PeerJ Computer Science

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, Taiwan
[2] Bachelor Program in Artificial Intelligence, Chang Gung University, Taoyaun, Taiwan
[3] Department of Physical Medicine and Rehabilitation, Chang Gung Memorial Hospital, Taoyuan, Taiwan
[4] Artificial Intelligence Research Center, Chang Gung University, Taoyuan, Taiwan

ABSTRACT

Speech emotion recognition (SER) systems have evolved into an important method for recognizing a person in several applications, including e-commerce, everyday interactions, law enforcement, and forensics. The 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 works 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.

Subjects Artificial Intelligence, Data Science, Natural Language and Speech
Keywords Speaker recognition, Data augmentation, Deep neural network, Multiple window size

Submitted 3 February 2022
Accepted 6 July 2022
Published 3 August 2022

Corresponding author
Hsien-Tsung Chang,
smallpig@widelab.org

Academic editor
Ali Kashif Bashir

Additional Information and Declarations can be found on page 15

DOI **10.7717/peerj-cs.1053**

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 by 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 & Rani, 2019*; *Khan et al., 2022a*). By 2025, the World Economic Forum predicts that the world will

create 463 exabytes of data every day. Finding the appropriate method to convert such a large volume of data into useful information is difficult.

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, Evin & Verrastro, 2016*; *Maghsoodi et al., 2019*; *Mouaz, Abderrahim & Abdelmajid, 2019*), and the hidden Markov model (HMM) technique (*Veena & Mathew, 2015*; *Bao & Shen, 2016*; *Chakroun et al., 2016*; *Maurya, Kumar & Agarwal, 2018*). However, as the data increases, the level of accuracy with these techniques drops rapidly, to the point where these traditional ML approaches suffer 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, Yamamoto & Nakagawa, 2015*; *Najafian et al., 2016*; *Matjka et al., 2016*; *Dumpala & Kopparapu, 2017*; *Snyder et al., 2018*; *Najafian & Russell, 2020*; *Rohdin et al., 2020*; *Khan et al., 2021*; *Amjad, Khan & Chang, 2021b*, *2021a*; *Khan et al., 2022b*) and convolutional neural network (CNN) methodologies used in the study (*Ravanelli & Bengio, 2019*) attained an overall accuracy of 85% with the TIMIT database and 96% with LibriSpeech. Using the deep learning technique, *An, Thanh & Liu (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 unsuitable for SER utilization 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, Chang & Amjad, 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, Ke & Wang, 2020*). For example, multiple SER studies using DA (*Schlüter & Grill, 2015*; *Salamon & Bello, 2017*; *Pandeya & Lee, 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, Jerez & Franco, 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*). As presented in *Ye et al. (2020)* pitch shifting is frequently utilized in DA and achieved 90% accuracy. In addition, *Damskägg & Välimäki (2017)* employed the time-stretched data augmentation approach when performing DA-based fuzzy identification on various audio signals. *Aguiar, Costa & Silla (2018)* incorporated 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)* reported 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 was 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 & 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, Fang & Wu, 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 article is divided into sections. The Introduction describes the significant issue and the studies done by the speaker; 'Related works' includes many existing works that support the proposed study; 'Data augmentation' describes data augmentation and several methodologies that are used in the research. The next section discusses DNNs, and the deep learning techniques employed. The methodology is covered in the next section, followed by the research outcomes and a discussion. Finally, the 'Conclusion' section covers various significant things about the conclusion of the research outcomes.

RELATED WORKS

The proposed study on multi-racial voice recognition was carried out in many nations, like China (*Nassif et al., 2019*), Africa (*Oyo & Kalema, 2014*), Italy (*Najafian & Russell, 2020*), Pakistan (*Syed et al., 2020*; *Qasim et al., 2016*), the United States (*Upadhyay & Lui, 2018*), and India, through CNN and MFCC (*Ashar, Bhatti & Mushtaq, 2020*). It is a vital technique that many researchers have chosen to enhance SER efficacy (*Chowdhury & 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 10-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, Adriani & Jatmiko, 2020*).

However, this investigation has limits regarding the number of emotional categories. Various types of SER studies have been conducted. For example, *Durrani & 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 two 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, Abderrahim & Abdelmajid, 2019*).

A single-layer DNN with a data augmentation approach was 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 & Frikha, 2020*). This accuracy is higher than the traditional GMM. Another approach was employed in a study (*Hanifa, Isa & Mohamad, 2020*) that used 52 recordings of Malaysian recorded samples utilizing the MFCC in the feature extraction, with 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.

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, Tao & Tang, 2019*). Furthermore, DA is a useful method for overcoming overfitting problems, enhancing model dependability, and increasing generalization (*Wang, Kim & Lee, 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, Jerez & Franco, 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.

Table 1 Detailed description of datasets.

Reference	Approach	Database	Classes	Accuracy
Wang, Wang & Liu (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 & Bello (2017)	SB-CNN	Urban- Sound8K	10	94%
Upadhyay & Lui (2018)	Deep Belief Network	FAS Database	6	90.2%
Singh (2018)	Fuzzy Vector Quantization	TIMIT	100	98.8%
Mouaz, Abderrahim & Abdelmajid (2019)	HMM One layer Deep Neural Network	VOCE *Corpus* Dataset	4	90%
Ashar, Bhatti & Mushtaq (2020)	CNN	Spontaneous Urdu dataset	–	87.5%
Azizah, Adriani & Jatmiko (2020)	DNNs	Indonesian speech *corpus*	4	98.96%
Chakroun et al. (2016)	GMM	TIMIT	8	98.44%
Hanifa, Isa & Mohamad (2020)	Support Vector Machine	speaker ethnicity	4	56.96%
Hanifa, Isa & Mohamad (2020)	DNN	OC16	2	86.10%

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 & Grill, 2015*; *Aguiar, Costa & Silla, 2018*; *Hu, Tan & Qian, 2018*). Furthermore, when white noise is added to an original sound gives a distinct sound effect, which increases the performance of SER.

Pitch shifting: is a commonly used method in an audio sample to increase or decrease the original tone of voice. Pitch variations are performed using this technique without affecting playback speed (*Mousa, 2010*). In addition, a method is utilized in pitch shifting to increase the pitch of the original sound without changing the duration of the recorded sound clip (*Rai & Barkana, 2019*). For example, various studies on singing voice detection (SVD) (*Gui et al., 2021*), environmental sound classification (ESC) (*Salamon & Bello, 2017*), and domestic cat classification have shown that pitch shifting may be highly effective for DA (*Pandeya & Lee, 2018*).

Time stretching: is a way to change the speed or length of an audio signal without changing the tone. Instead, it is used to manipulate audio signals (*Damskägg & Välimäki, 2017*). This technique is suitable for analyzing auditory signals that comprise tone, noise, and temporal elements. Numerous investigations used time stretching with other approaches such as synchronous overlap, fuzzy, and CNN to increase the efficiency of the suggested framework (*Sasaki et al., 2010*; *Kupryjanow & Czyżewski, 2011*; *Salamon &*

Bello, 2017). These studies used different techniques, such as the synchronous overlap algorithm, fuzzy logic, and CNN, to improve the performance of the proposed model.

Multiple window size: Multiple window size features are retrieved from a windowed signal called frames. The window strongly influences the obtained features retrieved from the voice signal-based functions width since signals are often steady for limited periods (*Kelly & 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 & 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, Khan & Tariq, 2016*; *Kirkpatrick, O'Brien & Scaife, 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, Byun & Jung, 2018*; *Tarantino, Garner & Lazaridis, 2019*; *Ramet et al., 2018*). On the other hand, some research has indicated that a larger window size improves emotion identification performance (*Chernykh & Prikhodko, 2018*; *Tripathi, Tripathi & Beigi, 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, Garner & Lazaridis, 2019*; *Chernykh & Prikhodko, 2018*). *Tarantino, Garner & Lazaridis (2019)* investigated the influence of overlap window size on SER. They discovered that a small step size leads to a lower test loss. *Chernykh & Prikhodko (2018)*, explored multiple window widths ranging from 30 to 200 ms before settling on a unique 200 ms window for the SER study.

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, Yamamoto & Nakagawa, 2015*; *Snyder et al., 2018*; *Novotny et al., 2018*; *Saleem & Irfan Khattak, 2020*).

The structure of a DNN approach is composed of input, hidden, dropout, and output layers (*Rajyaguru, Vithalani & Thanki, 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.

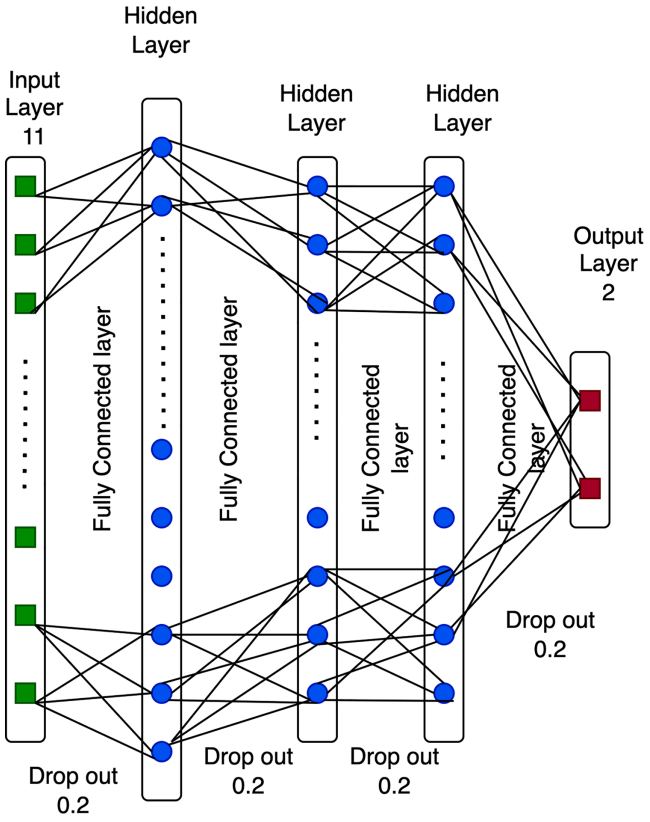

Figure 1 **Structure of a deep neural network.** Full-size DOI: 10.7717/peerj-cs.1053/fig-1

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.

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.

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 Fig. 1, since DNN obtains the highest efficiency with this value.

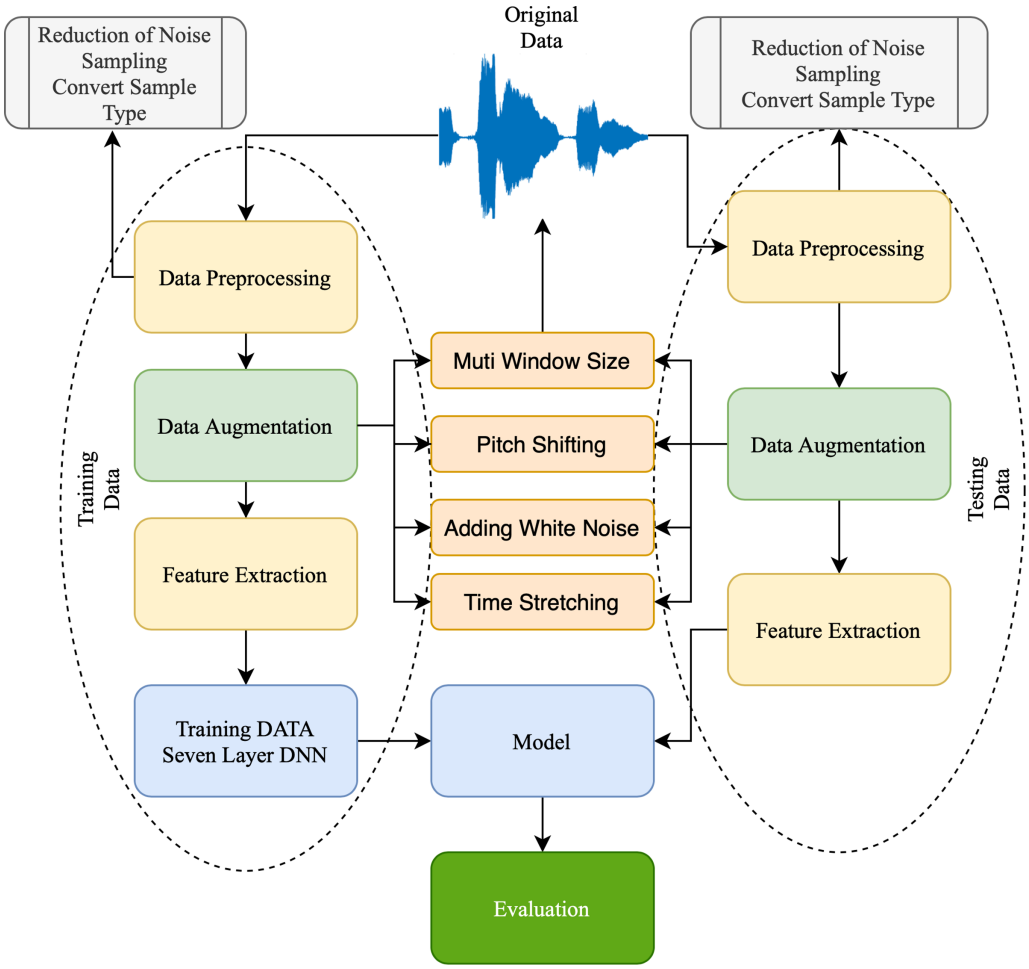

Figure 2 **Structure of proposed approach.** Full-size DOI: 10.7717/peerj-cs.1053/fig-2

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. Figure 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 and processes the features 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 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 & Guan, 2008*)	4 males and 4 females	42 s	500 samples	Speaker and text independent
Urdu (*Wang & Guan, 2008*; *Syed et al., 2020*)	4 males and 4 females	42 s	500 samples	Speaker and text independent
Sindhi (*Syed et al., 2020*)	32 males and 38 females	30 s	80 samples	Speaker and text independent
Saraiki	42 males and 28 females	30 s	80 samples	Speaker and text independent
Pashto	35 males and 35 females	30 s	80 samples	Speaker and text independent

DATASET AND PREPROCESSING

This study utilized a dataset of Pakistan's five most spoken local languages. The information was obtained to adjust for the numerous ethnicities. Various online resources were used to compile this dataset (*Wang & 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 s 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 of five racials, for 1,240 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 & Akagi, 2020*). Experiments on RAVDESS were carried out by *Livingstone & 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

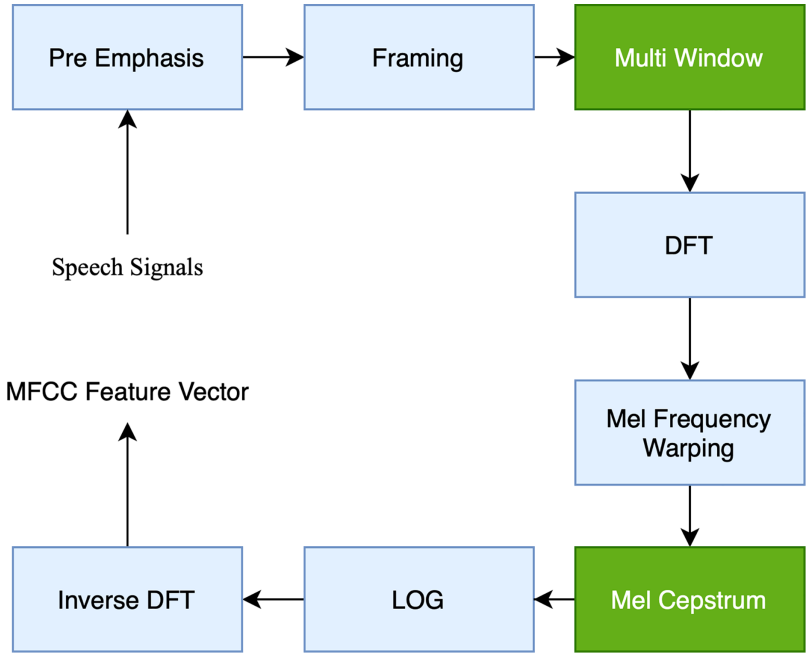

Figure 3 **Block diagram of the computation steps of MFCC.**
Full-size DOI: 10.7717/peerj-cs.1053/fig-3

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 10 frameshifts. The MFCC approach enhances the audio sound input during the preemphasis phase, increasing the signal-to-noise ratio (SNR) enough to ensure

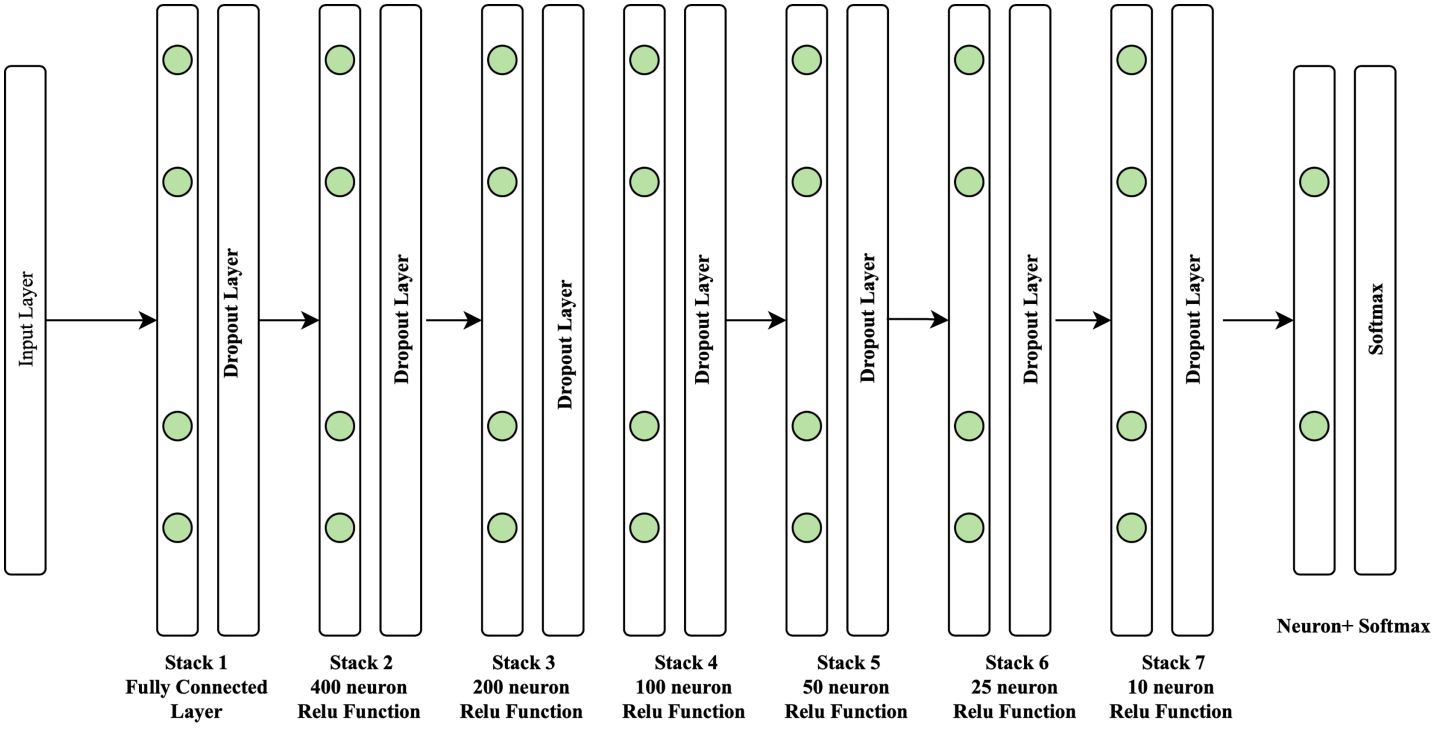

Figure 4 **Structure of proposed approach.** Full-size DOI: 10.7717/peerj-cs.1053/fig-4

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.

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 & 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 two, 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 10 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.

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

RESULTS AND DISCUSSION

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

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.

However, 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

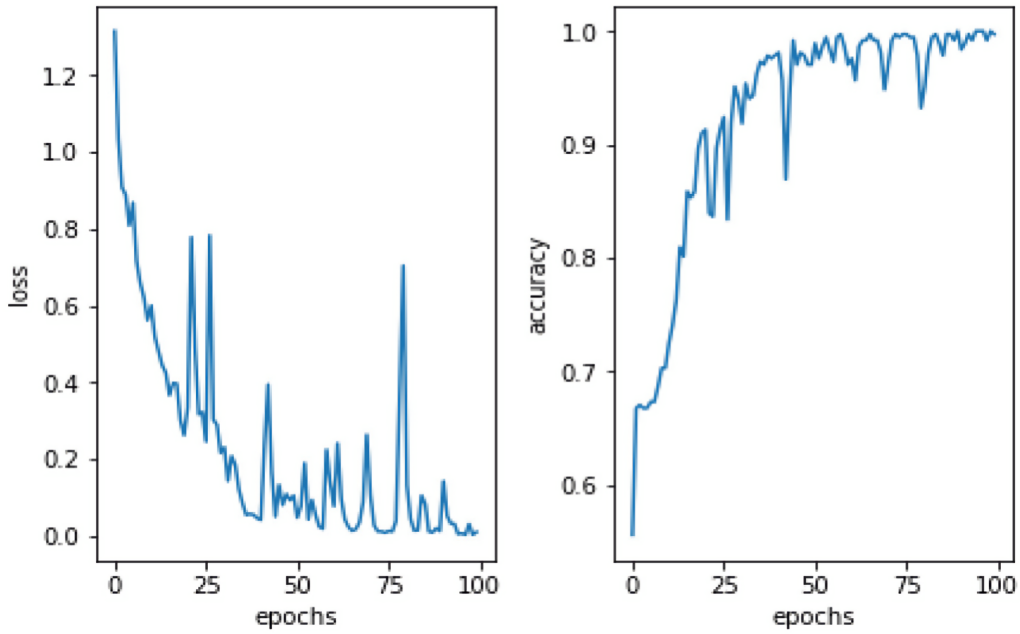

Figure 5 **Proposed model performance on training dataset.**
Full-size DOI: 10.7717/peerj-cs.1053/fig-5

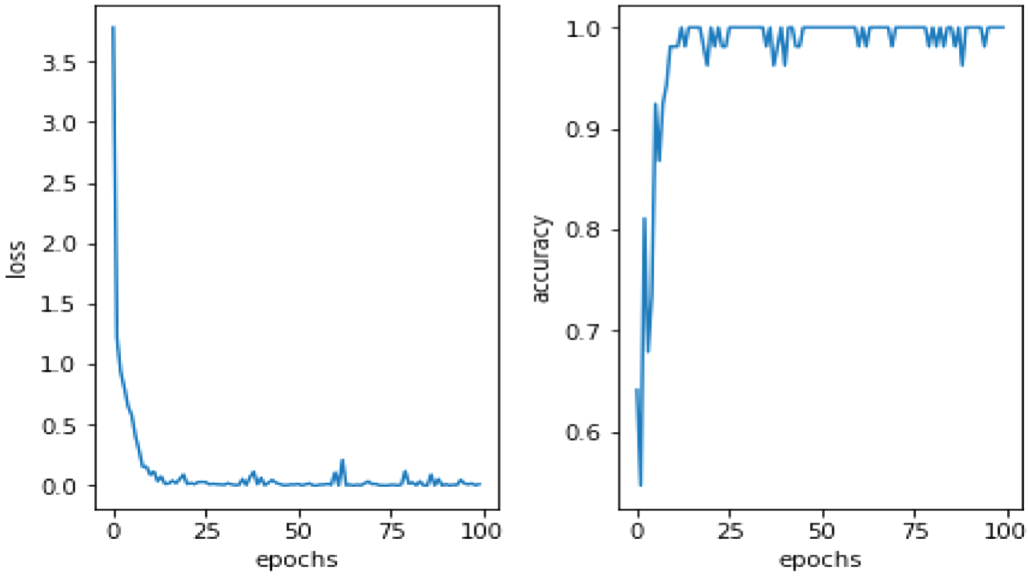

Figure 6 **Proposed model performance on testing dataset.**
Full-size DOI: 10.7717/peerj-cs.1053/fig-6

testing datasets. As shown in Tables 4–6, another test was conducted by adding 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,

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

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

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

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

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

Dataset	Classification accuracy	Accuracy
Pakistani racial speaker classification	KNN	81.99
	Random Forest	71.56
	Multilayer Perceptron (MLP)	91.45
	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

between many classes while generating predictions (*Silva-Palacios, Ferri & Ramírez-Quintana, 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.

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.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Ammar Amjad conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

- Lal Khan conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Hsien-Tsung Chang conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The code is available in the Supplemental File.

The third-party datasets are available at:

- http://shachi.org/resources/4965

- Zafi Sherhan Syed, Sajjad Ali Memon, Muhammad Shehram Shah and Abbas Shah Syed, "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 (IJACSA), 11(4), 2020. DOI 10.14569/IJACSA.2020.01104104.

- Z. Xie and L. Guan, "Multimodal Information Fusion of Audio Emotion Recognition Based on Kernel Entropy Component Analysis," 2012 IEEE International Symposium on Multimedia, 2012, pp. 1-8, DOI 10.1109/ISM.2012.9.

- Zhibing Xie and L. Guan, "Multimodal information fusion of audiovisual emotion recognition using novel information theoretic tools," 2013 IEEE International Conference on Multimedia and Expo (ICME), 2013, pp. 1-6, DOI 10.1109/ICME.2013.6607464.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.1053#supplemental-information.

REFERENCES

Afrillia Y, Mawengkang H, Ramli M, Fadlisyah, Fhonna RP. 2017. Performance measurement of mel frequency ceptral coefficient (MFCC) method in learning system Of Al- Qur'an based in *Nagham* pattern recognition. *Journal of Physics: Conference Series* **930**:12036 DOI 10.1088/1742-6596/930/1/012036.

Aguiar RL, Costa YM, Silla CN. 2018. Exploring data augmentation to improve music genre classification with convnets. In: *2018 International Joint Conference on Neural Networks (IJCNN)*. 1–8.

Amjad A, Khan L, Ashraf N, Mahmood MB, Chang H-T. 2022. Recognizing semi-natural and spontaneous speech emotions using deep neural networks. *IEEE Access* **10**:37149–37163 DOI 10.1109/ACCESS.2022.3163712.

Amjad A, Khan L, Chang H-T. 2021a. Effect on speech emotion classification of a feature selection approach using a convolutional neural network. *PeerJ Computer Science* **7(10)**:e766 DOI 10.7717/peerj-cs.766.

Amjad A, Khan L, Chang H-T. 2021b. Semi-natural and spontaneous speech recognition using deep neural networks with hybrid features unification. *Processes* **9(12)**:2286 DOI 10.3390/pr9122286.

An NN, Thanh NQ, Liu Y. 2019. Deep CNNs with self-attention for speaker identification. *IEEE Access* **7**:85327–85337 DOI 10.1109/ACCESS.2019.2917470.

Ashar A, Bhatti MS, Mushtaq U. 2020. Speaker identification using a hybrid CNN-MFCC approach. In: *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*. Piscataway: IEEE, 1–4.

Atmaja BT, Akagi M. 2020. On the differences between song and speech emotion recognition: effect of feature sets, feature types, and classifiers. In: *2020 IEEE Region 10 Conference (TENCON)*. Piscataway: IEEE, 968–972.

Azizah K, Adriani M, 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 DOI 10.1109/ACCESS.2020.3027619.

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

Chakroun R, Beltaïfa Zouari L, Frikha M, 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)*. 693–696.

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

Chernykh V, Prikhodko P. 2018. Emotion recognition from speech with recurrent neural networks. *ArXiv preprint*. DOI 10.48550/arXiv.1701.08071.

Chowdhury A, Ross A. 2020. Fusing MFCC and LPC features using 1D triplet CNN for speaker recognition in severely degraded audio signals. *IEEE Transactions on Information Forensics and Security* **15**:1616–1629 DOI 10.1109/TIFS.2019.2941773.

Damskägg E-P, Välimäki V. 2017. Audio time stretching using fuzzy classification of spectral bins. *Applied Sciences* **7(12)**:1293 DOI 10.3390/app7121293.

Dumpala SH, Kopparapu SK. 2017. Improved speaker recognition system for stressed speech using deep neural networks. In: *2017 International Joint Conference on Neural Networks (IJCNN)*. 1257–1264.

Durrani S, Arshad MU. 2021. Transfer learning based speech affect recognition in Urdu. *ArXiv preprint*. DOI 10.48550/arXiv.2103.03580.

Gui W, Li Y, Zang X, Zhang J. 2021. Exploring channel properties to improve singing voice detection with convolutional neural networks. *Applied Sciences* **11(24)**:11838 DOI 10.3390/app112411838.

Gupta D, Rani R. 2019. A study of big data evolution and research challenges. *Journal of Information Science* **45(3)**:322–340 DOI 10.1177/0165551518789880.

Hamidi M, Satori H, Zealouk O, Satori K. 2020. Amazigh digits through interactive speech recognition system in noisy environment. *International Journal of Speech Technology* **23(1)**:101–109 DOI 10.1007/s10772-019-09661-2.

Hanifa RM, Isa K, Mohamad S. 2020. Speaker ethnic identification for continuous speech in Malay language using pitch and MFCC. *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)* **19(1)**:207–214 DOI 10.11591/ijeecs.v19.i1.pp207-214.

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

Kelly AC, Gobl C. 2011. The effects of windowing on the calculation of MFCCS for different types of speech sounds. In: Travieso-González CM, Alonso-Hernández JB, eds. *Advances in Nonlinear Speech Processing*. Berlin, Heidelberg, Berlin Heidelberg: Springer, 111–118.

Khan L, Amjad A, Afaq KM, Chang H-T. 2022a. Deep sentiment analysis using CNN-LSTM architecture of English and Roman Urdu text shared in social media. *Applied Sciences* **12(5)**:2694 DOI 10.3390/app12052694.

Khan L, Amjad A, Ashraf N, Chang H-T. 2022b. Multi-class sentiment analysis of Urdu text using multilingual BERT. *Scientific Reports* **12(1)**:5436 DOI 10.1038/s41598-022-09381-9.

Khan L, Amjad A, Ashraf N, Chang H-T, Gelbukh A. 2021. Urdu sentiment analysis with deep learning methods. *IEEE Access* **9**:97803–97812 DOI 10.1109/ACCESS.2021.3093078.

Kingma DP, Ba J. 2017. Adam: a method for stochastic optimization. *ArXiv preprint*. DOI 10.48550/arXiv.1412.6980.

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

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

Kupryjanow A, Czyżewski A. 2011. A non-uniform real-time speech time-scale stretching method. In: *Proceedings of the International Conference on Signal Processing and Multimedia Applications*. 1–7.

Li Z, Wang S-H, Fan R-R, Cao G, Zhang Y-D, Guo T. 2019. Teeth category classification via seven-layer deep convolutional neural network with max pooling and global average pooling. *International Journal of Imaging Systems and Technology* **29(4)**:577–583 DOI 10.1002/ima.22337.

Li X, Zhang W, Ding Q, Sun J-Q. 2020. Intelligent rotating machinery fault diagnosis based on deep learning using data augmentation. *Journal of Intelligent Manufacturing* **31(2)**:433–452 DOI 10.1007/s10845-018-1456-1.

Liu J, Fang C, Wu C. 2016. A fusion face recognition approach based on 7-layer deep learning neural network. *Journal of Electrical and Computer Engineering* **2016**:8637260 DOI 10.1155/2016/8637260.

Livingstone SR, Russo FA. 2018. The Ryerson audio-visual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in north American English. *PLOS ONE* **13(5)**:e0196391 DOI 10.1371/journal.pone.0196391.

Long Y, Li Y, Zhang Q, Wei S, Ye H, Yang J. 2020. Acoustic data augmentation for Mandarin-English code-switching speech recognition. *Applied Acoustics* **161**:107175 DOI 10.1016/j.apacoust.2019.107175.

Ma R, Tao P, Tang H. 2019. Optimizing data augmentation for semantic segmentation on small-scale dataset. In: *Proceedings of the 2nd International Conference on Control and Computer Vision, ICCCV 2019*. New York, NY, USA: Association for Computing Machinery, 77–81.

Maghsoodi N, Sameti H, Zeinali H, 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 DOI 10.1109/TASLP.2019.2928143.

Marufo da Silva M, Evin DA, Verrastro S. 2016. Speaker-independent embedded speech recognition using hidden Markov models. In: *IEEE CACIDI 2016 – IEEE Conference on Computer Sciences*. Piscataway: IEEE, 1–6.

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

Maurya A, Kumar D, Agarwal R. 2018. Speaker recognition for Hindi speech signal using MFCC-GMM approach. *Procedia Computer Science* **125(2)**:880–887 The 6th International Conference on Smart Computing and Communications DOI 10.1016/j.procs.2017.12.112.

Mohammed MA, Abdulkareem KH, Mostafa SA, Khanapi Abd Ghani M, Maashi MS, Garcia-Zapirain B, Oleagordia I, Alhakami H, AL-Dhief FT. 2020. Voice pathology detection and classification using convolutional neural network model. *Applied Sciences* **10(11)**:3723 DOI 10.3390/app10113723.

Moreno-Barea FJ, Jerez JM, Franco L. 2020. Improving classification accuracy using data augmentation on small data sets. *Expert Systems with Applications* **161**:113696 DOI 10.1016/j.eswa.2020.113696.

Mouaz B, Abderrahim BH, 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 DOI 10.1016/j.procs.2019.04.138.

Mousa A. 2010. Voice conversion using pitch shifting algorithm by time stretching with PSOLA and re-sampling. *Journal of Electrical Engineering* **61(1)**:2011 DOI 10.2478/v10187-010-0008-5.

Najafian M, Russell M. 2020. Automatic accent identification as an analytical tool for accent robust automatic speech recognition. *Speech Communication* **122(10–11)**:44–55 DOI 10.1016/j.specom.2020.05.003.

Najafian M, Safavi S, Hansen JHL, 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)*. Piscataway: IEEE, 1–6.

Nassif AB, Shahin I, Attili I, Azzeh M, Shaalan K. 2019. Speech recognition using deep neural networks: a systematic review. *IEEE Access* **7**:19143–19165 DOI 10.1109/ACCESS.2019.2896880.

Nisar S, Khan OU, Tariq M. 2016. An efficient adaptive window size selection method for improving spectrogram visualization. *Computational Intelligence and Neuroscience* **2016(2)**:6172453 DOI 10.1155/2016/6172453.

Novotny O, Plchot O, Glembek O, Cernocky JH, Burget L. 2018. Analysis of DNN speech signal enhancement for robust speaker recognition. *Computer Speech and Language* **58**:403–421 DOI 10.1016/j.csl.2019.06.004.

Nurhaida I, Ayumi V, Fitrianah D, Zen RA, Noprisson H, 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 DOI 10.11591/ijece.v10i2.pp2045-2053.

Oyo B, Kalema BM. 2014. A preliminary speech learning tool for improvement of African English accents. In: *2014 International Conference on Education Technologies and Computers (ICETC)*. 44–48.

Pandeya YR, Lee J. 2018. Domestic cat sound classification using transfer learning. *The International Journal of Fuzzy Logic and Intelligent Systems* **18(2)**:154–160 DOI 10.5391/IJFIS.2018.18.2.154.

Qasim M, Nawaz S, Hussain S, 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)*. 28–32.

Rabiner LR, Schafer RW. 2007. Introduction to digital speech processing. *Foundations and Trends in Signal Processing* **1(1)**:1–194 DOI 10.1561/2000000001.

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

Rajyaguru V, Vithalani C, Thanki R. 2020. A literature review: various learning techniques and its applications for eye disease identification using retinal images. *International Journal of Information Technology* **2020**:1–12 DOI 10.1007/s41870-020-00442-8.

Ramet G, Garner PN, Baeriswyl M, Lazaridis A. 2018. Context-aware attention mechanism for speech emotion recognition. In: *2018 IEEE Spoken Language Technology Workshop (SLT)*. Piscataway: IEEE, 126–131.

Ravanelli M, Bengio Y. 2019. Speaker recognition from raw waveform with SincNet. In: *2018 IEEE Spoken Language Technology Workshop (SLT)*. Piscataway: IEEE, 1021–1028 DOI 10.1109/SLT.2018.8639585.

Rebai I, BenAyed Y, Mahdi W, 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 DOI 10.1016/j.procs.2017.08.003.

Rituerto-Gonzlez E, Mnguez-Snchez A, Gallardo-Antoln A, Pelez-Moreno C. 2019. Data augmentation for speaker identification under stress conditions to combat gender-based violence. *Applied Sciences* **9(11)**:2298 DOI 10.3390/app9112298.

Rohdin J, Silnova A, Diez M, Plchot O, Matjka P, Burget L, Glembek O. 2020. End-to-end DNN based text-independent speaker recognition for long and short utterances. *Computer Speech & Language* **59(2–3)**:22–35 DOI 10.1016/j.csl.2019.06.002.

Salamon J, Bello JP. 2017. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters* **24(3)**:279–283 DOI 10.1109/LSP.2017.2657381.

Saleem N, Irfan Khattak M. 2020. Deep neural networks based binary classification for single channel speaker independent multi-talker speech separation. *Applied Acoustics* **167(5)**:107385 DOI 10.1016/j.apacoust.2020.107385.

Sasaki T, Nakajima Y, ten Hoopen G, van Buuringen E, Massier B, Kojo T, Kuroda T, Ueda K. 2010. Time stretching: illusory lengthening of filled auditory durations. *Attention, Perception, & Psychophysics* **72(5)**:1404–1421 DOI 10.3758/APP.72.5.1404.

Schlüter J, Grill T. 2015. Exploring data augmentation for improved singing voice detection with neural networks. In: *International Society for Music Information Retrieval (ISMIR)*.

Seifert C, Aamir A, Balagopalan A, Jain D, Sharma A, Grottel S, Gumhold S. 2017. *Visualizations of deep neural networks in computer vision: a survey*. Netherlands: Studies in Big Data, Cham: Springer, 123–144.

Seki H, Yamamoto K, 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)*. 1222–1225.

Silva-Palacios D, Ferri C, Ramírez-Quintana MJ. 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 DOI 10.1016/j.procs.2017.05.218.

Singh S. 2018. Speaker recognition by Gaussian filter based feature extraction and proposed fuzzy vector quantization modelling technique. *Quantization Modelling Technique* **13(16)**:12798–12804.

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

Syed ZS, Memon SA, Shah MS, Syed AS. 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)**:1–6 DOI 10.14569/issn.2156-5570.

Tarantino L, Garner PN, Lazaridis A. 2019. Self-attention for speech emotion recognition. In: *INTERSPEECH*.

Tripathi S, Tripathi S, Beigi H. 2019. Multi-modal emotion recognition on IEMOCAP dataset using deep learning. *ArXiv preprint*. DOI 10.48550/arXiv.1804.05788.

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

Veena KV, Mathew D. 2015. Speaker identification and verification of noisy speech using multitaper MFCC and Gaussian mixture models. In: *2015 International Conference on Power, Instrumentation, Control and Computing (PICC)*. 1–4.

Wang Y, Guan L. 2008. Recognizing human emotional state from audiovisual signals*. *IEEE Transactions on Multimedia* **10(5)**:936–946 DOI 10.1109/TMM.2008.927665.

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

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

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

Xie J, Song Z, Li Y, Zhang Y, Yu H, Zhan J, Ma Z, Qiao Y, Zhang J, Guo J. 2018. A survey on machine learning-based mobile big data analysis: challenges and applications. *Wireless Communications and Mobile Computing* **2018(12)**:8738613 DOI 10.1155/2018/8738613.

Ye Y, Lao L, Yan D, Wang R. 2020. Identification of weakly pitch-shifted voice based on convolutional neural network. *International Journal of Digital Multimedia Broadcasting* **2020(1)**:8927031 DOI 10.1155/2020/8927031.

Yoon S, Byun S, Jung K. 2018. Multimodal speech emotion recognition using audio and text. In: *2018 IEEE Spoken Language Technology Workshop (SLT)*. Piscataway: IEEE.

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

Zhang Y-D, Zhang Y, Hou X-X, Chen H, 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 DOI 10.1007/s11042-017-4554-8.

Zheng Q, Ke Y, 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* **28(1)**:1–11 DOI 10.1080/10803548.2020.1730618.