Data Augmentation and Deep Neural Networks For the Classification of Pakistani Racial Speakers Recognition

Ammar Amjad [1] , **Lal Khan** [1] , **Hsien Tsung Chang** [Corresp. 1, 2, 3, 4]

[1] Department of Computer Science and Information Engineering, Chang Gung University, Taoyuan, Taiwan

[2] Bachelor Program in Artificial Intelligence, Chang Gung University, Taoyaun, Taiwan

[3] Department of Physical Medicine and Rehabilitation, Chang Gung Memorial Hospital, Taoyuan, T'ao-yuan, Taiwan

[4] Artificial Intelligence Research Center, Chang Gung University, Taoyuan, Taiwan

Corresponding Author: Hsien Tsung Chang
Email address: smallpig@widelab.org

Speech emotion recognition (SER) systems have evolved into an important method for recognizing a person in several applications, including e-commerce, e-commerce, everyday interactions, law enforcement, and forensics. However, an SER system's efficiency depends on the length of the audio samples used for testing and training. However, the different suggested models successfully obtained relatively high accuracy in this study. Moreover, the degree of SER efficiency is not yet optimum due to the limited database, resulting in overfitting and skewing samples. Therefore, the proposed approach presents a data augmentation method that shifts the pitch, uses multiple window sizes, stretches the time, and adds white noise to the original audio. In addition, a deep model is further evaluated to generate a new paradigm for SER. The data augmentation approach increased the limited amount of data from the Pakistani racial speaker speech dataset in the proposed system. The seven-layer framework was employed to provide the most optimal performance in terms of accuracy compared to other multilayer approaches. The seven-layer method is used in existing work to achieve a very high level of accuracy. The suggested system achieved 97.32\% accuracy with a 0.032\% loss in the 75\%:25\% splitting ratio. In addition, more than 500 augmentation data samples were added. Therefore, the proposed approach results show that deep neural networks with data augmentation can enhance the SER performance on the Pakistani racial speech dataset.

Data Augmentation and Deep Neural Networks for the Classification of Pakistani Racial Speakers Recognition

Ammar Amjad[1]**, Lal Khan**[1]**, and Hsien Tsung Chang**[1,2,3,4]

[1]**Department of Computer Science and Information Engineering, Chang Gung University, Taoyuan 333, Taiwan**

[2]**Bachelor Program in Artificial Intelligence, Chang Gung University, Taoyuan 333, Taiwan**

[3]**Department of Physical Medicine and Rehabilitation, Chang Gung Memorial Hospital, Taoyuan 333, Taiwan**

[4]**Artificial Intelligence Research Center, Chang Gung University, Taoyuan 333, Taiwan**

Corresponding author:

Hsien Tsung Chang[1]

Email address: smallpig@widelab.org

ABSTRACT

Speech emotion recognition (SER) systems have evolved into an important method for recognizing a person in several applications, including e-commerce, e-commerce, everyday interactions, law enforcement, and forensics. However, an SER system's efficiency depends on the length of the audio samples used for testing and training. However, the different suggested models successfully obtained relatively high accuracy in this study. Moreover, the degree of SER efficiency is not yet optimum due to the limited database, resulting in overfitting and skewing samples. Therefore, the proposed approach presents a data augmentation method that shifts the pitch, uses multiple window sizes, stretches the time, and adds white noise to the original audio. In addition, a deep model is further evaluated to generate a new paradigm for SER. The data augmentation approach increased the limited amount of data from the Pakistani racial speaker Speech dataset in the proposed system. The seven-layer framework was employed to provide the most optimal performance in terms of accuracy compared to other multilayer approaches. The seven-layer method is used in existing work to achieve a very high level of accuracy. The suggested system achieved 97.32% accuracy with a 0.032% loss in the 75%:25% splitting ratio. In addition, more than 500 augmentation data samples were added. Therefore, the proposed approach results show that deep neural networks with data augmentation can enhance the SER performance on the Pakistani racial speech dataset.

1 INTRODUCTION

Speaker emotion recognition (SER) is an attractive study since there are still many issues to address and many research gaps that need to be filled. However, deep learning (DL) and machine learning (ML) approaches have tackled SER challenges. Particularly in research that employs speech datasets with enormous volumes of data. The amount of data is increasing at the moment. Consequently, an expansion in the amount of data worldwide is inevitable. Social websites, personal archives, sensors, mobile devices, cameras, webcams, financial market data, and health data create hundreds of petabytes of data (Gupta and Rani (2019); Khan et al. (2022a)). By 2025, the World Economic Forum predicts that the world will create 463 exabytes of data every day. However, it is not easy to find the appropriate method to convert such a large volume of data into useful information.

, Therefore, artificial intelligence (AI) has been used in numerous fields of the latest studies. Previously, speech recognition studies utilizing ML achieved a high degree of precision by using the Gaussian Mixture Model (GMM) technique (Marufo da Silva et al. (2016); Maghsoodi et al. (2019); Mouaz et al. (2019)), and the Hidden Markov Model (HMM) technique (Veena and Mathew (2015); Bao and Shen (2016); Chakroun et al. (2016); Maurya et al. (2018)). However, as the data increases, the level of accuracy with

these techniques drops rapidly, to the point where this traditional ML approach suffers from low accuracy and generalization issues (Xie et al. (2018)). Nevertheless, this technique provides a reliable strategy for addressing data groupings, making it appropriate for various situations.

Several studies have been conducted regarding SER based on deep learning using different methods, such as the Deep Neural Network (DNN) (Seki et al. (2015); Najafian et al. (2016); Matjka et al. (2016); Dumpala and Kopparapu (2017); Snyder et al. (2018); Najafian and Russell (2020); Rohdin et al. (2020); Khan et al. (2021); Amjad et al. (2021b,a,b); Khan et al. (2022b)) and Convolutional Neural Network (CNN) methodologies used in the study (Ravanelli and Bengio (2019)) attained an overall accuracy of 85% with the TIMIT database and 96% with LibriSpeech. Using the deep learning technique, (An et al. (2019)) obtained 96.5 percent accuracy and significantly improved the ability to handle multiple issues in SER. However, DL requires a lot of training datasets, which are challenging to gather and expensive. Therefore, this approach is not suitable for utilization with SER because it will yield overfitting problems and may lead to skewed data. The use of data augmentation (DA) is one solution to the problem of small data in the SER study. A DA approach is a technique that can be used to create additional training datasets by altering the shape of a training dataset. DA is helpful in many investigations, such as digital signal processing, object identification, and image classification (Wu et al. (2020); Li et al. (2020); Amjad et al. (2022)).

The DA technique has been extensively used in various fields of study because a few samples in many different DA classes can help solve a problem more effectively (Zheng et al. (2020)). For example, multiple SER studies using DA (Schlüter and Grill (2015); Salamon and Bello (2017a,b); Pandeya et al. (2018)) showed a reduction of up to 30% in classification errors and obtained 86.194% accuracy. Data augmentation includes several approaches that have been effectively used in various research, including Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) approaches (Moreno-Barea et al. (2020)). The suggested approach obtained accuracy using limited data, with 87.7 percent. In another investigation, scientists employed an auditory DA strategy to achieve an 82.6 percent accuracy for Mandarin-English code flipping (Long et al. (2020)). Pitch shifting is frequently utilized in DA, as presented in (Ye et al. (2020)), and achieved 90% accuracy. In addition, (Damskgg and Vlimki (2017)) employs the time-stretched data augmentation approach when performing DA-based fuzzy identification on a variety of audio signals. (Aguiar et al. (2018)) incorporates Latin music's noise usage, shifting the pitch, loudness variation, and stretching the time to further enhance genre categorization. As a result, (Rituerto-Gonzlez et al. (2019)) reports an 89.45 percent accuracy using the database (LMD). We propose DA because it is proven to increase the quantity of the dataset so that it can help improve speaker recognition performance with an accuracy rate of 99.76

The proposed study presents a data augmentation method based on a seven-layer DNN for recognizing racial speakers in Pakistan by utilizing 400 audio samples from multiple classes of racial speakers in Pakistan. However, this kind of study may easily lead to multiclass difficulties due to the many classes it includes. On the other hand, DNN approaches are often utilized in SER (Nassif et al. (2019)). In addition, DNN is also a powerful model capable of achieving excellent performance in pattern recognition (Nurhaida et al. (2020)). The study undertaken by (Novotny et al. (2018)) in conjunction with Mel-frequency cepstral coefficients (MFCC) has shown the effectiveness of DNN in SER and improved network efficiency in busy and echo conditions. Furthermore, DNN with Mel-frequency cepstral coefficients has outperformed numerous other research approaches on SER single networks (Saleem and Irfan Khattak (2020)). Additionally, DNN has been effectively fusing with augmented datasets. The presented approach employs a seven-layer neural network because the seven-layer technique yields the highest efficiency and accuracy when used in previous works with an average precision above 90% (Liu et al. (2016); Zhang et al. (2018); Li et al. (2019)). Furthermore, including the Pakistani speakers with many classes employing DNN with DA would improve the identification efficiency of multiple emotional classes.

This paper is divided into three sections: Part 1 is an introduction that describes the significant issue and the studies done by the speaker; Part 2 is Related Works, which includes many existing works that support the proposed study; and Part 3 is DA, which describes DA and several methodologies that will be used in the research. Part 4 discusses DNNs, and the deep learning techniques employed. Next, the methodology is covered in Chapter 5. Then, Part 6 includes the research outcomes and a discussion. Finally, section 7 is the conclusion, which covers various significant things about the conclusion of the research outcomes.

2 RELATED WORKS

The proposed study on multi-racial voice recognition was carried out in many nations, like China (Nassif et al. (2019)), Africa (Oyo and Kalema (2014)), Italy (Najafian and Russell (2020)), Pakistan (Syed et al. (2020); Qasim et al. (2016)), United States (Upadhyay and Lui (2018)), and India, through CNN and MFCC (Ashar et al. (2020). It is a vital technique that many researchers have chosen to enhance SER efficacy (Chowdhury and Ross (2020)).

, In contrast, the limitations of multi-racial SER systems investigated in some studies included limited speech data and a lack of emotional classes. Therefore, weak data training methods may result from inaccurate outcomes. Nevertheless, some research in SER and multi-racial SER systems, such as automatic Urdu speech recognition using HMM, involves a ten-speaker category consisting of eight male and two female speakers with 78.2 percent accuracy. In addition, the study of multilingual, multi-speaker involves three classes, namely Javanese, Indonesian, and Sundanese (Azizah et al. (2020)). However, this investigation has limits regarding the number of emotional categories. Various types of SER studies have been conducted. For example, (Durrani and Arshad (2021)) used Deep Residual Network (DRN) with a 74.7 percent accuracy rate. Another study employing MFCC and Fuzzy Vector Quantization Modeling on hundred categories from the TIMIT database gives 98% accuracy, higher than other approaches such as Fuzzy Vector Quantization 2 and Fuzzy C-Means (Singh (2018)). The ML technique is still utilized in conjunction. The classic approaches, such as the HMM, recognize four Moroccan dialect speakers using 20 speakers; this research achieved a 90% accuracy rate for speaker recognition (Mouaz et al. (2019)).

A single-layer DNN with a data augmentation approach is also utilized to investigate the impact of stress on the performance of SER systems, obtaining an accuracy of 99.46% with the VOCE database (Rituerto-Gonzlez et al. (2019)). The VOCE database comprises 135 utterances from forty-five speakers. In addition, the GMM and MFCC with the TIMIT database were utilized to recognize short utterances from 64 different regions and obtained 98.44% accuracy (Chakroun and Frikha (2020)). This accuracy is higher than the traditional GMM. Another approach was employed in a study (Hanifa et al. (2020)) that used 52 recordings of Malaysian recorded samples utilizing the MFCC in the feature extraction, with obtained an accuracy of 57%. Along with machine learning, numerous works in SER and multi-racial utilize the DL technique, regarded as a rigorous approach to SER. The Deep Learning technique with a deep neural network is used with different techniques, one of which is DA, as demonstrated in a study presented by (Long et al. (2020)) on the OC16-CE80 dataset. This Mandarin-English mixlingual speech corpus successfully produced an effective model for SER with an 86% accuracy. The above research has several similarities with the proposed study: the dataset containing speakers from multi-racial backgrounds, DA, and the MFCC feature extraction method. However, some preceding studies differed from the proposed study in many ways, including the number of speech categories, the length of the utterance, and the identification techniques utilized. Table 1 explains the evolution of work on SER in further detail:

3 DATA AUGMENTATION

Researchers employ a method known as data augmentation to enhance the number of dataset samples. DA is an approach for increasing the number of training datasets useful for neural network training (Rebai et al. (2017)) and has a major influence on deep learning with limited datasets (Ma et al. (2019)). Furthermore, DA is a useful method for overcoming overfitting problems, enhancing model dependability, and increasing generalization (Wang et al. (2019)), which are common issues in machine learning. Research-based on deep learning with data augmentation techniques is critical for improving prediction accuracy while dealing with massive volumes of data (Moreno-Barea et al. (2020)). There are a few data augmentation methods, including adding white noise into an original sample, shifting the pitch, loudness variation, multiple window sizes, and stretching the time. The small size of the dataset is a problem when utilizing deep learning approaches. The proposed approach used to overcome this issue is to induce noise into the training data.

Adding white noise: Adding white noise to a speaker's data enhancements recognition effectively (Ko et al. (2017)). This approach involves the addition of random sound samples with similar amplitude but various frequencies (Mohammed et al. (2020)). Using white noise in a speech signal increases the performance of SER (Schlüter and Grill (2015); Aguiar et al. (2018); Hu et al. (2018)). Furthermore, when white noise is added to an original sound gives a distinct sound effect, which increases the performance of

Table 1. Detailed description of datasets

Reference	Approach	Database	Classes	Accuracy
Wang et al. (2014)	HMM and GMM	S-PTH database	4	13.8% and 24.6% Error Rate
Najafian et al. (2016)	DNNs	The First Accents of the British Isles Speech Corpus	14	3.91% and 10.5% error rate
Qasim et al. (2016)	Support Vector Machine,Random Forest and Gaussian Mixture Model	Recorded Pakistan ethnic speaker	6	92.55
Salamon and Bello (2017b)	SB-CNN	Urban-Sound8K	10	94
Upadhyay and Lui (2018)	Deep Belief Network	FAS Database	6	90.2
Singh (2018)	Fuzzy Vector Quantization	TIMIT	100	98.8
Mouaz et al. (2019)	HMM One layer Deep Neural Network	VOCE Corpus Dataset	4	90
Ashar et al. (2020)	CNN	Spontaneous Urdu dataset	–	87.5
Azizah et al. (2020)	DNNs	Indonesian speech corpus	4	98.96
Chakroun et al. (2016)	GMM	TIMIT	8	98.44
Hanifa et al. (2020)	Support Vector Machine	speaker ethnicity	4	56.96
Hanifa et al. (2020)	DNN	OC16	2	86.10

153 SER.

154 **Pitch shifting:** is a commonly used method in an audio sample to increase or decrease the original tone
155 of voice. Pitch variations are performed by using this technique without affecting playback speed (Mousa
156 (2010)). In addition, a method is utilized in pitch shifting to increase the pitch of the original sound
157 without changing the duration of the recorded sound clip (Rai and Barkana (2019)). For example, various
158 studies on Singing Voice Detection (SVD) (Gui et al. (2021)), Environmental Sound Classification (ESC)
159 (Salamon and Bello (2017b)), and domestic cat classification have shown that pitch shifting may be highly
160 effective for DA (Pandeya et al. (2018)).

161 **Time Stretching:** is a way to change the speed or length of an audio signal without changing the tone.
162 Instead, it is used to manipulate audio signals(Damskgg and Vlimki (2017)). This technique is suitable
163 for analyzing auditory signals that comprise tone, noise, and temporal elements. Numerous investigations
164 used time stretching with other approaches such as synchronous overlap, fuzzy, and CNN to increase the
165 efficiency of the suggested framework (Sasaki et al. (2010); Kupryjanow and Czyewski (2011); Salamon
166 and Bello (2017a)). These studies used different techniques, such as the Synchronous Overlap algorithm,
167 fuzzy logic, and CNN, to improve the performance of the proposed model.

168 **Multiple Window Size:** Multiple window size features are retrieved from a windowed signal called
169 frames. The window strongly influences the obtained features retrieved from the voice signal-based
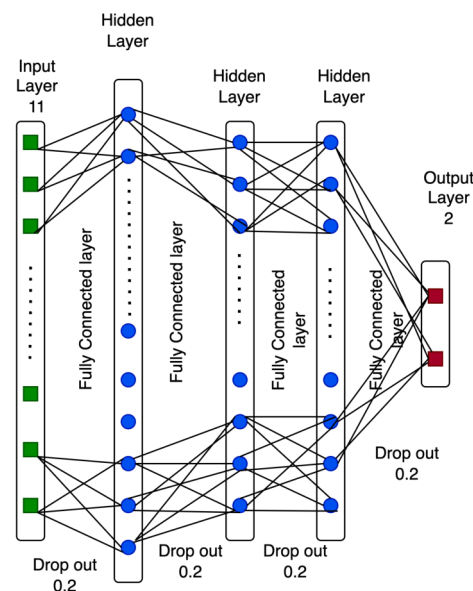
Figure 1. Structure of a Deep Neural Network

functions width since signals are often steady for limited periods (Kelly and Gobl (2011)). Suppose the length of the window is relatively small. In that case, insufficient training datasets are available to get an accurate spectrum for estimating the signals. On the other hand, if the window's length is set very wide, the signal may vary significantly across the frame. Thus, determining the width of the window function is a critical phase that is made more difficult by the lack of details about the original data (Rabiner and Schafer (2007); Zhang et al. (2019)). Several studies have demonstrated that the optimal window size selection contributes to the correlation between the acoustic representation and the human perception of a speech signal (Nisar et al. (2016); Kirkpatrick et al. (2006)). Three tuples express a window function: width of the window, offset, and shape. To extract a part of a signal, multiply the signal's value at the time "t," signal[t], by the value of the hamming window at a time "t," window[t], which is expressed as:
windowsignal[t] =window[t]* signal[t]

A windowed signal is utilized to create characteristics for emotion recognition. For SER, a standard size window of 25 ms is employed to extract features with a 10 ms overlap (Yoon et al. (2018); Tarantino et al. (2019); Ramet et al. (2018)). On the other hand, some research has indicated that a larger window size improves emotion identification performance (Chernykh and Prikhodko (2018); Tripathi et al. (2019)). In addition, other studies have assessed the significance of step size (overlap window size). However, SER analysis is conducted using a single-window (Tarantino et al. (2019); Chernykh and Prikhodko (2018)). Therefore, In (Tarantino et al. (2019)) investigated the influence of overlap window size on SER. They discovered that a small step size leads to a lower test loss. In (Chernykh and Prikhodko (2018)), explored multiple window widths ranging from 30 ms to 200 ms before settling on a unique 200 ms window for the SER study.

4 METHODOLOGY

Deep Learning has been used to create a variety of solid approaches for SER. The DNN is one of the most widely utilized deep learning approaches. In many SER studies, deep neural networks are employed because they have several benefits over conventional machine learning approaches. There are several benefits to using the DNN approach in many scientific domains, including object detection, geographic information retrieval, and voice classification Seifert et al. (2017). The DNN-based acoustic model was used in previous work to achieve high-level performance Seki et al. (2015); Snyder et al. (2018); Novotny et al. (2018); Saleem and Irfan Khattak (2020).

The structure of a DNN approach is composed of input, hidden, dropout, and output layers Rajyaguru et al. (2020) . The deep neural network is an evolution of the neural network (see Fig. 1), which is essentially a function in a mathematical measure R: A ⇒ B that may be stated as follows:
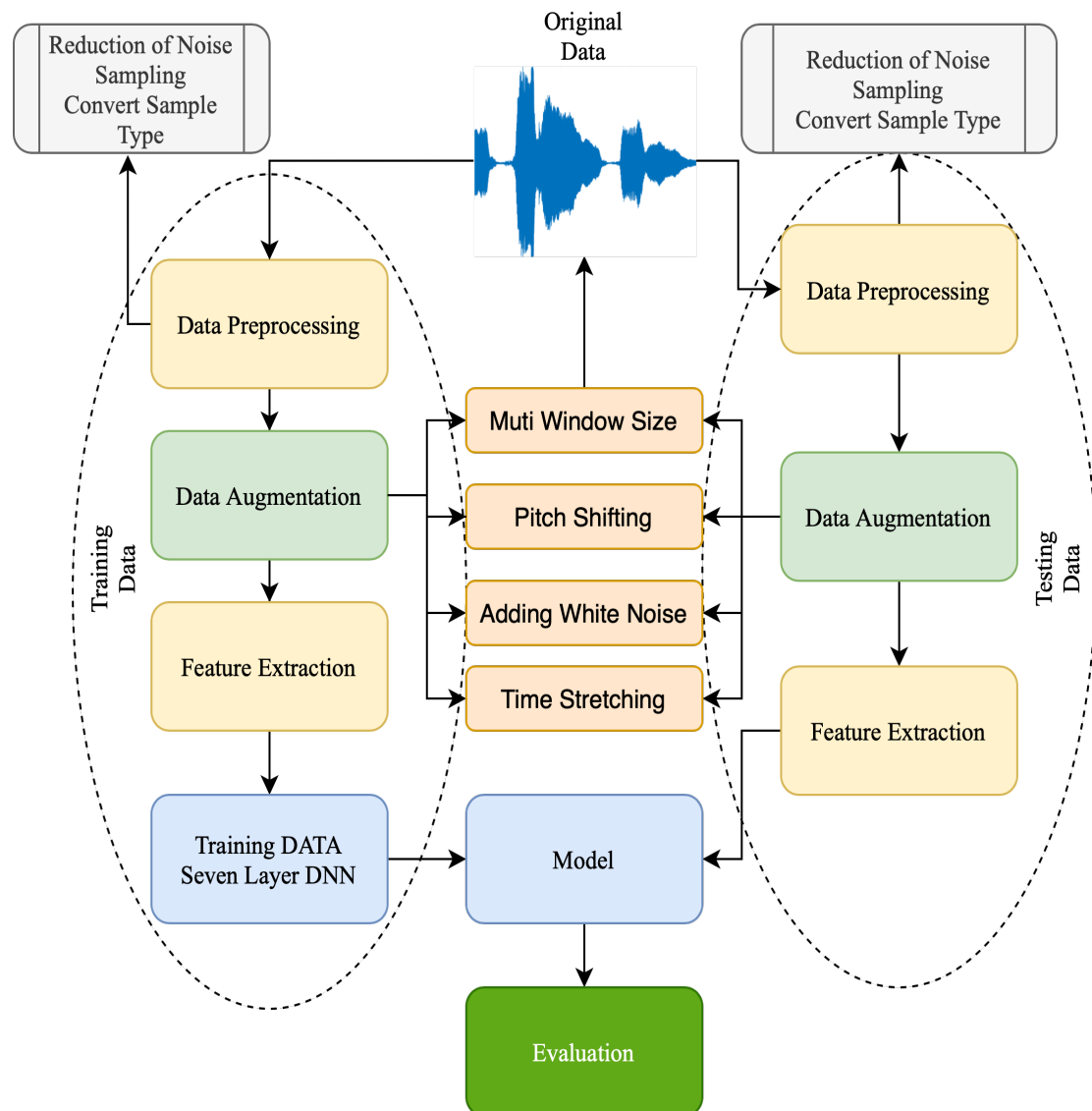
Figure 2. Structure of Proposed Approach

4.1 Input Layer

The input layer comprises nodes that obtain the inputted data from variable A. These nodes are directly connected to the hidden units. The generation of eleven input layer features is generated after a preprocessing step utilizing the Principal Component Analysis (PCA) algorithm.

4.2 Hidden layer

The hidden layer is composed of nodes that obtain data from the first layer. Previous studies have suggested that the volume of nodes in the hidden layer may be influenced by the dimensions of the input and output layers. For example, in Fig. 1, the size of the hidden neurons is 24.12, and 12 in the hidden units, which is the optimal number of deep neural network characteristics based on previous studies.

4.3 Dropout (DO)

A dropout is a single approach utilized to generate a range of system designs that may be used to address overfitting issues in the model. The dropout value ranges between 0 and 1. Dropout is set to a size of 0.2 for each layer in figure 1, since DNN obtains the highest efficiency with this value.
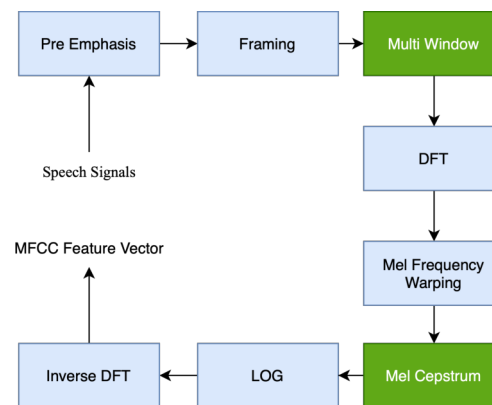
Figure 3. Block Diagram of the computation steps of MFCC
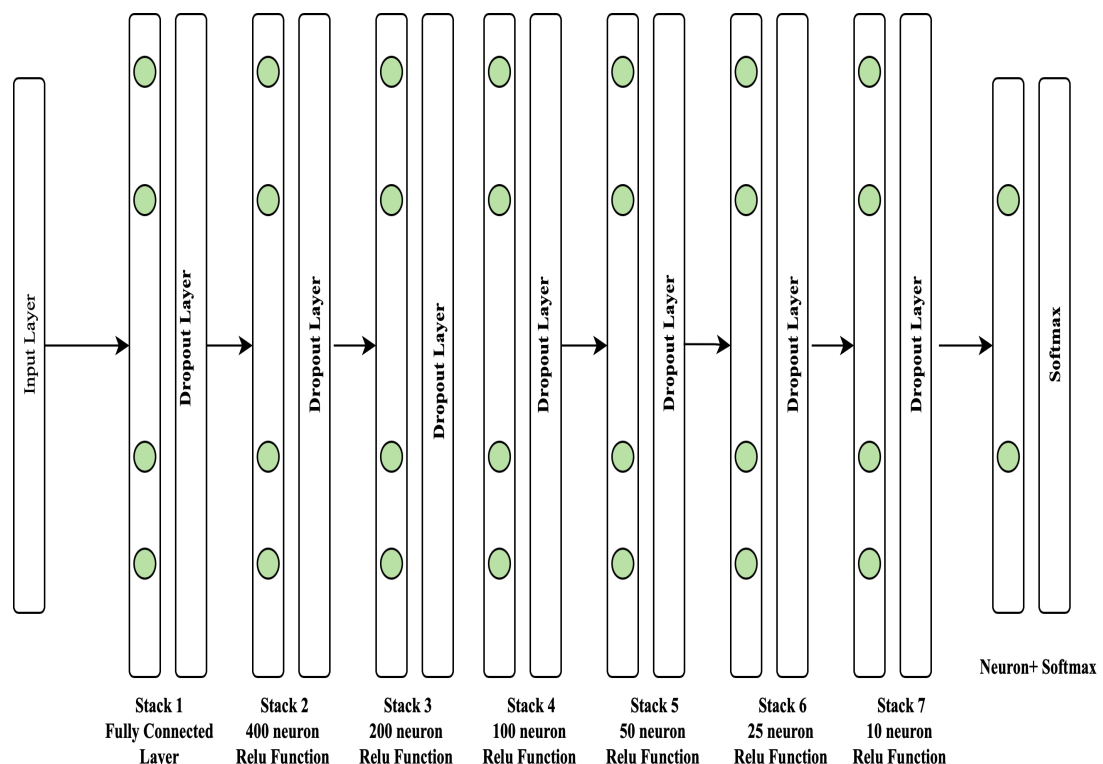
Figure 4. Structure of Proposed Approach

4.4 Output layer

The output layer comprises nodes that access data directly from the hidden or input layer. The output value provides a computation outcome from the A to B value. For example, the two output layer nodes in 1 represent the number of groups. The proposed technique improved the Pakistani racial speaker recognition accuracy. It was based on the seven-layer DNN architecture with a data augmentation approach. Fig. 2 illustrates the proposed method's architecture. The proposed SER using a seven-layer DNN-DA approach to the multi-language dataset, as shown in Fig. 2, is a robust approach. First, a dataset is divided into training data (75% of the dataset) and testing data (25% of the dataset). Then, the training data is preprocessed by trimming audio signals with identical temporal lengths and generating sample types with similar shapes and sizes. Moreover, four techniques of the data augmentation procedure are performed on the dataset to enhance audio data. Finally, the MFCC extracts the features and processes them with a seven-layer DNN-DA for classification. The testing dataset performs the same preprocessing steps, data augmentation, and feature extraction using MFCC. Furthermore, the proposed approach will be evaluated

using testing data to see how accurate it is in terms of Speaker Recognition.

Table 2. Duration of audio speech data in hours

Racial	Number of Male and Female Speakers	Duration per sample	Number of Samples	Nature of samples
Punjabi (Wang and Guan (2008))	4 males and 4 females	42 seconds	500 samples	Speaker and text independent
Urdu (Wang and Guan (2008); Syed et al. (2020))	4 males and 4 females	42 seconds	500 samples	Speaker and text independent
Sindhi(Syed et al. (2020))	32 males and 38 females	30 seconds	80 samples	Speaker and text independent
Saraiki	42 males and 28 females	30 seconds	80 samples	Speaker and text independent
Pashto	35 males and 35 females	30 seconds	80 samples	Speaker and text independent

228

5 DATASET AND PREPROCESSING

This study utilized a dataset of the six most spoken local languages in Pakistan. The information was obtained to adjust for the numerous ethnicities. A variety of online resources were used to compile this dataset(Wang and Guan (2008); Syed et al. (2020)) . This study aims to gather data from areas of Pakistan where Urdu and its five primary ethnicities (Punjabi, Sindhi, Urdu, Saraiki, and Pashto) are spoken. The audio samples were processed using PRAAT software. The dataset for the Urdu language is summarized in Table 2. The dataset is utilized only to recognize Urdu racials. The dataset contains 80 distinct utterances for each ethnicity type with different levels of education, ranging from semi-literate to literate. Each audio file is from an individual speaker, resulting in 80 distinct speakers per ethnic group. Each clip is 30 seconds long, in mono channel WAV format, and sampled at 16 kHz of Sindhi, Saraiki, and Pashto languages. Additionally, each utterance is distinct from others in the dataset. The dataset includes sounds from 80 speakers with five racials, for 1240 clips.

The dataset processing uses a segmentation process similar to that used for the dataset of The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). This multimodal recording dataset takes the form of emotional speech and songs recorded in audio and video formats Atmaja and Akagi (2020a). Experiments on RAVDESS were carried out by Livingstone and Russo (2018), and they involved the participation of 24 professional actors with North American accents. The research included speech and songs with various facial expressions, including neutral, calm, happy, sad, angry, fearful, surprised, and disgusted. In the data of Pakistani racial speakers, the complete audio utterances are segmented once again using the approach that is described below:

- Modality 001 = only-audio , 002 = only-video, 003 = audio-video

- Classes: 001 = disgust, 002 = neutral, 003 = fearful, 004 = angry, 005 = happy, 006 = surprised, 007 = sad, 008 = calm

- Vocal: 001 = song , 002 = speech

- Intensity: 001 = strong, 002 = normal

- The racial of the speakers as a class from 01 to 5

- Repetition: 001 = First, 002 = second

- Speaker sequence number per tribe/region from 01 to 10

5.1 Feature Extraction

We employed MFCC in the proposed study since it is one of the most robust approaches to extracting features from SER features. MFCC is the most widely used approach for obtaining spectral information from a speech by processing the Fourier Transform (FT) signal with a perception-based Mel-space filter bank. Additionally, in the proposed study, Librosa is used to extract MFCC features. This Python library has functionality for reading sound data and assisting in the MFCC feature extraction method. According to Hamidi et al. (2020), the MFCC technique is shown in Fig. 3: The MFCC approach enhances the audio sound input during the preemphasis phase and increases the signal-to-noise ratio (SNR) enough to ensure that the voice is not influenced by noise. The framing mechanism divides the audio signal into many frames with the same signal count. Windowing is the technique of employing the window function to weight the output frame. The following procedure is the DFT (Discrete Fourier Transform), which examines the frequency signal derived from the discrete-time signal. Then, the MFCC obtained from the original utterances is determined using the filter bank (FB). The wrapping of Mel Frequency is often used in conjunction with a FB. A FB is a kind of filter used to determine the amount of energy contained within a certain frequency range, Afrillia et al. (2017). Finally, the logarithmic (LOG) value is obtained by converting the DFT result to a single value. Inverse DFT is a technique for obtaining a perceptual autocorrelation sequence based on the linear prediction (LP) coefficient computation. The MFCC technique was employed in this study by setting frame lengths at 25 with a hamming window, 13 spectral and 22 lifter coefficients, and ten frameshifts. The MFCC approach enhances the audio sound input during the preemphasis phase, increasing the signal-to-noise ratio (SNR) enough to ensure that the voice is not influenced by noise. The framing mechanism divides the audio signal into many frames with the same signal count. Windowing is the technique of employing the window function to weigh the output frame. The following procedure is the DFT (Discrete Fourier Transform), which examines the frequency signal derived from the discrete-time signal. Then, the MFCC obtained from the original utterances is determined using the filter bank (FB). The wrapping of Mel Frequency is often used in conjunction with a FB.

5.2 Seven Layer DNN

In this study, the Rectified Linear Unit (Relu) activation function is utilized in conjunction with the adam optimizer (AO). Adam optimizer is used to improve the learning speed of deep neural networks. This algorithm was introduced at a renowned conference by deep learning experts, Kingma and Ba (2017), with a 0.2%dropout rate. A deep neural network comprises seven layers, with the structure shown in Fig. 4.

As seen in Fig. 4, the seven-layer architecture of the DNN consists of one fully connected layer with 400 neurons on layer 2, which is the expected volume of neurons identified in our investigation. The following layer has just half of the neurons from the preceding layer. Layer one is composed of dense functions that create a fully connected layer. The second layer comprises 400 neurons composed of the dense and dropout functions used in the neural network to avoid overfitting and accelerate the learning process. The third layer comprises 200 neurons. The fourth layer comprises 100 neurons, the fifth layer comprises 50 neurons, and the sixth layer comprises 25 neurons. It is also composed of dense and dropout functions. Finally, the seventh layer comprises ten neurons with dense and dropout functions. At the same time, softmax activation is used as the output layer. The seven-layer DNN architecture is employed in this work because it provides the maximum level of accuracy compared to the three-layer DNN and five-layer DNN.

Table 3. Comparison table of loss at dividing ratio with accuracy

Dividing Ratio	Classification Accuracy	Total Loss
90 : 10	93.55	0.105
80 : 20	95.767	0.093
75 : 25	97.32	0.032

5.3 Evaluation

Acted, semi-natural, and spontaneous datasets were employed in the proposed study. In addition, the split ratio method with train test split assessment was used to evaluate performance in ML. The proposed

Table 4. The accuracy and loss comparison table includes augmentation data with 75:25 ratio

Data augmentation	Accuracy	Loss
100	96.57	1.33
200	96.21	0.05
300	96.83	2.77
400	96.45	0.035
500	97.32	0.031

Table 5. The accuracy and loss comparison table includes augmentation data with 80:20 ratio

Data augmentation	Accuracy	Loss
100	95.12	6.33
200	95.99	0.04
300	96.13	0.19
400	96.29	0.66
500	97.09	2.77

303   approach separates the data into training for matching the ML architecture and testing the ML architecture.
304   The most utilized ratio is splitting training and testing data by 70%: 30%, 80%: 20%, or 90%: 10%.
305   Multiple factors determine the split ratios, namely the compute costs associated with the model training,
306   the computational costs associated with testing the model, and data analysis. Accuracy is a commonly
307   used metric for assessing the extent of incorrectly identified items in balanced and approximately balanced
308   datasets Atmaja and Akagi (2020b). It is one of the model performance assessment methodologies often
used in ML.
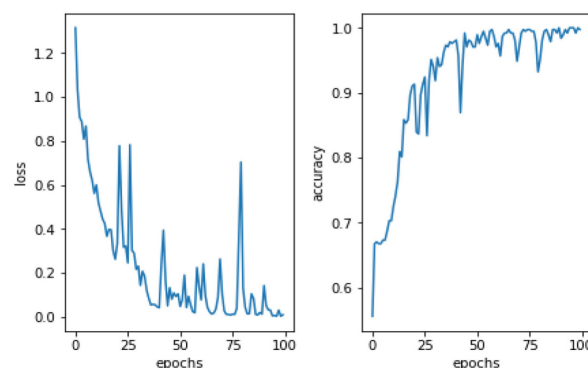
Figure 5. Proposed model performance on training dataset

309

6 RESULTS AND DISCUSSION

311   This study utilized DA methods to evaluate a Pakistani racial speech dataset using a 44,100 mono sample
312   rate. The testing efficacy of the seven-layer DNN-DA approach at epoch 100 with batch size two is
313   illustrated in Fig. 5. Testing a training dataset yields an accuracy of 97.32% with a total loss of 0.03. As
314   shown in Fig. 5, the total loss decreases from epoch 1 to 100. However, it has remained unstable at epochs
315   20, 28, 38, 64, 73, and 77, with loss increases that automatically decrease precision efficiency at epochs
316   20,28,38,64,73,77. It eventually stabilized above 90% in the 88th epoch. The graph in Fig. 6 illustrates
317   the outcomes of model testing utilizing data testing. Using 500 data wav files shows that the seven-layer
318   DNN-DA model produces a robust technique for SER. With highest efficiency of 97.32% and a low loss
319   rate of 0.032, the seven-layer DNN-DA model produces a robust approach for speaker recognition and
320   lacks overfitting in this model test. A split ratio is also used to assess the proposed approach performance,
321   as illustrated in Table 3.
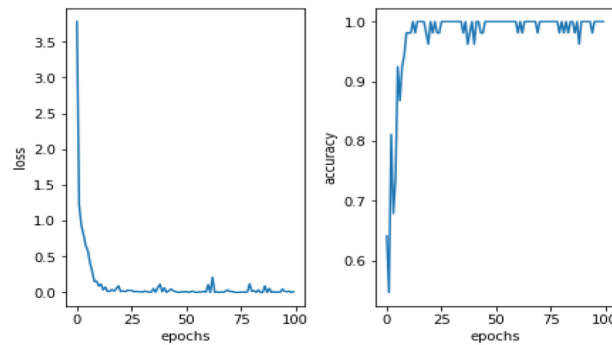
Figure 6. Proposed model performance on testing dataset

According to Table 4, when the split ratio is 75:25, the trained model achieves the highest accuracy and the lowest loss level. As shown in Table 5, the accuracy of the results decreases when the split ratio is 80:20. At the same time, the loss increases. Finally, when the split ratio is 90:10, the accuracy results increase while the loss rate decreases. Table 6 results illustrate that testing with a large amount of training data is beneficial since it exposes the model to many instances, allowing it to identify the optimal solution.

Table 6. The accuracy and loss comparison table includes augmentation data with 90:10 ratio

Data augmentation	Accuracy	Loss
100	95.21	0.13
200	96.90	0.28
300	96.34	3.22
400	96.99	6.23
500	97.01	5.232

Table 7. Comparison of outcomes with different ML and DL algorithms

Dataset	Classification Accuracy	Accuracy
	KNN	81.99
	Random Forest	71.56
	Multilayer Perceptron (MLP)	91.45
Pakistani Racial Speaker Classification	Decision Tree	67.45
	Three layers Deep Neural Network	92.56
	Five layers Deep Neural Network	94.78
	Seven Layer DNN-DA (Proposed)	97.732

On the other hand, if we utilize an insufficient training dataset, the model will lack expertise, resulting in inferior output during testing. The proposed approach will gain a more profound understanding and increase the model's generalizability by including many testing datasets. As shown in table 4-6, another test was conducted with the addition of 100 to 500 data samples to the original 400 wav data using the split ratio approach.

In the suggested method, a dataset with a data augmentation of 500 samples and a split ratio of 75:25 obtained the highest performance with a low total loss. However, as the sample of DA decreases, the SER model's performance decreases. In another comparison, accuracy improves when a large DA and a significant amount of training data are used. Additionally, as seen in Table 7, the study has the highest accuracy performance compared to numerous methodologies using ML and DL algorithms. The study performance on SER in Table 7 demonstrates that the seven-layer approach we presented is practical. DNN-DA is a robust approach for usage in SER that has achieved a high degree of accuracy. It is not straightforward to get accurate prediction findings while researching several classes. Certain aspects of multi-classes will be more challenging since they must discriminate between many classes while generating predictions

Silva-Palacios et al. (2017). However, seven layer DNN-DA outperforms conventional machine learning methods such as k-nearest Neighbors(KNN), Random Forest(RF), Multilayer Perceptron, Decision Tree, and DL approaches using three-layer DNN layer and five-layer DNN, as demonstrated by the highest accuracy performance compared to other approaches using three-layer DNN and five-layer layers DNN layer.

7 CONCLUSION

A study in SER that includes significant data is a challenging research issue; the Pakistani racial speech dataset is comprised of utterance groups. Therefore, seven-layer DNN-DA is the approach presented in this report, which combines the data augmentation technique with a DNN to improve performance and minimize overfitting issues. Finally, some of the contributions to our work include using a Pakistani racial speech dataset in this study. Furthermore, DA can increase the amount of data by using white noise, variable window widths, pitch-shifting, and temporal stretching methods to generate new audio data for the segments. Furthermore, classification with deep neural networks of seven layers is beneficial for improving the performance of the SER system when used with all Pakistani racial speech datasets. In addition, the proposed model with the seven-layer DNN-DA technique also has an accuracy advantage. Similar to some approaches using conventional ML and DL methods that also produce high accuracy performance.

REFERENCES

Afrillia, Y., Mawengkang, H., Ramli, M., Fadlisyah, and Fhonna, R. P. (2017). Performance measurement OfMel frequency ceptral coefficient(MFCC) method in learning system of al- qur'an based InNaghamPattern recognition. *Journal of Physics: Conference Series*, 930:012036.

Aguiar, R. L., Costa, Y. M., and Silla, C. N. (2018). Exploring data augmentation to improve music genre classification with convnets. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Amjad, A., Khan, L., Ashraf, N., Mahmood, M. B., and Chang, H.-T. (2022). Recognizing semi-natural and spontaneous speech emotions using deep neural networks. *IEEE Access*, 10:37149–37163.

Amjad, A., Khan, L., and Chang, H.-T. (2021a). Effect on speech emotion classification of a feature selection approach using a convolutional neural network. *PeerJ Computer Science*, 7:e766.

Amjad, A., Khan, L., and Chang, H.-T. (2021b). Semi-natural and spontaneous speech recognition using deep neural networks with hybrid features unification. *Processes*, 9(12).

An, N. N., Thanh, N. Q., and Liu, Y. (2019). Deep cnns with self-attention for speaker identification. *IEEE Access*, 7:85327–85337.

Ashar, A., Bhatti, M. S., and Mushtaq, U. (2020). Speaker identification using a hybrid cnn-mfcc approach. In *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*, pages 1–4.

Atmaja, B. T. and Akagi, M. (2020a). On the differences between song and speech emotion recognition: Effect of feature sets, feature types, and classifiers.

Atmaja, B. T. and Akagi, M. (2020b). On the differences between song and speech emotion recognition: Effect of feature sets, feature types, and classifiers. In *2020 IEEE REGION 10 CONFERENCE (TENCON)*, pages 968–972.

Azizah, K., Adriani, M., and Jatmiko, W. (2020). Hierarchical transfer learning for multilingual, multi-speaker, and style transfer dnn-based tts on low-resource languages. *IEEE Access*, 8:179798–179812.

Bao, L. and Shen, X. (2016). Improved gaussian mixture model and application in speaker recognition. In *2016 2nd International Conference on Control, Automation and Robotics (ICCAR)*, pages 387–390.

Chakroun, R., Beltafa Zouari, L., Frikha, M., and Ben Hamida, A. (2016). Improving text-independent speaker recognition with gmm. In *2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 693–696.

Chakroun, R. and Frikha, M. (2020). Robust text-independent speaker recognition with short utterances using gaussian mixture models. In *2020 International Wireless Communications and Mobile Computing (IWCMC)*, pages 2204–2209.

Chernykh, V. and Prikhodko, P. (2018). Emotion recognition from speech with recurrent neural networks.

Chowdhury, A. and Ross, A. (2020). Fusing mfcc and lpc features using 1d triplet cnn for speaker

392   recognition in severely degraded audio signals. *IEEE Transactions on Information Forensics and*
393   *Security*, 15:1616–1629.

394   Damskgg, E.-P. and Vlimki, V. (2017). Audio time stretching using fuzzy classification of spectral bins.
395   *Applied Sciences*, 7(12):1293.

396   Dumpala, S. H. and Kopparapu, S. K. (2017). Improved speaker recognition system for stressed speech
397   using deep neural networks. In *2017 International Joint Conference on Neural Networks (IJCNN)*,
398   pages 1257–1264.

399   Durrani, S. and Arshad, M. U. (2021). Transfer learning based speech affect recognition in urdu.

400   Gui, W., Li, Y., Zang, X., and Zhang, J. (2021). Exploring channel properties to improve singing voice
401   detection with convolutional neural networks. *Applied Sciences*, 11(24).

402   Gupta, D. and Rani, R. (2019). A study of big data evolution and research challenges. *Journal of
403   Information Science*, 45(3):322–340.

404   Hamidi, M., Satori, H., Zealouk, O., and Satori, K. (2020). Amazigh digits through interactive speech
405   recognition system in noisy environment. *International Journal of Speech Technology*, 23(1):101–109.

406   Hanifa, R. M., Isa, K., and Mohamad, S. (2020). Speaker ethnic identification for continuous speech in
407   Malay language using pitch and MFCC. *Indonesian Journal of Electrical Engineering and Computer
408   Science (IJEECS)*, 19(1):207–214.

409   Hu, H., Tan, T., and Qian, Y. (2018). Generative adversarial networks based data augmentation for noise
410   robust speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal
411   Processing (ICASSP)*, pages 5044–5048.

412   Kelly, A. C. and Gobl, C. (2011). The effects of windowing on the calculation of mfccs for different types
413   of speech sounds. In Travieso-González, C. M. and Alonso-Hernández, J. B., editors, *Advances in
414   Nonlinear Speech Processing*, pages 111–118, Berlin, Heidelberg. Springer Berlin Heidelberg.

415   Khan, L., Amjad, A., Afaq, K. M., and Chang, H.-T. (2022a). Deep sentiment analysis using cnn-lstm
416   architecture of english and roman urdu text shared in social media. *Applied Sciences*, 12(5).

417   Khan, L., Amjad, A., Ashraf, N., and Chang, H.-T. (2022b). Multi-class sentiment analysis of urdu text
418   using multilingual bert. *Scientific Reports*, 12(1):5436.

419   Khan, L., Amjad, A., Ashraf, N., Chang, H.-T., and Gelbukh, A. (2021). Urdu sentiment analysis with
420   deep learning methods. *IEEE Access*, 9:97803–97812.

421   Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization.

422   Kirkpatrick, B., O'Brien, D., and Scaife, R. (2006). A comparison of spectral continuity measures as a
423   join cost in concatenative speech synthesis. In *2006 IET Irish Signals and Systems Conference*, pages
424   515–520.

425   Ko, T., Peddinti, V., Povey, D., Seltzer, M. L., and Khudanpur, S. (2017). A study on data augmentation of
426   reverberant speech for robust speech recognition. In *2017 IEEE International Conference on Acoustics,
427   Speech and Signal Processing (ICASSP)*, pages 5220–5224.

428   Kupryjanow, A. and Czyewski, A. (2011). A non-uniform real-time speech time-scale stretching method.
429   *Proceedings of the International Conference on Signal Processing and Multimedia Applications*, pages
430   1–7.

431   Li, X., Zhang, W., Ding, Q., and Sun, J.-Q. (2020). Intelligent rotating machinery fault diagnosis based
432   on deep learning using data augmentation. *Journal of Intelligent Manufacturing*, 31(2):433–452.

433   Li, Z., Wang, S.-H., Fan, R.-R., Cao, G., Zhang, Y.-D., and Guo, T. (2019). Teeth category classification
434   via seven-layer deep convolutional neural network with max pooling and global average pooling.
435   *International Journal of Imaging Systems and Technology*, 29(4):577–583.

436   Liu, J., Fang, C., and Wu, C. (2016). A fusion face recognition approach based on 7-layer deep learning
437   neural network. *Journal of Electrical and Computer Engineering*, 2016:8637260.

438   Livingstone, S. R. and Russo, F. A. (2018). The ryerson audio-visual database of emotional speech and
439   song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english.
440   *PLOS ONE*, 13(5):1–35.

441   Long, Y., Li, Y., Zhang, Q., Wei, S., Ye, H., and Yang, J. (2020). Acoustic data augmentation for
442   mandarin-english code-switching speech recognition. *Applied Acoustics*, 161:107175.

443   Ma, R., Tao, P., and Tang, H. (2019). Optimizing data augmentation for semantic segmentation on
444   small-scale dataset. In *Proceedings of the 2nd International Conference on Control and Computer
445   Vision*, ICCCV 2019, page 7781, New York, NY, USA. Association for Computing Machinery.

446   Maghsoodi, N., Sameti, H., Zeinali, H., and Stafylakis, T. (2019). Speaker recognition with random digit

strings using uncertainty normalized hmm-based i-vectors. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(11):1815–1825.

Marufo da Silva, M., Evin, D. A., and Verrastro, S. (2016). Speaker-independent embedded speech recognition using hidden markov models. In *IEEE CACIDI 2016 - IEEE Conference on Computer Sciences*, pages 1–6.

Matjka, P., Glembek, O., Novotn, O., Plchot, O., Grzl, F., Burget, L., and Cernock, J. H. (2016). Analysis of dnn approaches to speaker identification. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5100–5104.

Maurya, A., Kumar, D., and Agarwal, R. (2018). Speaker recognition for hindi speech signal using mfcc-gmm approach. *Procedia Computer Science*, 125:880–887. The 6th International Conference on Smart Computing and Communications.

Mohammed, M. A., Abdulkareem, K. H., Mostafa, S. A., Khanapi Abd Ghani, M., Maashi, M. S., Garcia-Zapirain, B., Oleagordia, I., Alhakami, H., and AL-Dhief, F. T. (2020). Voice pathology detection and classification using convolutional neural network model. *Applied Sciences*, 10(11).

Moreno-Barea, F. J., Jerez, J. M., and Franco, L. (2020). Improving classification accuracy using data augmentation on small data sets. *Expert Systems with Applications*, 161:113696.

Mouaz, B., Abderrahim, B. H., and Abdelmajid, E. (2019). Speech recognition of moroccan dialect using hidden markov models. *Procedia Computer Science*, 151:985–991. The 10th International Conference on Ambient Systems, Networks and Technologies (ANT 2019) / The 2nd International Conference on Emerging Data and Industry 4.0 (EDI40 2019) / Affiliated Workshops.

Mousa, A. (2010). Voice conversion using pitch shifting algorithm by time stretching with psola and re-sampling.

Najafian, M. and Russell, M. (2020). Automatic accent identification as an analytical tool for accent robust automatic speech recognition. *Speech Communication*, 122:44–55.

Najafian, M., Safavi, S., Hansen, J. H. L., and Russell, M. (2016). Improving speech recognition using limited accent diverse british english training data with deep neural networks. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6.

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., and Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. *IEEE Access*, 7:19143–19165.

Nisar, S., Khan, O. U., and Tariq, M. (2016). An efficient adaptive window size selection method for improving spectrogram visualization. *Computational Intelligence and Neuroscience*, 2016:6172453.

Novotny, O., Plchot, O., Glembek, O., Cernocky, J. H., and Burget, L. (2018). Analysis of dnn speech signal enhancement for robust speaker recognition.

Nurhaida, I., Ayumi, V., Fitrianah, D., Zen, R. A., Noprisson, H., and Wei, H. (2020). Implementation of deep neural networks (dnn) with batch normalization for batik pattern recognition. *International Journal of Electrical and Computer Engineering (IJECE)*, 10(2):2045–2053.

Oyo, B. and Kalema, B. M. (2014). A preliminary speech learning tool for improvement of african english accents. In *2014 International Conference on Education Technologies and Computers (ICETC)*, pages 44–48.

Pandeya, Y. R., , and Lee, J. (2018). Domestic cat sound classification using transfer learning. *The International Journal of Fuzzy Logic and Intelligent Systems*, 18(2):154–160.

Qasim, M., Nawaz, S., Hussain, S., and Habib, T. (2016). Urdu speech recognition system for district names of pakistan: Development, challenges and solutions. In *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 28–32.

Rabiner, L. R. and Schafer, R. W. (2007). Introduction to digital speech processing. *Found. Trends Signal Process.*, 1(1):1194.

Rai, A. and Barkana, B. D. (2019). Analysis of three pitch-shifting algorithms for different musical instruments. In *2019 IEEE Long Island Systems, Applications and Technology Conference (LISAT)*, pages 1–6.

Rajyaguru, V., Vithalani, C., and Thanki, R. (2020). A literature review: various learning techniques and its applications for eye disease identification using retinal images. *International Journal of Information Technology*.

Ramet, G., Garner, P. N., Baeriswyl, M., and Lazaridis, A. (2018). Context-aware attention mechanism for speech emotion recognition. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages

126–131.

Ravanelli, M. and Bengio, Y. (2019). Speaker recognition from raw waveform with sincnet.

Rebai, I., BenAyed, Y., Mahdi, W., and Lorr, J.-P. (2017). Improving speech recognition using data augmentation and acoustic model fusion. *Procedia Computer Science*, 112:316–322. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 21st International Conference, KES-20176-8 September 2017, Marseille, France.

Rituerto-Gonzlez, E., Mnguez-Snchez, A., Gallardo-Antoln, A., and Pelez-Moreno, C. (2019). Data augmentation for speaker identification under stress conditions to combat gender-based violence. *Applied Sciences*, 9(11).

Rohdin, J., Silnova, A., Diez, M., Plchot, O., Matjka, P., Burget, L., and Glembek, O. (2020). End-to-end dnn based text-independent speaker recognition for long and short utterances. *Computer Speech Language*, 59:22–35.

Salamon, J. and Bello, J. P. (2017a). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3):279–283.

Salamon, J. and Bello, J. P. (2017b). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3):279283.

Saleem, N. and Irfan Khattak, M. (2020). Deep neural networks based binary classification for single channel speaker independent multi-talker speech separation. *Applied Acoustics*, 167:107385.

Sasaki, T., Nakajima, Y., ten Hoopen, G., van Buuringen, E., Massier, B., Kojo, T., Kuroda, T., and Ueda, K. (2010). Time stretching: Illusory lengthening of filled auditory durations. *Atten Percept Psychophys*, 72(5):1404–1421.

Schlüter, J. and Grill, T. (2015). Exploring data augmentation for improved singing voice detection with neural networks. In *ISMIR*.

Seifert, C., Aamir, A., Balagopalan, A., Jain, D., Sharma, A., Grottel, S., and Gumhold, S. (2017). *Visualizations of Deep Neural Networks in Computer Vision: A Survey*, pages 123–144. Studies in Big Data. Springer, Netherlands.

Seki, H., Yamamoto, K., and Nakagawa, S. (2015). Deep neural network based acoustic model using speaker-class information for short time utterance. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1222–1225.

Silva-Palacios, D., Ferri, C., and Ramrez-Quintana, M. J. (2017). Improving performance of multiclass classification by inducing class hierarchies. *Procedia Computer Science*, 108:1692–1701. International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland.

Singh, S. (2018). Speaker recognition by gaussian filter based feature extraction and proposed fuzzy vector quantization modelling technique.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333.

Syed, Z. S., Memon, S. A., Shah, M. S., and Syed, A. S. (2020). Introducing the urdu-sindhi speech emotion corpus: A novel dataset of speech recordings for emotion recognition for two low-resource languages. *International Journal of Advanced Computer Science and Applications*, 11(4).

Tarantino, L., Garner, P. N., and Lazaridis, A. (2019). Self-attention for speech emotion recognition. In *INTERSPEECH*.

Tripathi, S., Tripathi, S., and Beigi, H. (2019). Multi-modal emotion recognition on iemocap dataset using deep learning.

Upadhyay, R. and Lui, S. (2018). Foreign english accent classification using deep belief networks. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pages 290–293.

Veena, K. V. and Mathew, D. (2015). Speaker identification and verification of noisy speech using multi-taper mfcc and gaussian mixture models. In *2015 International Conference on Power, Instrumentation, Control and Computing (PICC)*, pages 1–4.

Wang, H., Wang, L., and Liu, X. (2014). Multi-level adaptive network for accented mandarin speech recognition. In *2014 4th IEEE International Conference on Information Science and Technology*, pages 602–605.

Wang, J., Kim, S., and Lee, Y. (2019). Speech augmentation using wavenet in speech recognition. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6770–6774.

Wang, Y. and Guan, L. (2008). Recognizing human emotional state from audiovisual signals*. *IEEE Transactions on Multimedia*, 10(5):936–946.

Wu, C.-H., Chang, H.-T., and Amjad, A. (2020). Eye in-painting using WGAN-GP for face images with mosaic. In Su, R., editor, *2020 International Conference on Image, Video Processing and Artificial Intelligence*, volume 11584, pages 146 – 149. International Society for Optics and Photonics, SPIE.

Xie, J., Song, Z., Li, Y., Zhang, Y., Yu, H., Zhan, J., Ma, Z., Qiao, Y., Zhang, J., and Guo, J. (2018). A survey on machine learning-based mobile big data analysis: Challenges and applications. *Wireless Communications and Mobile Computing*, 2018:8738613.

Ye, Y., Lao, L., Yan, D., and Wang, R. (2020). Identification of weakly pitch-shifted voice based on convolutional neural network. *International Journal of Digital Multimedia Broadcasting*, 2020:8927031.

Yoon, S., Byun, S., and Jung, K. (2018). Multimodal speech emotion recognition using audio and text.

Zhang, S., Loweimi, E., Bell, P., and Renals, S. (2019). Windowed attention mechanisms for speech recognition. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7100–7104.

Zhang, Y.-D., Zhang, Y., Hou, X.-X., Chen, H., and Wang, S.-H. (2018). Seven-layer deep neural network based on sparse autoencoder for voxelwise detection of cerebral microbleed. *Multimedia Tools and Applications*, 77(9):10521–10538.

Zheng, Q., Ke, Y., and Wang, H. (2020). Design and evaluation of cooling workwear for miners in hot underground mines using pcms with different temperatures. *International Journal of Occupational Safety and Ergonomics*, 0(0):1–11. PMID: 32276569.

Figure 1

Structure of a Deep Neural Network
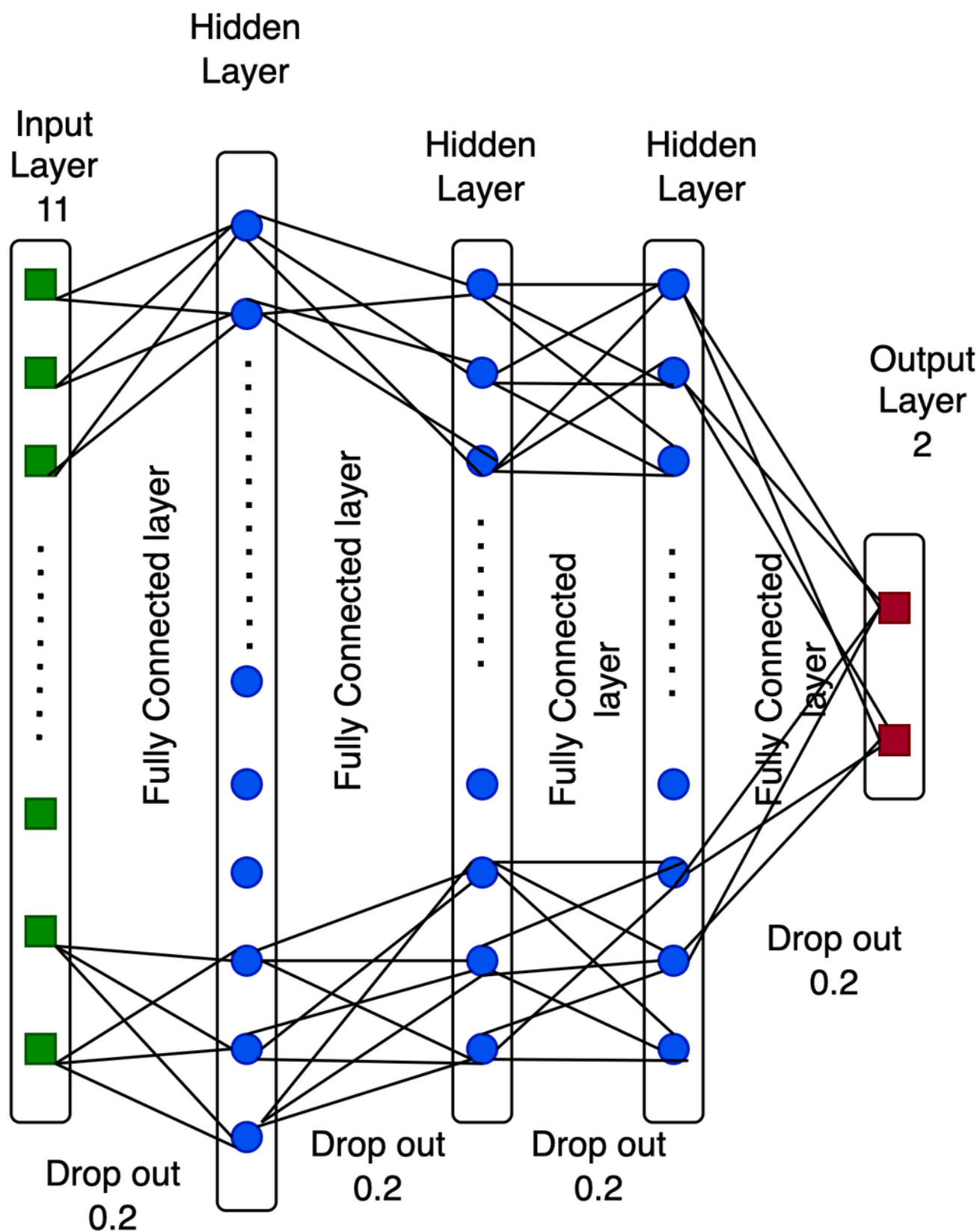
Figure 2

Structure of Proposed Approach
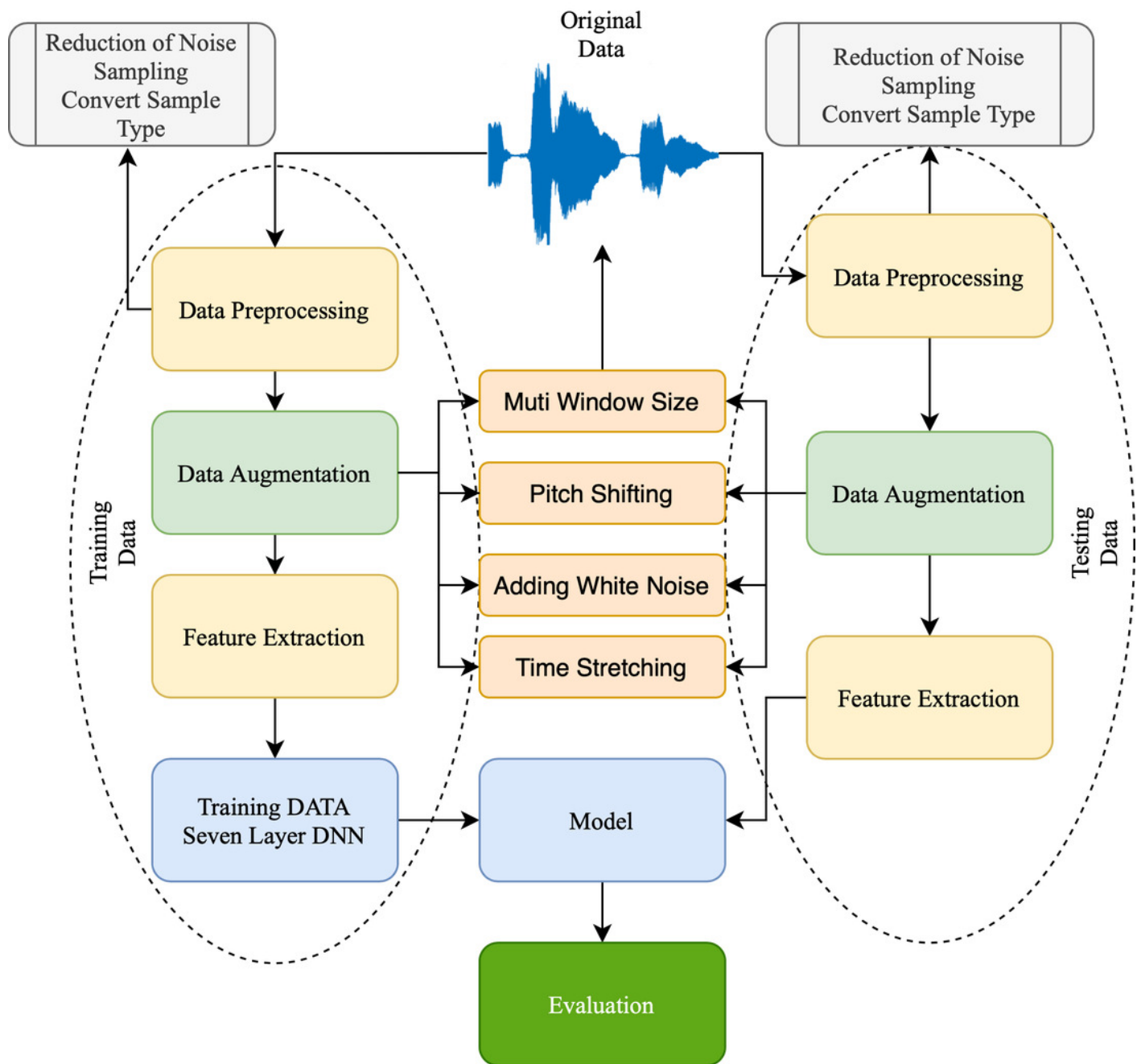
Figure 3

Block Diagram of the computation steps of MFCC
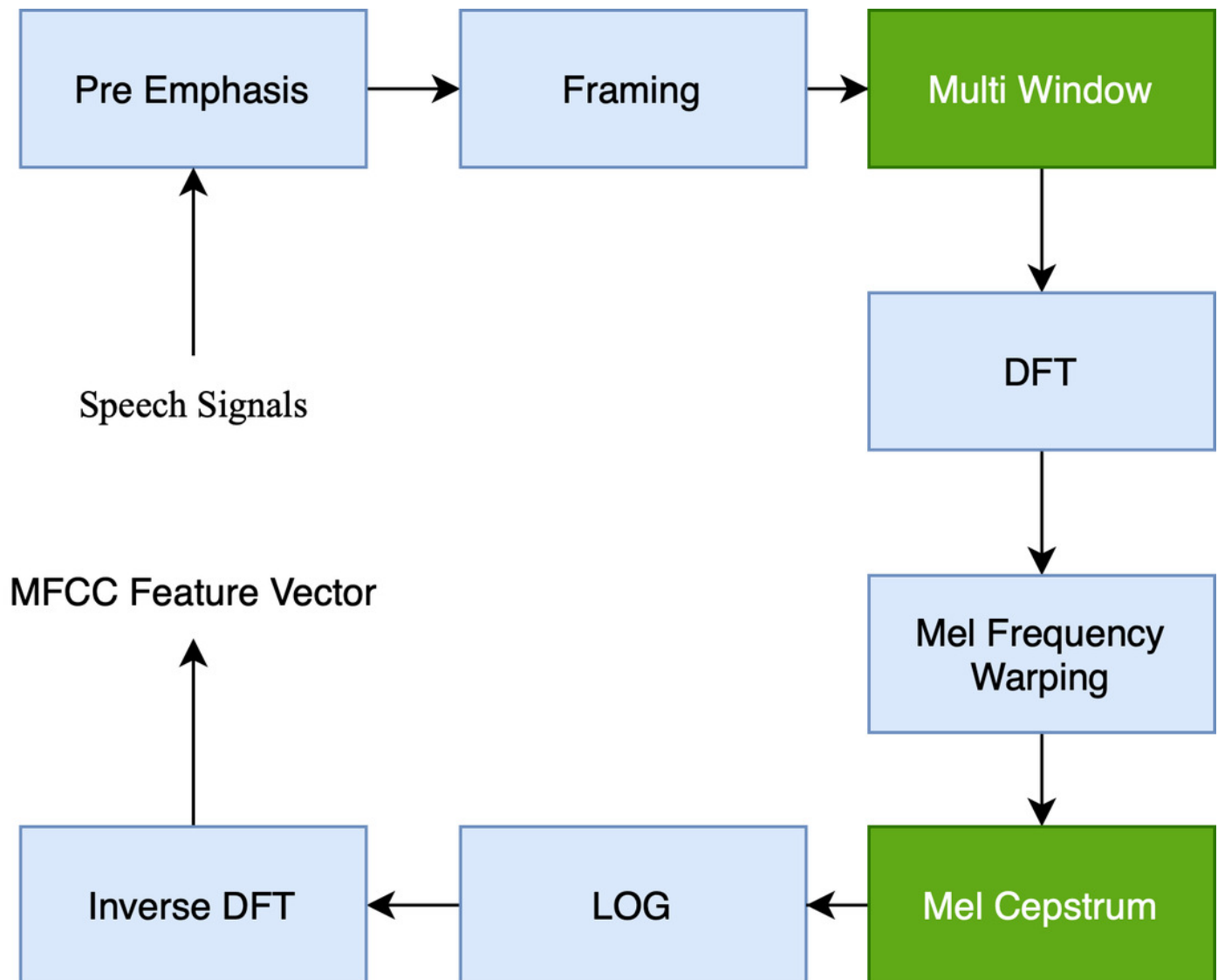
Figure 4

Structure of Proposed Approach
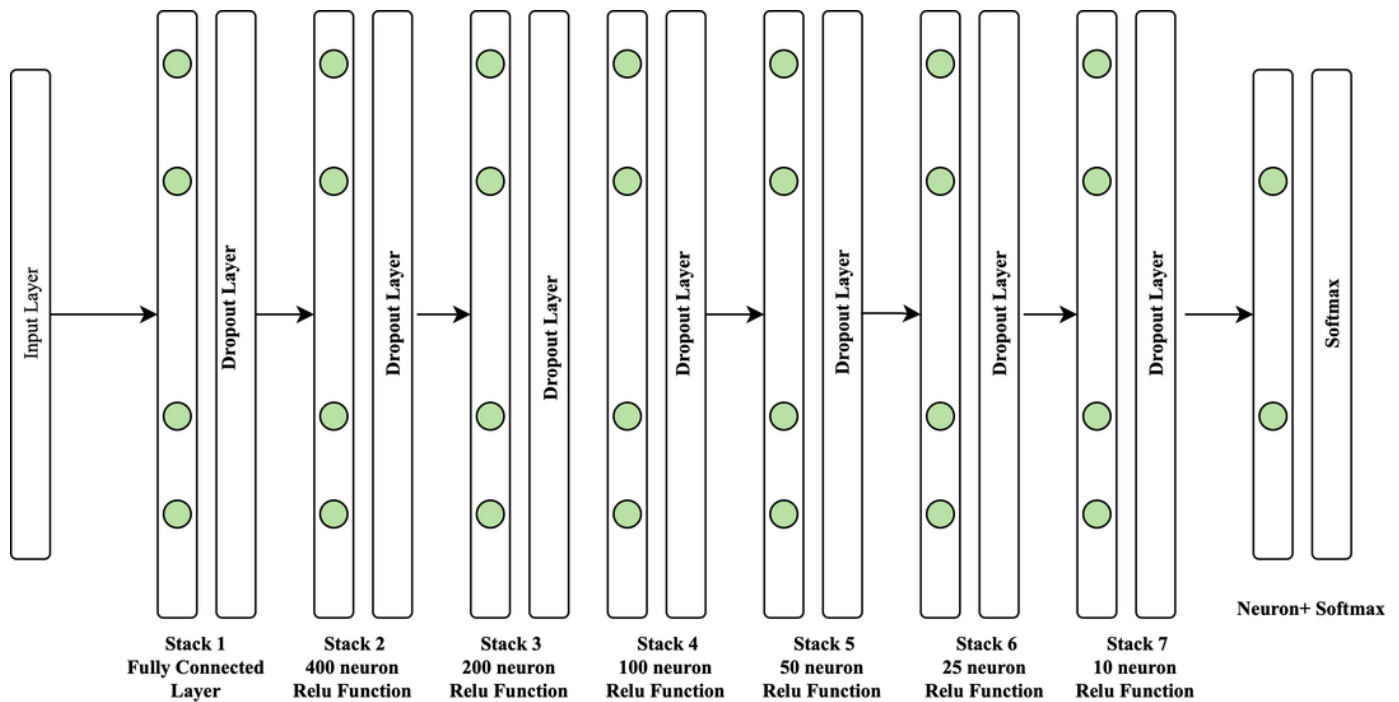
Figure 5

Proposed model performance on training dataset

Figure 6

Proposed model performance on testing dataset