introduction

Birds not only enhance nature's charm and beauty but also help maintain the balance of the new environment of the world. Because they are essential parts of natural systems, birds have ecological importance. Birds manage insects and rodents, pollinate crops, spread seeds, and serve humans directly. Bird vocalizations are very noticeable, which makes them a helpful tool for population monitoring and biodiversity assessment. Bird vocalization includes both calls and songs. Birds are essential to our ecology. For instance, birds keep our globe beautiful by controlling pests, pollinating crops, and preserving the ecology of an island. There are around 10,000 species on earth, according to [1]. Birds make sounds for many reasons, including locating territories, which is important for male birds, inviting a mate to mate, reacting to their environment, and determining whether or not they are in danger [2]. People often find it difficult to distinguish between a bird's song and a call, especially if they are unfamiliar with birds. An audio recording of a bird's voice is an essential tool for identifying the species of a bird for a biologist who is interested in the study, management, and conservation of birdlife [3]. There are many bird calls, and it is hard for people to figure out which ones have a place with animal categories. The manual recording and recognition of avian sounds is inconvenient and can sabotage bird conservation efforts. As a result, accurate, scalable, and automated bird species recognition is essential for wildlife monitoring and can help conserve avian biodiversity [4]-[7]. The identification of bird species is a classic pattern recognition problem, and most research includes sections on signal pre-processing, feature extraction, and classification [8], [9]. Deep learning has received increased attention from researchers recently since it has been successfully used in a number of practical applications. To stop the rapid loss of avian variety in this area, deep-learning algorithms for bird detection are appropriate [10]. In this context, several automated bird detection models were developed. Additionally, a test has been performed on a system that can recognize new bird songs and learn from previously recorded annotated bird sounds. This system may provide accurate information on the presence or absence of a target species as well as the overall biodiversity status of a region. Deep learning algorithms are superior to conventional machine learning techniques because they can extract high-level characteristics from enormous amounts of data [11]. On the other hand, traditional machine learning approaches need users to construct features, which demands significant manual work.

On the other hand, deep learning approaches automatically extract data features using a hierarchical feature extraction method and an unsupervised or semi-supervised feature learning methodology [12]-[14]. Deep learning can be defined as a representation learning

* Corresponding author.
E-mail address: nabanita.das@giet.edu

Please cite this article in press as:
N. Das, N. Padhyb, N. Deyc, S. Bhattacharyad, J. Manuel R.S. Tavarese. Deep Transfer Learning-Based Automated Identification of Bird Song, International Journal of Interactive Multimedia and Artificial Intelligence, (2023), http://dx.doi.org/10.9781/ijimai.2023.01.003

algorithm in machine learning that is based on large data. Although deep learning models can achieve good predictive performance, such models require a huge number of unique data points to achieve this performance and this turns out to be challenging for endangered or endemic birds, as inadequate data overwhelms deep learning models. One of the fundamental problems of deep learning is data dependency. Deep learning is more dependent on training data than traditional machine learning methods since it needs a lot of data to find latent patterns in the data. Inadequate training data is unavoidable in some deep-learning applications. For instance, the high cost of data collection and expensive annotation, which impedes development, make it difficult to produce a sizable, thoroughly annotated dataset for each sample in a bioinformatics dataset [15], [16],[17]. The issue of overfitting in a deep model can be solved using a transfer learning technique [18]. Because transfer training makes the condition that the training data be independent and distributed equally with the test data simpler, it can address the issue of a lack of training data. Transfer learning drastically reduces the amount of training data and time needed for the target domain, because it does not require training and testing data or starting from scratch to train the target domain model.

In this experimental study, 7 different bird species were correctly identified using 16387 test samples from the xeno-canto database. Due to the limited sample size, several data augmentation techniques have been studied, and the underlying hypotheses were thoroughly evaluated. It was interesting to note that such type of augmentation techniques results in overfitting of the models. In light of these considerations, the idea of transfer learning was chosen for this investigation. By using transfer learning, a total of 36 species were classified rather than 7. Because of the limited availability of high-quality data, pre-trained models have been used in the identification of 37 different categories of birds. To develop this investigation on the local bird recognition in Sundarban, West Bengal, India, two deep learning models were used. Hence, InceptionV3 and MobileNet were initially tested without the use of transfer learning technology, and then they were tested once again with it. Finally, the findings were compared using MobileNet and transfer learning, employing performance evaluation metrics such as accuracy and F1 score. In the experiment, the result showed that in the VGG16 model the training accuracy was 75%, while the test one accuracy was 78%. Respectively in the VGG19 model, the training accuracy was 64%, while the test accuracy was 61.9%. On the InceptionV3 model, which was employed in additional tests, an accuracy of 95% was reached during training, while an accuracy of 85% was achieved which obtained the best result. In the InceptionV3 model, ImageNet was used as a weight, and average pooling was employed. The rest of this article is organized in the following way. Section 2 describes the related research. The methodologies used are detailed in Section 3. Section 4 shows the results and their analysis, which are followed by a discussion and conclusion in section 5, and concludes what future work can be done.

## II. Literature Survey

Different researchers have proposed different features for the audio sounds of birds, and artificial intelligence techniques have been used to voice classification. CNN models that use Mel spectrogram or mel frequency cepstral coefficient (MFCC) derived from audio data have been observed to dominate the most promising solutions [19]. However, recent trends show that the best results were achieved by the works that used Convolutional Neural Networks with transfer learning [20]. The best results were for the most part from using Resnet, Inception, and VGG models Additionally, Fritzler et al. [21] propose the Inception-v3 pre-trained convolutional neural network-based bird recognition system. The technology was enhanced with 36,492 audio recordings of 1,500 different bird species for the BirdCLEF 2017 task. The audio recordings were afterward transformed into spectrograms and used for data augmentation. According to this study, optimizing a pre-trained convolutional neural network trumps starting from scratch in terms of performance. For acoustic bird detection, Ntalampiras [22] introduced a transfer learning framework employing the probability density distribution of ten musical genres to determine the degree of affinities between different bird species and various musical genres. Deep learning models based on CNNs are efficient categorization models. However, getting numerous training samples in specialist disciplines like bird acoustics is expensive and difficult as they require a large amount of data for training. To address this issue, transfer learning is one method that can classify data with a limited number of training examples. DB Efremov et al. [23] assessed the effectiveness of birdcall classification utilizing transfer learning from a bigger base dataset to a smaller target dataset using a ResNet-50 CNN in this regard. A bird recognition model built on Inception-v3 was presented by J. Bai et al. [24] can identify and categorize 659 different bird species from supplied audio recordings. Inception-v3 is used to recognize bird sounds by using log-Mel spectrograms as features.

To enhance the model's performance, several data augmentation strategies were employed. In order to categorize the cries of 24 species of birds and amphibians discovered in environmental field recordings, Zhong M. et al. [25] created a deep convolutional neural network. Their primary objective was to prepare enough training data, which is a significant difficulty for many deep-learning applications. To tackle this problem, they created a pre-trained deep convolutional neural network by fusing the idea of transfer learning with a supervised pseudo-labeling technique and an eigen loss function. In order to categorize grouper species based on the courtship-related noises they make during spawning aggregations, Ibrahim suggests a transfer learning technique, A. K. [26]. On the other hand, Rajan R. et al. [27] suggested a method for learning bird vocalizations utilizing sliding window analysis on the Mel spectrogram and a pre-trained Deep Convolutional Neural Network (DCNN), a VGG16 model. Using a deep learning model, Henri, E. J. et al. [28] created a method for classifying Mauritius bird sounds from audio recordings. Many categorized recordings from the birdsong-sharing website Xeno-canto were utilized as input for this model. Following that, they improved three previously trained CNN models: InceptionV3, MobileNetV2, and RestNet50, as well as a brand-new model. With 84% of accuracy, transfer learning was successfully applied to develop the study's model. However, to create an effective deep-learning classifier, a substantial amount of data is needed. It is typically difficult to gather such vast amounts of data about endemic or endangered organisms. By separating two acoustic features, mainly, the Mel spectrogram and the Mel frequency cepstral coefficient, from each data point, Gunawan, K.W. et al. [29] established a transfer learning model that restricts overfitting in deep models and a method to maximize the dataset used. In order to incorporate and learn from both audio data, the researchers employed a two-input scalable convolutional neural network constructed from EfficientNet. On the test set, they had 99.9% of accuracy. A classification system for the sounds of 17 species of Indian owls was developed by Nayak S. et al. [30]. For the transfer learning model created in this study, four model architectures were used: InceptionV3, Resnet152, InceptionResnetV2, and VGG16, with all models sharing the same model parameters. The InceptionV3 network, which had an accuracy of 85.3%, produced the most precise results. ResNet50, DenseNet201, InceptionV3, Xception, and Efficient Net were just a few of the deep transfer learning models employed by Kumar Y. et al. [31] to create an intelligent system for predicting various bird species from a massive collection of audio data sets. DenseNet201 has the highest classification accuracy in the group, which was of 97.43%. A methodology for automatically classifying and

TABLE I. State of Art Studies on Automated Bioacoustics Bird Species Identification

| Author(s) | Dataset(s) | Technique(s) | Limitation(s) | Results |
|---|---|---|---|---|
| Sprengel, E. et al. (2016) | LifeCLEF plant challenge 2016 Dataset | CNN | Longer files create chunks | Accuracy: 84% |
| M Lasseck (2018) | LifeCLEF 2018 | DCNNs pre-trained on ImageNet | Results can be further enhanced by combining models with various features | Accuracy: 93% |
| Ntalampiras, S. (2018) | GTZAN corpus and http://www.Xeno-canto.org/ | Transformation based on Reservoir Networks | It is necessary to assess whether the Transfer Learning-based approach can handle feature spaces with a wide range of sizes | There were obtained 92.5 and 81.3% classification accuracy on average |
| Efremova et al. (2019) | From http://www. Xeno-canto.org: Base "SoundNet" Dataset, Target Dataset, Negative Dataset | ResNet-50 CNN | Results can be further improved | In 5-fold cross-validation, the target dataset's average validation accuracy was of 79% |
| Bai, J., et al. (2019) | BirdCLEF2019 | Inception-v3 | Ensemble of networks could significantly improve the results | The classifications mean average precision was of 0.055 (c-mAP) |
| Rahman, M. M., et al. (2020) | Seven local birds' images | MobileNet and Inception-v3 | Need to evaluate whether this model is suitable for a large number of various species | Accuracy: 91% |
| R Rajan., st al. (2021) | Xeno-canto bird sound database | VGG16 through a sliding window analysis on Mel spectrogram | The classification of multiple-label birds is a difficult undertaking because of vocalization that overlaps | Average F1-score: 0.65 |
| Henri, E. J., et al. (2021) | Xeno-canto bird sound database | InceptionV3, MobileNetV2 and RestNet50 | Misclassifications were detected in some classes | Accuracy: 84% |
| Gunawan, K. W., et al. (2021) | Xeno-canto database | Scalable with two inputs, EfficientNet's Convolutional Neural Network (CNN) | It is difficult to gather the vast amount of high-quality data on endemic or threatened animals that are required to create a powerful model | Accuracy: 99.27% |
| Nayak, S., et al. (2022) | Xeno-canto database | The ImageNet dataset was used to train the pre-trained InceptionV3 network | Need to detect the calls in poor quality audio | Accuracy: 85.3% |
| Kumar, Y., et al. (2022) | https://www.kaggle.com/c/birdsongrecognition/data | InceptionV3, Xception, ResNet50, DenseNet201, and Efficient Net | To increase the recognition rate, various noise reduction filtering must be applied during the pre-processing stage | DenseNet201 and ResNet50 classification models achieved an accuracy of 97.43% on the validation set. |
| Sharma, N., et al. (2022) | With 264 bird species, Cornell Bird Call Identification - 200 dataset offers roughly 150 recordings for each one | ResNet50V2 and EfficientNetB0 | Need to detect the calls in poor quality audio and need to remove ambient noise | EfficientNetB0 accuracy: 92.4% |

processing images and sounds to identify bird species from bird videos was presented by Sharma N. et al. [32]. On image and sound datasets containing recordings of 137 different bird species, classification models for images and sounds were developed using pre-trained neural networks ResNet 50V2 and EfficientNet B0. The final model's overall accuracy was equal to 90%, while the test accuracy for the two models was 97.1 and 92.4%, respectively.

Deep learning techniques would be a practical solution, according to the aforementioned discussion of previously developed methodologies [33], [34]. The creation of a useful classification model that optimizes performance for numerous species using transfer learning and convolutional neural networks is the major contribution of the current study. Ornithologists and other researchers are aware of the potential benefits that may come from combining developments in bioacoustics with transfer learning models, which could provide a new study dimension. Additionally, there have been a few works completed, some of which we have discussed here; nonetheless, all of those efforts have certain restrictions. Table I lists state of art studies on automated bioacoustics bird species identification by using transfer learning models. It has been noted that exceptionally lengthy files may sometimes break apart into pieces. It is essential to determine whether or not the Transfer Learning-based strategy can manage feature spaces that span a broad range of sizes. Following the deployment of transfer learning models, it was shown that the categorization of multiple-label birds might be a challenging task at times due to overlaps in their vocalizations. Additionally, misclassifications were found in certain classes. Utilizing transfer learning models has not resulted in a significant amount of additional work being done for the purpose of recognizing calls from low-quality audio. During the pre-processing step, a number of noise-reduction filtering techniques need to be employed in order to get a higher recognition rate. Our main focus of this work is to provide a technique that can identify a large number of species from their audio and also the system must be cost-effective and scalable.

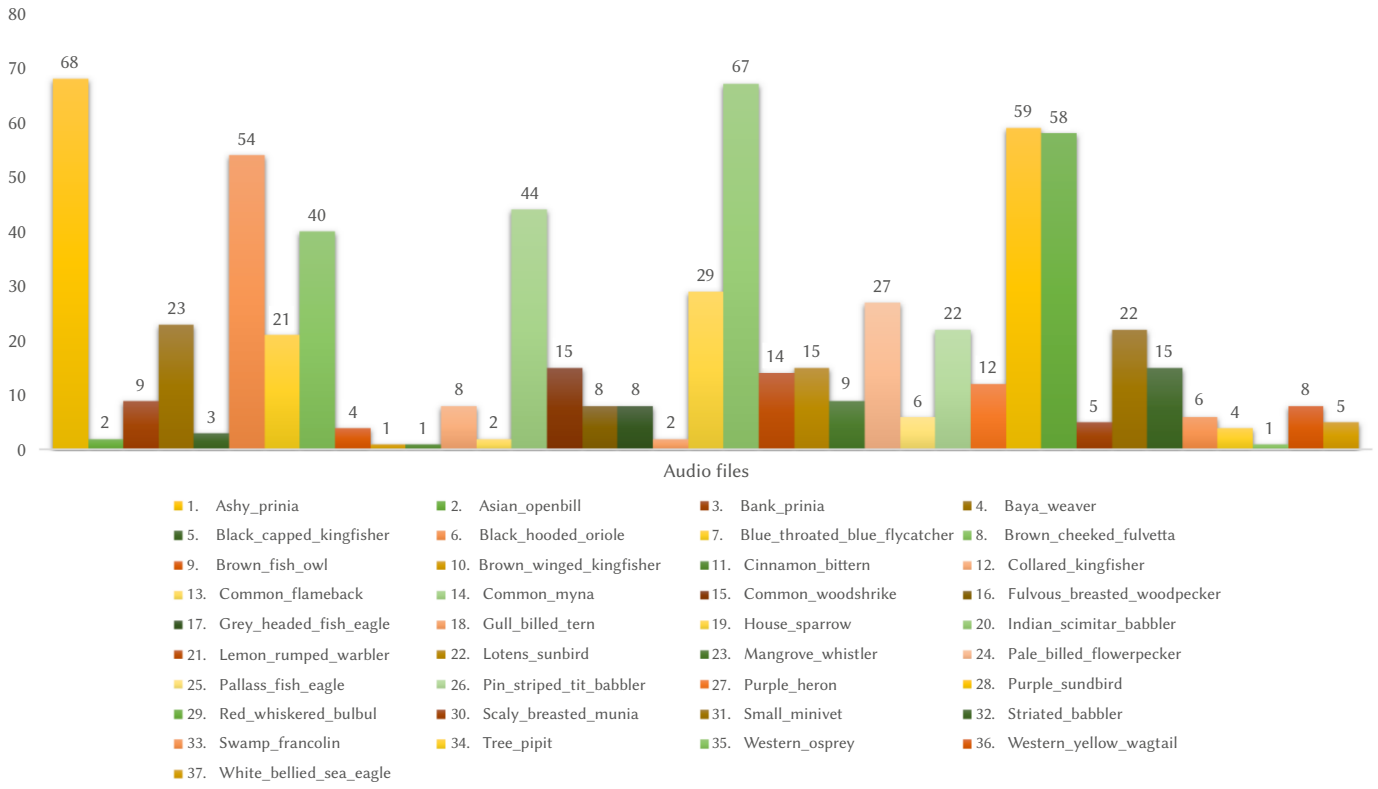


Fig. 1. List of the studied species data.

## III. Methodology

This section outlines the procedure followed in this study. The employed methodology incorporates transfer learning, deep learning, and audio-processing ideas. First, input comes from an audio recording of the bird under analysis. After that, features are extracted from the audio input using signal pre-processing techniques. The processed components are then fed into a powerful classification model that makes use of Convolutional Neural Networks [35], [36] and the idea of Transfer Learning [37], [38] to produce the best results for a wide range of species. The used three models are built on pre-trained networks called VGG16, VGG19, and InceptionV3, which were trained using data from 37 different bird species. In Fig. 2, the implementation process for this study is shown. First, from Xeno Canto, particular regional data is chosen, then data cleaning is performed, and after that data augmentation technique is used. Then, all the data is inputted into pre-trained models and classified.

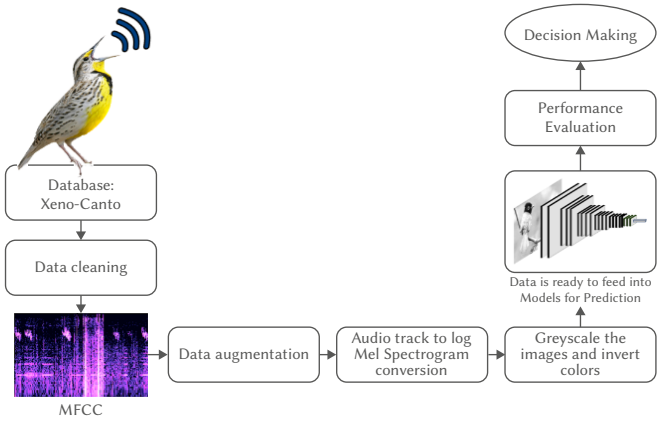


Fig. 2. Proposed bird sound identification solution.

### A. Data Collection

The widely used Xeno-canto bird sound database served as the foundation for this study's dataset. Volunteers from all across the world can record bird calls and sounds for the Xeno-canto Foundation, an online database of bird noises that includes more than a million bird sounds from more than 10,000 distinct species. Birdsong captured at Sundarban, West Bengal, India, served as the particular dataset for this study. Information on 37 different species was gathered. Fig. 1 shows the dataset utilized in this study. Each audio file was modified to contain a single vocalization lasting 1.5 seconds (sampling rate: 16000 Hz). In total, 11325 files were included. The models were trained with augmented data, which were validated using the original 453 files.

### B. Data Cleaning

Data cleaning is the process of removing inaccurate, corrupted, malformed, duplicate, or incomplete data from a dataset. There is a substantial risk of data duplication or mislabeling when merging multiple data sources. Background noise in the downloaded audio files was minor, which was confirmed manually. Hence background noise treatment was unnecessary. Parts of the audio files that had no or minimal sound were eliminated as follows: firstly, it was determined what the median sound power was, and the audio segments whose energy level or functional ability was below 50% of the median were removed, and lastly, the remaining audio files were reassembled.

### C. Feature Extraction Technique

#### 1. Mel Spectrogram

The audio sample was converted to Mel spectrogram in a different way and at different frequencies. Humans always perceive frequency logarithmically. A time-frequency representation, a perceptually appropriate amplitude representation, and ultimately a perceptually

relevant frequency representation make up ideal sound qualities. For the pitch, Mel is crucial. Convert the frequencies to the Mel scale, extract the short-time Fourier transform, and then convert the amplitude to Db.

The Mel scale conversion procedures for frequencies are:

- Determine the number of Mel Scales;
- Create banks of Mel filters;
- Use Mel filter banks for the spectrogram.

### D. Data Set Pre-Processing

Data pre-processing is the first and most crucial stage in developing a classification model. The audio classification task is an image classification challenge in this study. Here, MFCCs are employed in sound identification tasks and can accurately map auditory information in a visual domain (Fig. 3.). In order to be used, CNN models for classification audio recordings must be represented in the optical environment. Different processing is frequently required to make the dataset acceptable for usage with a CNN model [39]. The data pre-processing steps include data sizing, labeling, and expansion. The database consists of audio of the 37 birds' songs of Sundarban; among them, 19 birds' themes are included, which are very few (below 10). This significant imbalance may influence the model's performance and can lead to issues such as overfitting and difficulty learning the model.

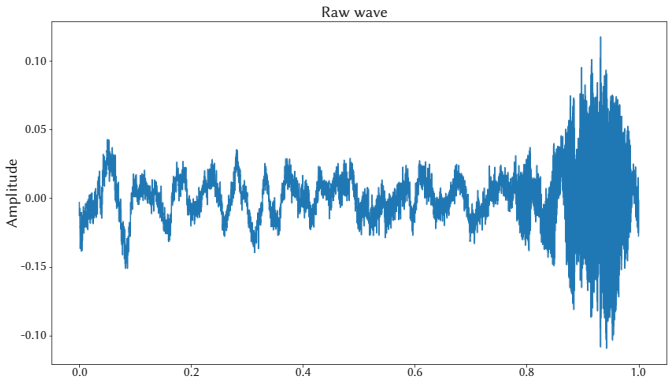


Fig. 3. Time domain representation of original audio of the ashy_prinia dataset.

### 1. Data Augmentation

A CNN model could not be used since the data collected for certain species was insufficient. Therefore, for those specific species, data augmentation was used. On the other hand, in order to prevent overfitting, data augmentation is needed. The term "data augmentation" refers to an increase in available data. Time shifting, adding noise, time stretching, and pitch augmentation is examples of audio data augmentation techniques. Time stretching, pitch scaling, and the addition of white noise were the three data augmentation methods used in this study. The aforementioned data-cleaning procedure has been applied to all used data.

#### a) Time Stretching

A method is known as "time stretching" allows one to increase the length or speed of an audio stream without changing its pitch or other parameters. For example, one can extend a sound to 200 milliseconds by decoding twice as many samples from each frame if uttered for 100 milliseconds (10 frames) [40]. Librosa, a python utility for music modification, applies the time stretching simple. The rate settings can change the audio's pace and duration. Fig. 4 represents the time stretching of 0.8 times of original audio.

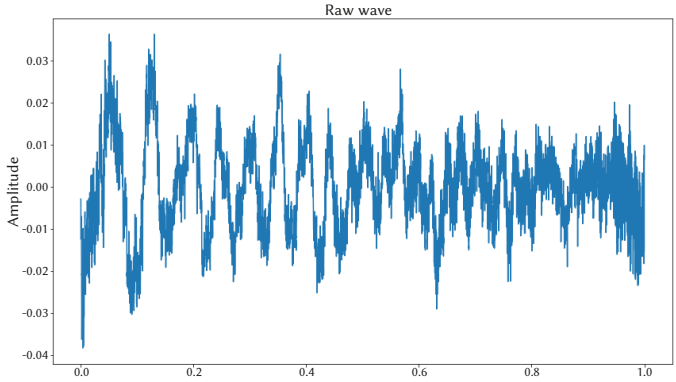


Fig. 4. Time Stretching of an original bird call audio.

#### b) Pitch Scaling

This technique serves as a wrapper for the librosa function. The pitch veers all over the place. When applying different rate values without altering the duration of the signal, pitch scaling is the reverse of time stretching [41], as can be seen in Fig. 5.

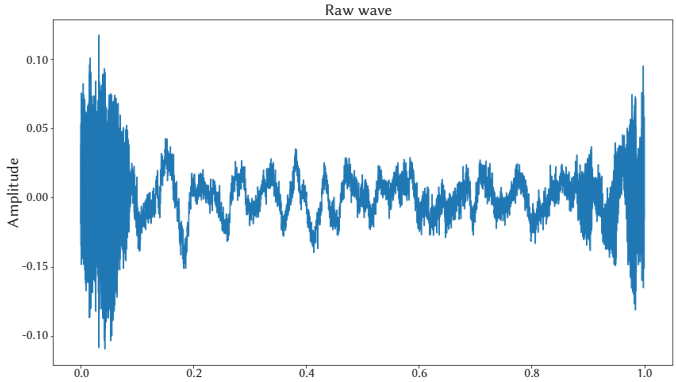


Fig. 5. Pitch Scaling of an original bird call audio.

#### c) Noise Addition

Noise addition can generate syntactic audio data for the data augmentation process. Numpy makes it simple to deal with noise addition by adding a random value to the date. In Fig. 6, this technique can be seen.

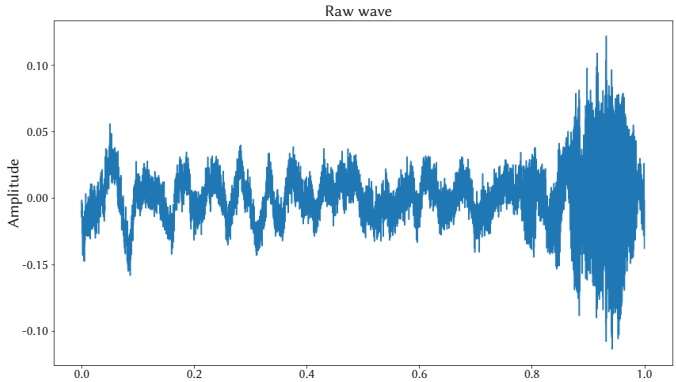


Fig. 6. Noise addition of an original bird call audio.

### 2. Dataset Splitting: Training & Testing

The total number of files that were selected from Sundarbans's set consisted of 2265; after doing data augmentation, the total number of files was 11325. Then, the used dataset was split into 80% and 20%, for training and testing, respectively.

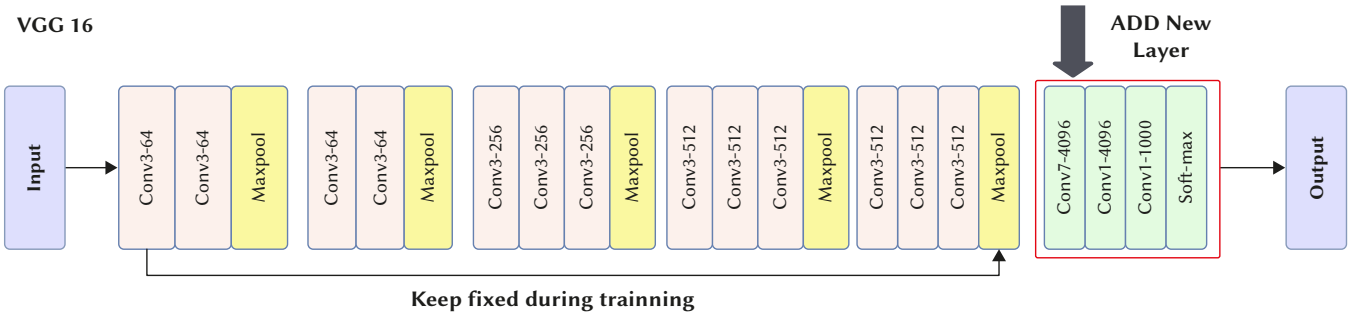


Fig. 7. VGG16 model architecture.

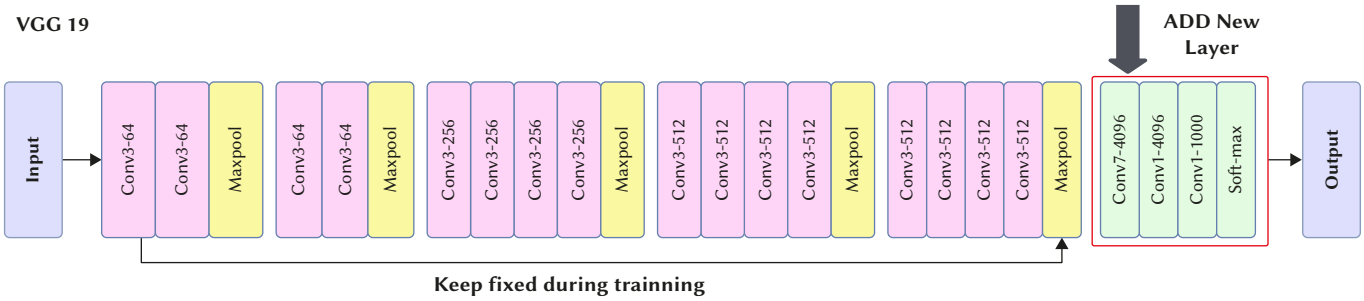


Fig. 8. VGG19 model architecture.

### *E. Model Description*

The use of deep learning in audio recognition is well-recognized. Neural networks have been applied to numerous facets of audio recognition since the development of deep understanding [42], [43]. The effectiveness of neural learning for sound recognition is influenced by the adaptability and predictive power of the increasingly accessible deep neural networks. The deep learning models utilized in the study are described in the following.

#### *1. VGG16*

VGGNet-16 has a relatively homogeneous architecture with 16 convolutional layers. It only has 3x3 convolutions but a lot of filters [44]. The Visual Geometry Organization, or VGG for short, was a group that replaced Alex Net which was established in Oxford. It adopts and enhances some concepts from its forerunners and uses deep convolutional neural layers to increase accuracy. Comparatively, managing VGGNet's 138 million parameters can be challenging.

VGG16 has thirteen convolutional layers, five Max Pooling layers, and three Dense layers for a total of twenty-one layers, but only sixteen weight layers or trainable parameters layers [45]. Each of the 16 layers has one convolution and one pooling layer, Fig. 7. VGG16 can be enhanced through transfer learning.

Following the rectified linear unit (ReLu) activations, the image data is transmitted through the first of two convolutional layers with a minimum receiving area of 3X3. In each of these two layers, there are 64 filters. One pixel serves as padding, while one pixel always serves as the convolution step. The first convolutional layer is responsible for capturing low-level information such as gradient and edge orientation, among other information. The spatial maxima are then binned with a step of 2 pixels in a 2x2 pixel window for activation maps. An activation's size is cut in half. Consequently, the activations at the base of the first stack are 112x112x64 long. The activations then proceed via the 128 filters in the second stack as opposed to the 64 in the first one.

The size is 56x56x128 as a result after the second layer. A maximum pool layer and three convolutional layers make up the third layer. Because 256 filters are employed, the output stack size is 28x28x256. The following two sets of three convolutional layers have each 512 filters. The final stack is of 7x7x512 size for both. Following stacks of convolutional layers with a flattened layer in between are the three fully connected layers. The last completely connected layer serves as the output layer, and has 1000 neurons, or 1000 potential classifications of the ImageNet dataset. The previous two fully connected layers have each 4096 neurons. The SoftMax activation layer, which is utilized for categorization, comes after the output layer. In order to adapt the architecture to high-level characteristics, additional layers are also helpful. The spatial size of the convolved feature is decreased by the pooling layer. The amount of processing power needed to process the data lowers as its dimension increases. Smooth training is made possible by the VGG16 model, which is useful for extracting rotation- and position-invariant dominating features.

#### *2. VGG19*

A 19-layer version of the VGG model is known as the VGG19 model, which has 16 convolution layers, three fully connected layers, 5 Max Pool layers, and 1 SoftMax layer, Fig. 8 [46]. An RGB image of fixed size (224*224) was provided to this network as input, indicating that the matrix was of the form (224,224,3). The only preprocessing was to take the mean RGB value for the entire training set and subtract it from each pixel [47]. The complete visual concept was then covered using kernels of size (3*3) with a step size of 1 (one) pixel. Spatial padding was then applied to preserve the spatial resolution of the image. Step two was then used to create maximum pooling in two * 2-pixel windows. Then, instead of using tanh or sigmoid functions, a ReLu was used to induce non-linearity and improve processing speed. Three final connected layers are then implemented, the first two of which are 4096 in size, followed by a 1000-channel ILSVRC classification layer, and finally a SoftMax activation layer, which is used for category classification. It has been used as a good classification architecture for various other datasets. The models were publicly available, so they can be used as is or with minor modifications for other similar work.

#### *3. InceptionV3*

Convolutional neural networks are the foundation of the deep learning model known as InceptionV3, which was first developed as a Google network module for image analysis and object detection.

Inception Networks (Google Net/Inception v1) are more cost- and time-effective computationally than VGGNet in terms of the number of network parameters produced. It has 42 layers and a lower error rate than previous models, Fig. 9. To improve model adaptation, the InceptionV3 model uses a number of mesh optimization strategies.

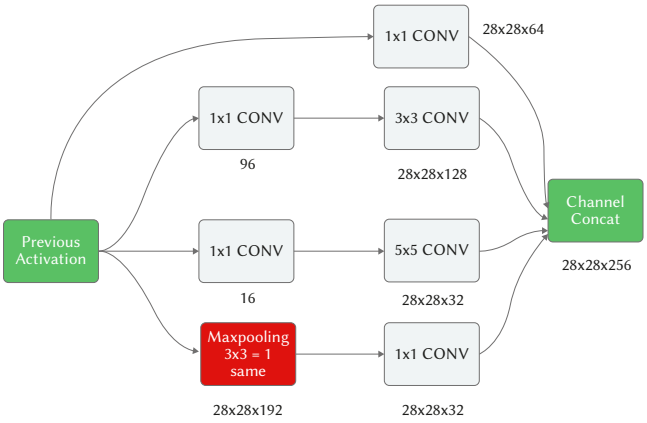


Fig. 9. Layers used in the InceptionV3 model.

The used approaches are factorized convolution, regularization, dimensionality reduction, and parallelized calculations [48]. The number of parameters in the network is decreased via factorized convolutions, which enhances computational effectiveness. It also benefits the network performance. Training becomes faster as smaller convolutions take the place of bigger ones. For instance, replacing a 5 5 convolution with two 3 3 filters only requires 18 (3*3+3*3) parameters. In asymmetric convolutions, a 3 3 convolution can be swapped out for a 1 3 convolution followed by a 3 1 convolution. If the 3 3 convolutions were switched out for a 2 2, there would be a lot more parameters than in the case of the described asymmetric convolution. The network suffers a considerable loss as a result of the losses caused by the little CNN that was added between the layers during training. In InceptionV3, a third classifier acts as a regularization term. Last but not least, pooling procedures are frequently used to achieve a grid size reduction strategy. The final building incorporates all of the principles previously mentioned. The InceptionV3 was used in this study because, while not slower than the Inception V1 and V2 models, it is more effective and has a deeper network [49]. The InceptionV3 model is less expensive to calculate.

In Fig. 10, the proposed customized model is shown. First, the model was built with a standard structure, and later it was fine-tuned for respective models. Methodologies such as feature extraction, data augmentation, and three transfer learning models were used for the comparison purpose in this study. As because of the transfer learning concept is employed therefore there are no overfitting issues with the model. First, input comes from an audio recording of the bird under analysis. After that, features are extracted from the audio input using signal pre-processing techniques. After that, the data augmentation task is accomplished. The processed components are then fed into a powerful classification model and the idea of Transfer Learning to produce the best results for a wide range of species. The used three models are built on pre-trained networks called VGG16, VGG19, and InceptionV3, which were trained using data from 37 different bird species.

In the proposed VGG16 model, there are five convolutions' blocks. Each block contains a convolution 2D model and max-pooling 2D layer. The input of the model is 224, 224 with three dimensions; after one complete convolution, the output size is (112, 112,64). Following another convolution, the output is (56,56,120) after block three, (28,28,256) in partnership four, (14,14,512) in partnership five, and (7,7,512) as input and output are 512 in partnership six. The included dropout layer has a very slight change, and the final dense layer has 256 as an input and 37 as an output. Generally, all the layers of VGG16 were frozen and a customized layer was added. The Sigmoid function is used as an activation function and optimizer. As a loss function, an Adam optimizer with cross-entropy was used. For the VGG19 model, ImageNet was used as a weight, and average pooling, a customized base layer, and convolution layers with 256 dense layers with activation function as ReLu with dropout 0.1 were used. Additionally, SoftMax with a learning rate of 0.00005 was employed in the final layer. During the course of the model-building procedure, the Adam optimizer and the loss function were used as the categorical cross-entropy. Lastly, for the InceptionV3 model, ImageNet was used as a weight, and average pooling was employed. Lastly, a customized model was built by adding custom layers. In the customized model, 256 dense layers with an activation function ReLU were used, and a dropout of 0.4 was used. With the Adam optimizer, the model was built using categorical cross-entropy as the loss function.

## IV. Experimental Results & Analysis

We have experimentally chosen three transfer learning models in this study: VGG16, VGG19, and InceptionV3 model. Table II, Table III,

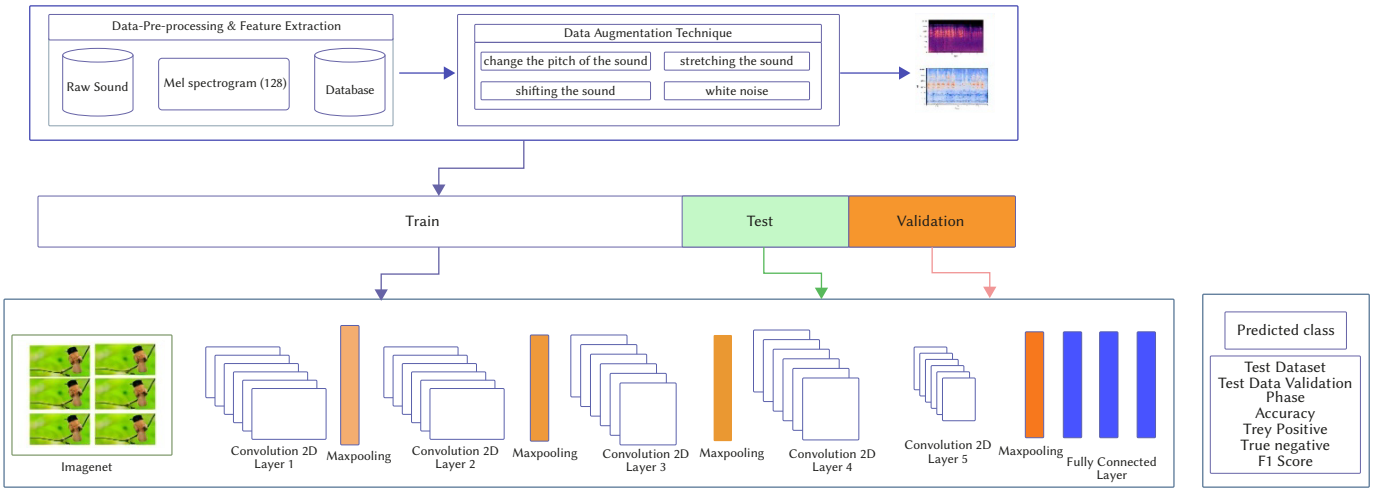


Fig. 10. Proposed deep learning model.

and Table IV show the individual performance of the VGG16, VGG19, and InceptionV3 models respectively. From these three tables, it is observed that the performance of the proposed InceptionV3 model shows better performance when it is compared with the VGG16 and VGG19 models. The experimental results of the InceptionV3 model are reported as 86% precision, 86% of recall, and 85% of F1-score. The classification results of VGG16 are as follows: 18 out of 37 bird sounds: ashy_prinia, brown_fish_owl, brown_winged_kingfisher, cinnamon bittern, collared kingfisher, common_woodshrike, fulvous_breasted_woodpecker, grey_headed_fish_eagle, lotens_sunbird, red_whiskered_bulbul, striated_babbler, swamp_francolin, tree_pipit, western_osprey, asian_openbill, baya_weaver, brown_cheeked_fulvetta, and western_yellow_wagtail, were 100% detected from the test data. The overall accuracy achieved using the VGG16 model was equal to 78%. For the VGG19 model, the training accuracy obtained was of 64%, and the test accuracy was 61.9%. According to the categorization results, VGG19 obtained 100% of recognition in 17 of the 37 test cases. In the InceptionV3 model with a batch size of 32, the obtained train accuracy was 95%, and the test accuracy of 85%. As to the classification results, 24 of the 37 species were 100% detected in the test dataset.

### A. Accuracy

TABLE II. VGG16 Classification Model

| Class | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.80 | 0.29 | 0.42 | 14 |
| 1 | 1.00 | 1.00 | 1.00 | 12 |
| 2 | 0.73 | 1.00 | 0.85 | 11 |
| 3 | 0.67 | 0.57 | 0.62 | 14 |
| 4 | 1.00 | 1.00 | 1.00 | 14 |
| 5 | 0.50 | 0.55 | 0.52 | 11 |
| 6 | 0.42 | 0.62 | 0.50 | 13 |
| 7 | 0.71 | 0.62 | 0.67 | 16 |
| 8 | 1.00 | 1.00 | 1.00 | 12 |
| 9 | 1.00 | 1.00 | 1.00 | 10 |
| 10 | 1.00 | 1.00 | 1.00 | 10 |
| 11 | 0.93 | 1.00 | 0.96 | 13 |
| 12 | 0.87 | 1.00 | 0.93 | 13 |
| 13 | 0.80 | 0.67 | 0.73 | 18 |
| 14 | 0.90 | 0.75 | 0.82 | 12 |
| 15 | 100 | 0.85 | 0.92 | 13 |
| 16 | 1.00 | 1.00 | 1.00 | 13 |
| 17 | 1.00 | 1.001 | 100 | 11 |
| 18 | 0.50 | 0.25 | 0.33 | 12 |
| 19 | 0.70 | 0.54 | 0.61 | 13 |
| 20 | 0.62 | 0.45 | 0.53 | 11 |
| 21 | 0.56 | 0.42 | 0.48 | 12 |
| 22 | 1.00 | 1.00 | 1.00 | 11 |
| 23 | 0.37 | 0.64 | 0.47 | 11 |
| 24 | 0.92 | 1.00 | 0.96 | 12 |
| 25 | 1.00 | 0.46 | 0.63 | 13 |
| 26 | 0.92 | 0.92 | 0.92 | 12 |
| 27 | 0.50 | 0.42 | 0.45 | 12 |
| 28 | 0.37 | 0.64 | 0.47 | 11 |
| 29 | 1.00 | 1.00 | 1.00 | 10 |
| 30 | 0.45 | 0.77 | 0.57 | 13 |
| 31 | 0.71 | 1.00 | 0.83 | 12 |
| 32 | 1.00 | 1.00 | 1.00 | 12 |
| 33 | 1.00 | 1.00 | 1.00 | 11 |
| 34 | 1.00 | 1.00 | 1.00 | 10 |
| 35 | 0.92 | 0.85 | 0.88 | 13 |
| 36 | 1.00 | 1.00 | 1.00 | 10 |
| accuracy | | | 0.78 | 451 |
| macro avg | 0.81 | 0.79 | 0.78 | 451 |
| weighted avg | 0.80 | 0.78 | 0.78 | 451 |

TABLE III. VGG19 Classification Model

| Class | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.33 | 0.36 | 0.34 | 14 |
| 1 | 1.00 | 1.00 | 1.00 | 12 |
| 2 | 1.00 | 0.36 | 0.53 | 11 |
| 3 | 0.26 | 0.36 | 0.30 | 14 |
| 4 | 1.00 | 1.00 | 1.00 | 14 |
| 5 | 0.50 | 0.18 | 0.27 | 11 |
| 6 | 0.67 | 0.15 | 0.25 | 13 |
| 7 | 0.39 | 0.44 | 0.41 | 16 |
| 8 | 0.60 | 1.00 | 0.75 | 12 |
| 9 | 0.91 | 1.00 | 0.95 | 10 |
| 10 | 1.00 | 1.00 | 1.00 | 10 |
| 11 | 0.62 | 1.00 | 0.76 | 13 |
| 12 | 0.93 | 1.00 | 0.96 | 13 |
| 13 | 0.67 | 0.22 | 0.33 | 18 |
| 14 | 0.33 | 0.67 | 0.44 | 12 |
| 15 | 0.86 | 0.46 | 0.60 | 13 |
| 16 | 0.92 | 0.85 | 0.88 | 13 |
| 17 | 1.00 | 1.00 | 1.00 | 11 |
| 18 | 0.40 | 0.17 | 0.24 | 12 |
| 19 | 0.50 | 0.31 | 0.38 | 13 |
| 20 | 0.44 | 0.64 | 0.52 | 11 |
| 21 | 0.52 | 0.92 | 0.67 | 12 |
| 22 | 0.91 | 0.91 | 0.91 | 11 |
| 23 | 0.33 | 0.36 | 0.35 | 11 |
| 24 | 0.57 | 1.00 | 0.73 | 12 |
| 25 | 0.15 | 0.15 | 0.15 | 13 |
| 26 | 0.67 | 0.83 | 0.74 | 12 |
| 27 | 0.43 | 0.25 | 0.32 | 12 |
| 28 | 0.00 | 0.00 | 0.00 | 11 |
| 29 | 0.77 | 1.00 | 0.87 | 10 |
| 30 | 1.00 | 0.23 | 0.38 | 13 |
| 31 | 0.50 | 0.58 | 0.54 | 12 |
| 32 | 0.71 | 1.00 | 0.83 | 12 |
| 33 | 0.86 | 0.55 | 0.67 | 11 |
| 34 | 1.00 | 1.00 | 1.00 | 10 |
| 35 | 0.50 | 0.54 | 0.52 | 13 |
| 36 | 0.53 | 1.00 | 0.69 | 10 |
| accuracy | | | 0.62 | 451 |
| macro avg | 0.64 | 0.63 | 0.6 | 451 |
| weighted avg | 0.64 | 0.62 | 0.59 | 451 |

In order to meaningfully evaluate a machine learning model's performance, accuracy is a metric frequently used. A model's accuracy is usually calculated once the parameters are specified and represented in terms of percentage, which is a statistic that shows how accurately the model's performance contrasts with actual data. Figs. 11, 12, and 13 show the accuracy curves for the built learning models. In the experiment using the baseline model of VGG16, only ten epochs with a batch size 32 were run, and the obtained train accuracy was of 75% and the test one was 78%. Similarly, the VGG19 model also run in 10 periods with a batch size of 32, and 64% was train accuracy and 61.9% the test one. On the InceptionV3 model, which was used in further experiments, with 10 epochs, a training accuracy of 95% and a test accuracy of 85% were obtained.

TABLE IV. INCEPTIONV3 CLASSIFICATION MODEL

| Class | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.46 | 0.43 | 0.44 | 14 |
| 1 | 1.00 | 1.00 | 1.00 | 12 |
| 2 | 0.85 | 1.00 | 0.92 | 11 |
| 3 | 0.88 | 1.00 | 0.93 | 14 |
| 4 | 0.93 | 1.00 | 0.97 | 14 |
| 5 | 0.35 | 0.55 | 0.43 | 11 |
| 6 | 0.90 | 0.69 | 0.78 | 13 |
| 7 | 0.70 | 0.88 | 0.78 | 16 |
| 8 | 1.00 | 1.00 | 1.00 | 12 |
| 9 | 1.00 | 1.00 | 1.00 | 10 |
| 10 | 1.00 | 1.00 | 1.00 | 10 |
| 11 | 1.00 | 1.00 | 1.00 | 13 |
| 12 | 1.00 | 1.00 | 1.00 | 13 |
| 13 | 0.91 | 0.56 | 0.69 | 18 |
| 14 | 1.00 | 1.00 | 1.00 | 12 |
| 15 | 1.00 | 1.00 | 1.00 | 13 |
| 16 | 0.93 | 1.00 | 0.96 | 13 |
| 17 | 0.92 | 1.00 | 0.96 | 11 |
| 18 | 0.75 | 0.50 | 0.60 | 12 |
| 19 | 0.78 | 0.54 | 0.64 | 13 |
| 20 | 0.92 | 1.00 | 0.96 | 11 |
| 21 | 0.86 | 1.00 | 0.92 | 12 |
| 22 | 0.79 | 1.00 | 0.88 | 11 |
| 23 | 0.83 | 0.91 | 0.87 | 11 |
| 24 | 1.00 | 1.00 | 1.00 | 12 |
| 25 | 0.71 | 0.77 | 0.74 | 13 |
| 26 | 1.00 | 1.00 | 1.00 | 12 |
| 27 | 0.43 | 0.25 | 0.32 | 12 |
| 28 | 0.30 | 0.27 | 0.29 | 11 |
| 29 | 0.91 | 1.00 | 0.95 | 10 |
| 30 | 0.90 | 0.69 | 0.78 | 13 |
| 31 | 1.00 | 1.00 | 1.00 | 12 |
| 32 | 1.00 | 1.00 | 1.00 | 12 |
| 33 | 0.92 | 1.00 | 0.96 | 11 |
| 34 | 1.00 | 1.00 | 1.00 | 10 |
| 35 | 1.00 | 1.00 | 1.00 | 13 |
| 36 | 0.91 | 1.00 | 0.95 | 10 |
| accuracy | | | 0.86 | 451 |
| macro avg | 0.86 | 0.87 | 0.86 | 451 |
| weighted avg | 0.86 | 0.86 | 0.85 | 451 |

## B. Loss

A more accurate model is indicated by lower loss values. The loss is not expressed as a percentage, in contrast to accuracy. The built learning models' loss curves are shown in Figs. 14, 15 and 16. The training loss of the VGG16, VGG19, and InceptionV3 models decreased over time, but the validation data revealed frequent variations and substantial loss. The loss function shown was in the 0.9 to 1.5 in range in the three studied models. In the training of the studied models, categorical_crossentrophy was used.

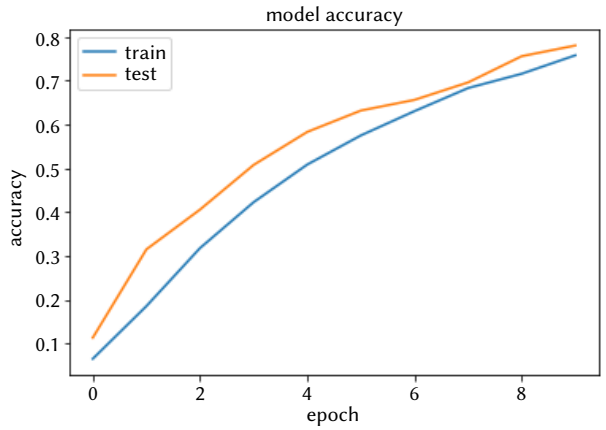


Fig. 11. VGG16 model's accuracy.

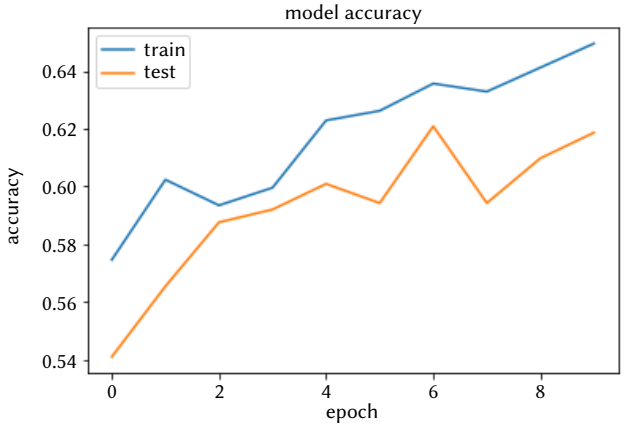


Fig. 12. VGG19 model's accuracy.

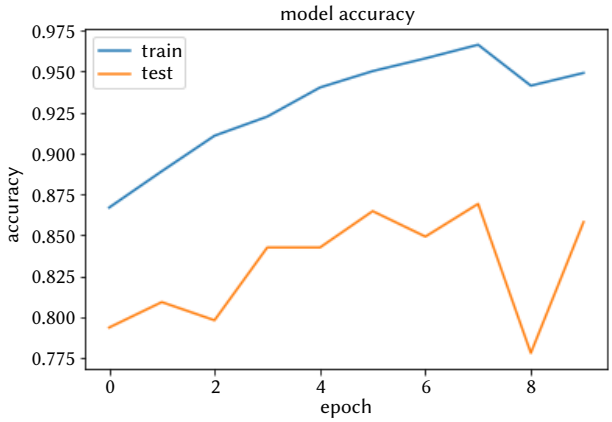


Fig. 13. INCEPTIONV3 model's accuracy.

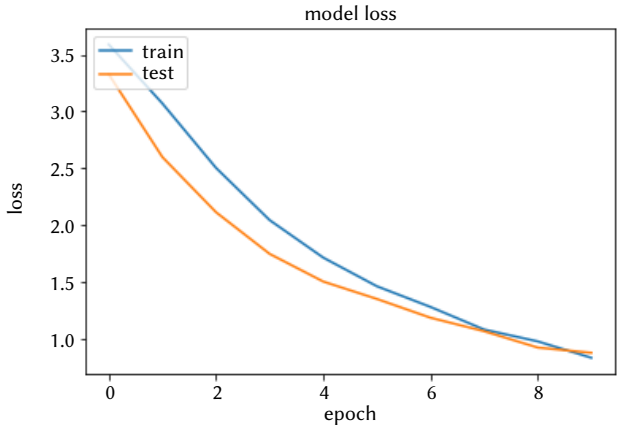


Fig. 14. VGG16 model's loss.

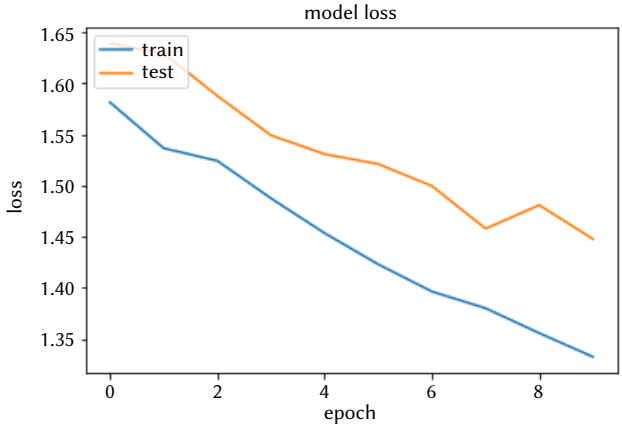


Fig. 15. VGG19 model's loss.

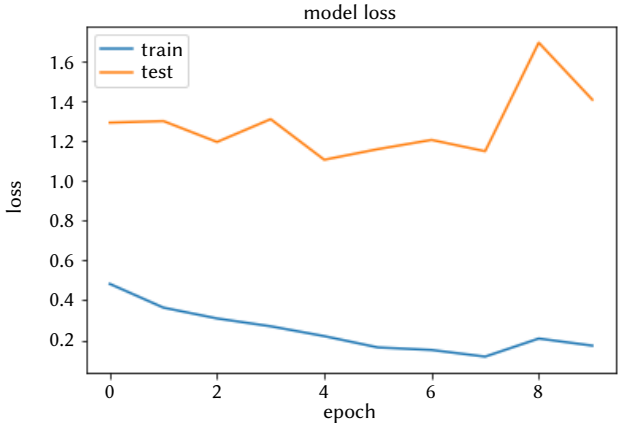


Fig. 16. INCEPTIONV3 model's loss.

## C. Confusion Matrix

An evaluation of the performance of a classification model, or "classifier", on a set of test data for which the true values are known is given by a confusion matrix, which is a table. The matrix also allows a comparison between the targets' actual values and the model projections. To properly comprehend the classification findings, the confusion matrix for each of the three classification architectures were built, Figs. 17, 18 and 19.

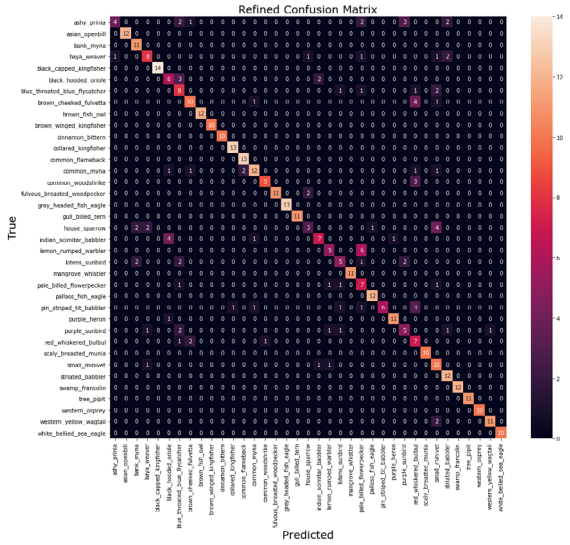


Fig. 17. Confusion matrix obtained by VGG16.

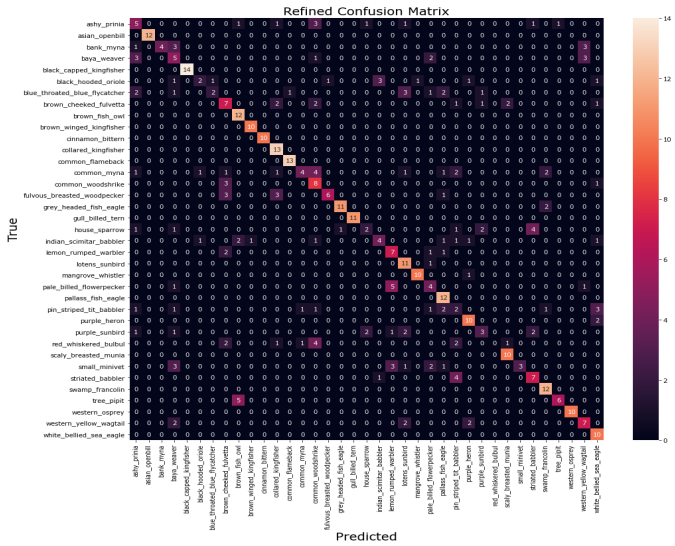


Fig. 18. Confusion matrix obtained by VGG19.

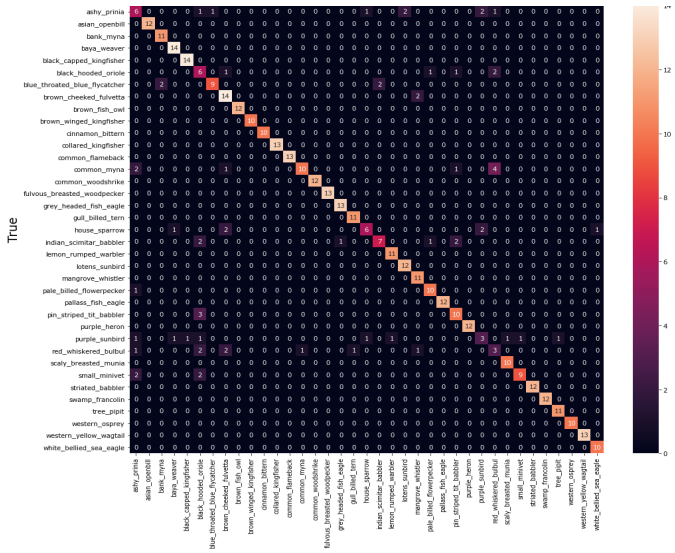


Fig. 19. Confusion matrix obtained by INCEPTIONV3.

## V. Discussion

Working with bird species of a more significant number of types is challenging. Deep learning architectures have improved speech recognition accuracy, and automated learning approaches have been developed. Transfer learning technique was used in this study as 37 bird species were addressed, and a large dataset with a wide range of bird sounds is required. In this research, the categorization of bird noises is accomplished via the utilization of three different deep-learning frameworks. These frameworks are VGG16, VGG19, and InceptionV3. All the models use the same model parameters. As was shown, the InceptionV3 model obtained the best result. However, M Lasseck et al. [50] showed 93% accuracy using ensemble models with deep convolution neuronal networks with a pre-trained model but using a more significant number of epochs. The proposed model outperforms the solutions proposed by earlier work that was carried out by other researchers. Previously, various models gave a visual representation of the sound, but the suggested model is capable of working directly with the unprocessed audio file. According to another finding, the InceptionV3 model performs better than the other two models in this regard. In addition, acoustic properties were gathered from bird calls and were classified using various feature extraction techniques. It has been demonstrated that the proposed strategy is capable of boosting prediction accuracy. A novel method for identifying a large number of bird species in the Sundarban region of West Bengal, India was devised using existing recordings of their sounds.

## VI. Conclusion

The suggested model may be put into low-cost devices via the use of a technique that is both cost-effective and scalable; hence, more devices can be employed to cover more land. In this experiment, it was shown that a transfer-learned network that had previously been trained on ImageNet shows a better predictive capability and accelerates convergence when compared with the same network architecture that is trained from scratch. The experiment was conducted in order to demonstrate this. When there are a limited number of high-quality datasets available, it is advantageous to utilize a model that has already been pre-trained because of the benefits it provides. In terms of practical uses, the suggested approach may be of great use to ornithologists by making the identification of bird species a straightforward process, In the future, in order to enhance the recognition rate, we want to use a variety of noise reduction filtering techniques during the pre-processing step. In addition, another problem that should be addressed is the overlapping of sounds.

## References


[1] M. A. Tabur and Y. Ayvaz, "Ecological importance of birds," in Second International Symposium on Sustainable Development Conference, 2010, Jun., pp. 560-565.

[2] S. D. H. Permana et al., "Classification of bird sounds as an early warning method of forest fires using Convolutional Neural Network (CNN) algorithm," Journal of King Saud University - Computer and Information Sciences. Inf. Sci., 2021.

[3] G. F. Budney and R. W. Grotke, "Techniques for audio recording vocalizations of tropical birds," Ornithological Monographs, no. 48, pp. 147-163, 1997, doi:10.2307/40157532.

[4] Available at: https://www.environmentalscience.org/birds-environmental-indicators (last access date: 18/18/2022).

[5] Available at: https://www.ck12.org/biology/bird-ecology/lesson/Importance-of-Birds-MS-LS/ (last access date: 18/12/2022).

[6] Available at: https://www.thespruce.com/bird-courtship-behavior-386714 (last access date: 18/12/2022).

[7] Available at: https://www.birdlife.org/worldwide/news/why-we-need-birds-far-more-they-need-us (last access date: 18/12/2022).

[8] S. Fagerlund, "Bird species recognition using support vector machines," EURASIP Journal on Advances in Signal Processing, vol. 2007, no. 1, pp. 1-8, 2007, doi:10.1155/2007/38637.

[9] N. Das et al., Machine Learning Models for Bird Species Recognition Based on Vocalization: A Succinct Review. Information Technology and Intelligent Transportation Systems, 2020, pp. 117-124.

[10] C. Yüksel, 2020, Bird call detection using deep learning (Master's thesis, Fen Bilimleri Enstitüsü).

[11] S. Bhattacharya et al., "Deep classification of sound: A concise" in Proceeding of First Doctoral Symposium on Natural Computing Research: DSNCR 2020, vol. 169. Springer Nature, 2021, Mar.

[12] N. Das et al., "Building of an edge-enabled drone network ecosystem for bird species identification," Ecological Informatics, vol. 68, p. 101540, 2022, doi: 10.1016/j.ecoinf.2021.101540.

[13] S. Bhattacharya et al., "Deep analysis for speech emotion recognization" in Second International Conference on Computer Science, Engineering and Applications (ICCSEA), vol. 2022. IEEE, 2022, Sept., pp. 1-6, doi:10.1109/ICCSEA54677.2022.9936080.

[14] K. Lan et al., "A survey of data mining and deep learning in bioinformatics," Journal of Medical Systems, vol. 42, no. 8, pp. 139, 2018, doi:10.1007/s10916-018-1003-9.

[15] Y. Wu et al., "Learning models for semantic classification of insufficient plantar pressure images," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 1, pp. 51-61, 2020, doi:10.9781/ijimai.2020.02.005.

[16] H. Chang et al., "Unsupervised transfer learning via multi-scale convolutional sparse coding for biomedical applications," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 5, pp. 1182-1194, 2018, doi:10.1109/TPAMI.2017.2656884.

[17] R. Wald et al., "Hidden dependencies between class imbalance and difficulty of learning for bioinformatics datasets" in 14th International Conference on Information Reuse & Integration (IRI), vol. 2013. IEEE. IEEE, 2013, Aug., pp. 232-238, doi:10.1109/IRI.2013.6642477.

[18] C. Tan et al., "A survey on deep transfer learning" in International conference on artificial neural networks. Cham: Springer, 2018, Oct., pp. 270-279.

[19] L. Muda et al., 2010, Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. arXiv preprint arXiv:1003.4083.

[20] C. Y. Koh et al., 2019, Sept., "Bird sound classification using convolutional neural networks" in Clef [Working notes].

[21] A. Fritzler et al., 2017, "Recognizing bird species in audio files using transfer learning" in Clef [Working notes].

[22] S. Ntalampiras, "Bird species identification via transfer learning from music genres," Ecological Informatics, vol. 44, pp. 76-81, 2018, doi: 10.1016/j.ecoinf.2018.01.006.

[23] D. B. Efremova et al., "Data-efficient classification of birdcall through convolutional neural networks transfer learning" in Digital Image Computing: Techniques and Applications (DICTA), vol. 2019. IEEE, 2019, Dec., pp. 1-8, doi:10.1109/DICTA47822.2019.8946016.

[24] J. Bai et al., 2019, "Inception-v3 based method of LifeCLEF," vol. 2019 Bird Recognition in Clef [Working notes].

[25] M. Zhong et al., "Multispecies bioacoustic classification using transfer learning of deep convolutional neural networks with pseudo-labeling," Applied Acoustics, vol. 166, p. 107375, 2020, doi: 10.1016/j.apacoust.2020.107375.

[26] A. K. Ibrahim et al., "Transfer learning for efficient classification of grouper sound," Journal of the Acoustical Society of America, vol. 148, no. 3, pp. EL260, 2020, doi:10.1121/10.0001943.

[27] R. Rajan and A. Noumida, "Multi-label bird species classification using transfer learning" in 2021 International Conference on Communication, Control and Information Sciences (ICCISc), vol. 1. IEEE, 2021, Jun., doi:10.1109/ICCISc52257.2021.9484858.

[28] E. J. Henri and Z. Mungloo-Dilmohamud, "A deep transfer learning model for the identification of bird songs: A case study for Mauritius" in International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), vol. 2021. IEEE, 2021, Oct., pp. 1-6, doi:10.1109/ICECCME52200.2021.9590917.

[29] K. W. Gunawan et al., "A transfer learning strategy for owl sound

classification by using image classification model with audio spectrogram," International Journal on Electrical Engineering and Informatics, vol. 13, no. 3, pp. 546-553, 2021, doi:10.15676/ijeei.2021.13.3.3.

[30] S. Nayak et al., "Whose hoot? Identification of owl species using call recognition with neural networks,", SSRN Journal, 2022, doi:10.2139/ssrn.4020038.

[31] N. Sharma et al., "Automatic identification of bird species using audio/video processing" in International Conference for Advancement in Technology (ICONAT), vol. 2022. IEEE, 2022, Jan., pp. 1-6, doi:10.1109/ICONAT53423.2022.9725906.

[32] Y. Kumar et al., "A novel deep transfer learning models for recognition of birds sounds in different environment," Soft Computing, pp. 1-14, 2022.

[33] S. Bhattacharya et al., "Emotion detection from multilingual audio using deep analysis," Multimedia Tools and Applications , pp. 1-30, 2022.

[34] E. Sprengel et al., 2016, Audio-based bird species identification using deep learning techniques (No. CONF, pp. 547-559).

[35] E. Cakir et al., "Convolutional recurrent neural networks for bird audio detection," 25th European Signal Processing Conference EUSIPCO, vol. 2017, 2017. 2017-Janua, pp. 1744-1748, doi:10.23919/EUSIPCO.2017.8081508.

[36] J. Kim et al., "Acoustic classification of mosquitoes using convolutional neural networks combined with activity circadian rhythm information,", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 7, no. 2, 2021, doi:10.9781/ijimai.2021.08.009.

[37] S. Ahuja et al., "Deep transfer learning-based automated detection of COVID-19 from lung CT scan slices," Applied intelligence (Dordrecht, Netherlands), vol. 51, no. 1, pp. 571-585, 2021, doi:10.1007/s10489-020-01826-w.

[38] M. Singh et al., "Transfer learning–based ensemble support vector machine model for automated COVID-19 detection using lung computerized tomography scan data," Medical & Biological Engineering & Computing, vol. 59, no. 4, pp. 825-839, 2021, doi:10.1007/s11517-020-02299-2.

[39] D. A. Pitaloka et al., "Enhancing CNN with preprocessing stage in automatic emotion recognition," Procedia Computer Science, vol. 116, pp. 523-529, 2017 [doi:10.1016/j.procs.2017.10.038].

[40] M. Morrison et al., 2021, Neural pitch-shifting and time-stretching with controllable LPCNet. arXiv preprint arXiv:2110.02360.

[41] P. B. Baptista and C. Antunes, "Bioacoustic classification framework using transfer learning," Model Decision Artificial Intelligence, vol. 35, 2021.

[42] A. Bhaik et al., "Detection of improperly worn face masks using deep learning-A preventive measure against the spread of COVID-19," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 7, no. 7, 2021, doi:10.9781/ijimai.2021.09.003.

[43] K. He et al., "Deep residual learning for image recognition" in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2016, vol. 7, no. 3, 2016, pp. 770-778, doi:10.1109/CVPR.2016.90

[44] Available at: https://iq.opengenus.org/vgg16/ [Last Access Date: 18.12.2022].

[45] S. K. Rahut et al., "Bengali abusive speech classification: A transfer learning approach using" VGG-16 in Emerging Technology in Computing, Communication and Electronics (ETCCE), vol. 2020. IEEE, 2020, Dec., pp. 1-6.

[46] A. Ashurov et al., "Environmental sound classification based on transfer-learning techniques with multiple optimizers," Electronics, vol. 11, no. 15, p. 2279, 2022 [doi:10.3390/electronics11152279].

[47] M. J. Horry et al., "COVID-19 detection through transfer learning using multimodal imaging data," IEEE Access, vol. 8, pp. 149808-149824, 2020 [doi:10.1109/ACCESS.2020.3016780].

[48] Available at: https://blog.paperspace.com/popular-deep-learning-architectures-resnet-inceptionv3-squeezenet/ [Last Access Date: 18.12.2022].

[49] Y. Shen et al., "Urban acoustic classification based on deep feature transfer learning," Journal of the Franklin Institute , vol. 357, no. 1, pp. 667-686, 2020 [doi:10.1016/j.jfranklin.2019.10.014].

[50] M. Lasseck, "Acoustic bird detection with deep convolutional neural networks" in Proc. Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018), 2018, Nov., pp. 143-147.

Nabanita Das

Nabanita Das is a Ph.D. Research Scholar with the Department of Computer science and engineering, GIET University, Gunupur, Orissa, India. Currently, she is an Asst. Professor in the Department of Computer Science and Engineering, Bengal Institute of Technology, India. She received the M. Tech. degree from MAKAUT, West Bengal, India, and has more than ten years of teaching experience. She is actively involved in research in the domains of Machine Learning, Deep Learning, IoT, Software Engineering, and Computer Aided Diagnosis.

Neelamadhab Padhy

Neelamadhab Padhy received his Ph.D. in 2018 from Sri Satya Sai University of technology and medical science, Sehore, India. He is now employed as an Associate Professor in the Department of Computer science and engineering, GIET University, Gunupur. His research topics are machine learning, deep learning software engineering, image processing, etc. He published more than 30 peer-reviewed journal and conference papers. He is a life member of CSI and a member of the IE and Soft Computing Society.

Nilanjan Dey

Nilanjan Dey is an Associate Professor in the Department of Computer Science and Engineering, Techno International New Town, Kolkata, India. He is a visiting fellow of the University of Reading, UK. He also holds a position of Adjunct Professor at Ton Duc Thang University, Ho Chi Minh City, Vietnam. Previously, he held an honorary position of Visiting Scientist at Global Biomedical Technologies Inc., CA, USA (2012–2015). He was awarded his PhD from Jadavpur University in 2015. He is the Editor-in-Chief of the International Journal of Ambient Computing and Intelligence, IGI Global, USA. He is the Series Co-Editor of Springer Tracts in Nature-Inspired Computing (SpringerNature), Data-Intensive Research (SpringerNature), Advances in Ubiquitous Sensing Applications for Healthcare (Elsevier). He is an associate editor of IET Image Processing and editorial board member of Complex & Intelligent Systems, Springer Nature, Applied Soft Computing, Elsevier etc. He is working in the area of medical imaging, machine learning, computer aided diagnosis, data mining, etc. He is the Indian Ambassador of the International Federation for Information Processing—Young ICT Group and Senior member of IEEE.

Sudipta Bhattacharya

Sudipta Bhattacharya is an Asst. Professor in the Department of Computer Science and Engineering, Bengal Institute of Technology, India. He is a Ph.D. Research Scholar with the Department of Computer science and engineering, GIET University, Gunupur, Orissa, India. He received his Bachelor of Technology, (IT) from West Bengal University of Technology, India, and Master of Technology (IT) from the Indian Institute of Engineering Science and Technology, Shibpur, India. His area of research interest is Pattern Recognition and Speech emotion Recognition.

João Manuel R.S. Tavares

João Manuel R.S. Tavares received the degree in mechanical engineering and the M.Sc. and Ph.D. degrees in electrical and computer engineering from the Universidade do Porto, Portugal, in 1992, 1995, and 2001, respectively, and the Habilitation degree in mechanical engineering, in 2015. He is currently a Senior Researcher in the Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial (INEGI), and a Full Professor in the Department of Mechanical Engineering (DEMec), Faculdade de Engenharia da Universidade do Porto (FEUP). He is the co-editor of more than 80 books, the co-author of more than 50 book chapters, 650 articles in international and national journals and conferences, and three international and three national patents. He has been a committee member of several international and national journals and conferences. He is the Co-Founder and the Co-Editor of the book series Lecture Notes in Computational Vision and Biomechanics (Springer), the Founder and Editor-

in-Chief of the journal Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization (Taylor & Francis), the Editor-in-Chief of the journal Computer Methods in Biomechanics and Biomedical Engineering (Taylor & Francis), and the Co-Founder and the Co-Chair of the International Conference Series, such as CompIMAGE, ECCOMAS VipIMAGE, ICCEBS, and BioDental. Additionally, he has been the co-supervisor of several M.Sc. and Ph.D. thesis and a supervisor of several postdoctoral projects. He has participated in many scientific projects both as a Researcher and as a Scientific Coordinator. His research interests include computational vision, medical imaging, computational mechanics, scientific visualization, human-computer interaction, and new product development. (More information can be found at https://www.fe.up.pt/~tavares).