Multimedia Tools and Applications (2022) 81:41309–41338
https://doi.org/10.1007/s11042-022-12411-3

Emotion detection from multilingual audio using deep analysis

Sudipta Bhattacharya[1] · Samarjeet Borah[2] · Brojo Kishore Mishra[1] · Atreyee Mondal[3]

Received: 15 November 2020 / Revised: 11 February 2021 / Accepted: 25 January 2022 /
Published online: 20 May 2022

Abstract

Human emotion detection from multiple languages is a very challenging job. In this work, we have used language emotional databases of various languages such as – Ryerson-Audio-Visual database (RAVDESS), Berlin Database (EmoDb) and Italian Database (Emo-Vo) which are in English, German and Italian languages respectively. The proposed model extract MFCC, chroma, Tonnetz, Contrast from the raw audio file, which is further taken as input in the CNN model to identify emotions correctly. We are not using any visual representation of sound only direct from natural sound data. An extensive comparison is made with some of the previous approaches on emotion detection from speech. The experimental result shows that; the proposed model has successfully worked with all the selected databases with higher accuracy. The same also has been tested with the augmented database. We secure 70.46% for RAVDESS, 70.37% Emo-Db and 73.47% for Emo-Vo in the initial database and best model work in the augmented database. However, test with Original test dataset, secured 96.53% in RAVDESS 96.22% in Emo-Db and Emo-Vo 96.11% respectively. Multilingual Emotion detection, a state of art model, has been discussed with an accuracy of 97.89%. The proposed model is a speaker-independent as well as language-independent emotion detection system.

✉ Samarjeet Borah
samarjeetborah@gmail.com

Sudipta Bhattacharya
sudipta.bhattacharya@giet.edu

Brojo Kishore Mishra
bkmishra@giet.edu

Atreyee Mondal
atreyee.shakshi@gmail.com

[1] School of Computer Engineering (SOCE), GIET University, Gunupur, India

[2] Department of Computer Applications, SMIT, Sikkim Manipal University, Rangpo, Sikkim, India

[3] Department of Information Technology, Techno India College of Technology, Kolkata, India

Keywords Deep learning · Convolution neural network · Speech · Emotion detection · RAVDESS · EmoDb · Emo-vo · MFCC

1 Introduction

Human is emotion-driven, so it plays an essential role to take any action and decision. It helps us match and understand others' feelings by conveying our compassion to others and getting feedback. Emotions expressed by an individual carry information about their mental state. Raw emotions like angry, neutral, disgust, sad, happy, surprise, fear is dependent upon the value of valance and arousal. Value of valance may be positive or negative, and the amount of arousal is high or low represents emotions, emotion detection through, speech recognition from valance and arousal value. A new research area has been open automated emotion recognition to understand desired feelings.

Speech emotion recognition is one kind of pattern recognition system where pattern matched with the desired method. Emotion recognition from the speech information may be the speaker-dependent or speaker-independent. [51]

There are some significant obstacles to implement a successful speech emotion system, which are as follows:

- Proper emotional speech database has been to choose.
- Suitable features have to extract
- Appropriate reliable classifier and machine learning algorithm are required.

Proper implementation of speech emotion recognition system (SER), we need a robust and sustainable model. Emotions have different psychological states like physiological, behavioural and communicative reactions [2].

Emotion detection from the raw sound file itself is a challenging job. Our motivation is to do this work; several studies have been done for sound emotion recognition from the last few decades [7]. However, from raw sound data, significantly less work has been done. Over the last few years, researchers have been using deep learning models to recognize emotion face, voice, and image recognition by deep learning models to solve the problem. [4, 9, 12, 14–17, 26, 29, 39, 41] In deep learning models, features are automatically selected; this is why the deep learning model is trendy [41]. In Literature review, we have mentioned several emotional frameworks which are combined with different features types. By experiment, we have chosen a mix of features from raw sound emotion detection has been proposed, which is a powerful reflection to identification from other models. Especially harmony and pitch class is enriched than before, which is poorly distinguish previously. In this work, we have discussed 5 different features extracted from the sound file, into the 1D CNN model as an input; we have also mixed the elements to get a better result [8] .

We have shown a baseline model firstly and check the performance of the model in the different corpus, we have used EmoVo [11] (Italian Database), EmoDb [10] (German Database), RAVDESS Database [33]. (English Languages) This corpus has two types of database, namely song and speech. We directly use the sound file to extract features and fed to 1D CNN and compare results with various speech emotion framework. Issa. D et al. [20] proposed a model, in which accuracy is better than all previous work except Zhao. et al. [50], Issa.D et al.

[20] extensively done analysis on Emo-Db [10] database, which is very popular; most of the researchers has experimented on it.

In the present work, we proposed a simple 1D convolution neural network model with k fold cross-validation and compared with the three corpora Emo-Vo [11], Emo-Db [10], RAVDNESS [33]. Multilingual database (one mixed with common set emotion) shows better accuracy than Issa.D et al. [20].

The rest of the paper is organized as follows -in section 2, the relevant research on speech emotion is concisely reviewed. In section 3, the database used in the proposed model is mentioned. Section 4 includes the discussion of different features. In section 5, the methodology of the proposed algorithm is illustrated. Also, we discussed a baseline model and some upgrade with model and database after comparing with the result in section 6. Finally, section 7 depicts the conclusion and possible future direction.

2 Literature review

Over the last few decades, many emotion detection techniques have been incorporated [4, 9, 29]. Tables 1, 2 and 3 represent some extensive work done by several researchers, with the limitation of different aspects. Most of the researcher has done several experiments on EmoDb [10] (German Database). Different classifier like SVM, GMM, Binary tree using the linear kernel, time distributed CNN, 2D CNN LSTM, semi-supervised neural network, Artificial Neural Network etc. are applied with features like a spectral representation of speech, MFCC, the first-order derivative of MFCC, pitch, energy etc. Also follows literature survey has been done for RAVDESS Database and Emo-Vo (Italian Database) respectively.

Emotion like anger, disgust, fear, happiness, joy and sadness have their identity [4]. For example, it can be disgust by pitch, intensity, speaking rate and voice quality, pitch abruption on stress identity much higher and voice quality chest tone. Fear broad typical pitch average intensity lower and irregular voicing happiness and joy more or less same, but happiness speaking rate is faster or slower, and satisfaction speaking rate is always faster. Corpus information of various databases under consideration are presented in Table 4.

3 Database

Lots of emotion database are available, both commercially and non-commercially. Some database is based on different languages like German, Japanese, English and many more. We have chosen three other languages and one different speech of limited actor database in this case. These all are well-known database in the research community. Emo-DB, Emo-Vo, RAVDESS has been used into our model as a database.

3.1 RAVDESS

RAVDESS [33] (Ryerson Audio-Visual Database), a North -American British Accent Audio and emotional video database, is chosen. In RAVDESS, dataset categories into some part, audio and video. The audio element has two types, namely a song and speech database. In this database, a total of 24 male and female actor consists of speech and song database. Actor

Table 1 Review of works on EmoDB [10] database

Year	Literature	Database description	Augmented data	Features	Classifier	Accuracy	Limitation
2010	Luengo et al. [24]	German Languages Database Emotion presents into that database is boredom anger, sadness, fear, disgust, neutral, happiness and number of the audio file is 535.	No	Spectral, intonation and intensity regression features, sentence end, voice quality features statistics, speech rate etc.	SVM	78.30%	The original database has 535 utterances with 16KHz with 16 bits per sample, but in this paper, it has been subsampled in 8KHz.
2011	Wu et al. [47]		Yes	Modulation spectral Features (MSF_S) with prosodic Features	Multiclass linear discriminant analysis(LDA) Classifier.	85.8%	No proper limitation has been mention.
2012	Lampropoulos and Tshisintzis [30]		No	Combine MPEG-7dscriptor, MFCC and Timbral Features	SVM with RBF kernel using leave one out evaluation	83.93%	No proper limitation has been found.
2014	Pohjalainen et al. [27]		No	MFCC +1st and 2nd derivatives	GMM	68.49%	Authors claim that further improvement can be possible to automatically select training data at the time of training of GMM based model.
2014	Huang et al. [23]		yes	Spectrogram	CNN input layer with the last layer with SVM classifier	88.3% speaker-dependent and 85.2% speaker--independent	Utilize Spectogram
2014	Yüncü et al. [21]		No	Principal components (five)	SVM utilizing binary decision tree	72.30%	No proper limitation has been mentioned.
2015	Kadiri et al. [28]		No	MFCC, MSF, PLP, and prosody features	SVM	75.22%	From Emo-Db database they have studied only 339 out of 535 utterances only consider 4 emotions
2015	Sinith et al. [44]		No	MFCC, pitch and energy	Binary tree using a linear kernel	73.75%	The best result shows their creation of Malayalam

Table 1 (continued)

Year	Literature	Database description	Augmented data	Features	Classifier	Accuracy	Limitation
							database and MFCC features which is 96 sample.
2016	Lim et al. [23]		No	short-time Fourier transform.	time distributed CNN	Precision recall and f1 score 88.01%,86.86%86.6-5% respectively	Authors claim that model study is required in case of multimodel database
2016	Sidorov et al. [34]		No	Power, mean, root mean square, jitter, shimmer, 12 MFCCs and 5 formants. And pitch, intensity and harmonicity	GMM	72.00%	Authors Claim that Hidden Markov model (HMM) and Neuro-Fuzzy System is a robust algorithm; this type of classification can be implemented.
2017	Khan et al. [5]			Pitch and MFCC	Naïve Bayes classifier	72.34%	Gender dependency
2017	Yang et al. [48]		No	F0, energy +1st and 2nd derivatives, four formants, MFCC, speaking rate	SVM	59.66%	the applications of the pQPSO, massive number of computational complexity and huge memory consumption and are the limitations of the proposed method and authors mention that deep extreme learning machine and elliptical basis function networks could be implemented
2017	Badshah et al. [6].			Spectrograms analysis	CNN	84.30%	The satisfactory result show Except fear Emotion
2017	WeiBsskirchen N vet al. [46]		Spectrogram created and mirroring with the	Spectrogram analysis	CNN	52%	original input has been recreated, and that is the reason input data effected with noise,

Table 1 (continued)

Year	Literature	Database description	Augmented data	Features	Classifier	Accuracy	Limitation
			relevant part				
2017	Wang et al. [45]		No	Fourier Parameter (FP) inspired by Fourier Analysis and MFCC features	SVM classifier	average accuracy of 73.3%	The authors extracted new features from only 6 out of 7 emotion classes by eliminating the "disgust" class.
2018	Zhao et al. [50]		No	utilized log-mel spectrograms as input data to their	2-D CNN LSTM network.	speaker-independent 95.89% (Best model till now)	In speaker-independent model accuracy showing lowest accuracy in except Emo-DB database
2018	Demircan and Kahramanli [13]		No	Mel Frequency Cepstral coefficients (MFCCs) The framework then implemented the feature reduction utilizing the fuzzy C-means clustering	artificial neural network (ANN), support vector machines (SVM) and k-nearest neighbours (kNN)	accuracies on the test set were 90%, 92.86%, and 92.86%, respectively	520 sample has been taken instead of 535 from Emo-Db
2019	Zhao et al. [51]		No	Long -mel Spectrograms	2D CNN LSTM Classifier	95.33% speaker dependent & 95.89% Speaker indepent	No Limitation Found
2020	issh. D et.al [20]		Time stretching and noise addition	MFCC, Melseptrogram, Contrast, Chroma, Torrentz	CNN	61%	The best model works for specific database Specficafic Model
2020	Pereira et al. [25]		No	Prosodic and low-level features	semi-supervised neural network	72.00%	High training Time and to increase accuracy metaheuristic and the dynamic classifier can be implemented

Table 2 Review of works on RAVDESS [33] database

Year	Literature	Database Description	Augmented data	Features	Classifier	Accuracy	Limitation
2016	Shegokar and Sircar [43]	8 different emotions like sad, happy, angry, calm, fearful, surprised, neutral and disgust, Total no of	No	Continuous Wavelet Transform (CWT)	Quadratic SVM with a 5-fold cross-validation technique	the best result with 60.1%	Male Speech sample and use selected features with different types of SVM
2016	Zhang et al. [50].	utterances 2452 (Song database 1012 and Speech Database 1440) Each Actor has 60 utterness (12	Dataset by indicating that classifiers using the song-to-speech relationship can achieve higher accuracy	group multi-task feature	selection (GMTFS) model.	the accuracy of 57.14%	Out of 8 emotions author only utilize 4 class: angry, happy, neutral and sad.
2017	Zeng et al. [49]	Actors & 12 Actresses)	No	spectrograms generated from the songs and the speech utterances	multi-task gated Residual Networks (GResNets)	65.97%	Authors claim that this model is task-specific
2017	Popova et al. [38]		No	Mel spectrograms obtained from the speech	using Convolutional Neural Network VGG-16 as a classifier	The authors obtained the accuracy of 71%	Melspectrogram co-efficient is not included in this work

Table 3 Review of works on Emo-vo [11] database

Year	Literature	Database description	Augmented Data	Features	Classifier	Accuracy	Limitation
2019	Latif. et al. [31]	This dataset is the first Italian language emotional corpus and contains 588 recordings. There are 6 actors whose scripts of 14 different sentences in 7 other emotional states, including disgust, fear, anger, joy, surprise, sadness and neutral. Two separate groups of listeners evaluated the recordings to validate the performance of emotional actors. All the recordings in this corpus were made with equipment in the Fondazione Ugo Bordoni laboratories.	Yes	Latent Codes+eGeMAPS	SVM	61.80%	No Limitation reported yet
2020	Haider, F.et al. [22]		No	emobase and eGeMAPs features	SVM	80%	Only consider two emotions instead of seven emotion

Table 4 Different corpus

Sl. No	Database	Languages	Emotions	Size
1	Emo-DB	German	anger, neutral, sadness, fear/anxiety, happiness, disgust, and boredom	535 utterances
2	Emo-Vo	Italian	disgust, fear, anger, joy, surprise, sadness and neutral	588 utterances
3	RAVDESS	English	happy, sad, calm, angry, fearful, surprised, neutral and disgust	Total no of utterances 2452 (Song database 1012 and Speech Database 1440) Each Actor has 60 utterness (12 Actors & 12 Actresses)

number 18 does not have a song database and disgust, neutral and surprised emotions are not included in the song database—the total sample of 2452. (Song database 1012 and Speech Database, 1440). Each actor has 60 utterness with eight different emotions: sad, happy, angry, calm, fearful, surprised, neutral and disgust. Two various emotional intensity statements except for neutral emotion in neutral emotion, no vigorous intensity is present.

3.2 EMODB

The second database which we used here is Emo-DB [10]. Another name of EmoDb is Berlin Database. It has 535 utterness in German languages, and it has a seven-emotion classification like fear, sadness, anger, neutral, happiness, disgust, and boredom, with ten different speakers with ten additional texts. Each audio file consists of 16KHz and 16 bits.

3.3 EMOVO

The third database is Emo-Vo [11] which is an Italian Speech Database. This database consists of 6 actors, which is 3 male, 3 female and 14 sentences. In Italian languages database, it has six emotions like neutral, disgust, fear, anger, joy, surprise, sadness including neutral sentiment having 14 sentences categories into four types (non-sense sentences, short and long sentences, questions). A total number of utterances is 588. Each audio file sampling frequency is 48Khz and 16bit stereo.

4 Methods

4.1 Feature extraction

All audio features can be categorized into a few segments like time-domain features, frequency domain features, perceptual features, windowing features. Example of time-domain features includes RMSE of the waveform [19]. In frequency domain example amplitude of individual frequency, perceptual frequency example MFCC and hamming distance of window are examples of windowing features. Features extraction play a significant role to detect proper classification of any machine learning algorithm. Useful features selection could lead to better analysis and train the model. Here we are using 5 spectrogram related features [35].

MFCC features MFCC is the most used spectral features in speech emotion recognition [14–17, 41, 42]. It is beneficial features to collect benefits from raw audio. It also works very effectively in noise. In this study, 39 characteristics of MFCC has been incorporated. The human hearing system is not linearly; therefore, MFCC works the Mel scale [41]. Hearing in humans is not on a linear scale, and therefore MFCC uses the Mel scale. Figure 1. represents the diagrammatic overview of MFCC pipeline.

Mel-spectrogram Mel-spectrogram very significant features in sound analysis. Spectrogram input mapped directly onto Mel basis function. This Mel–spectrogram representation is increased auditory clarity of the system [17].

Chroma features Chroma features or chromatogram is very closely related to the 12 different pitch classes; these features are potent tools for analyzing music. Chroma features are very authentic to capture harmonic and melodic characteristics of music and instrument [42].

Contrast It is estimated mean energy of a sound, and high contrast value is apparent and low contrast value corresponds to broadband noise [16].

Tonnetz Tonnetz features are similarly working like chromatogram. It is also related to harmony and pitch class. [42]

4.2 Convolution neural network baseline model

CNN is a compelling non-linear deep learning classification model. CNN consists of three basic win units other than the input layer, such as the Convolution layer, Pooling layer, and fully connected layer. A general algorithm for CNN is presented in Fig. 2.

- Convolution layer: The convolution layer uses a different filter to change input value data. Its' hyper-parameter consists of the filter size F and stride S. The output O is named feature map or activation map. The convolution operation of ξ *and* ϑ is denoted with the operator $*$, defined as an integral product of two functions after one is reversed and shifted.

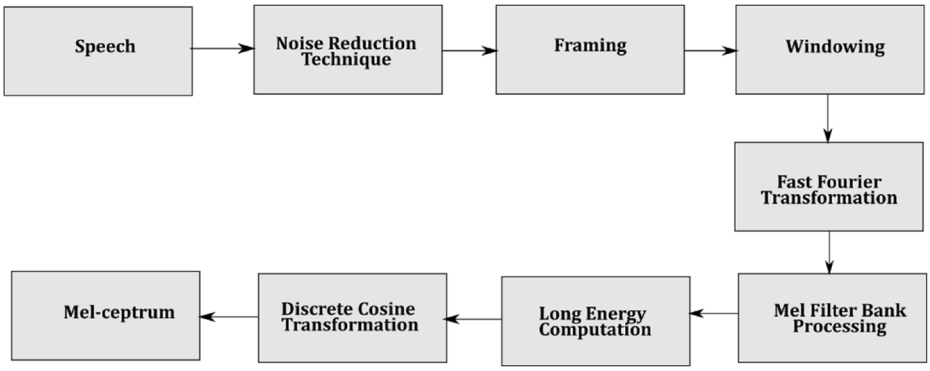


Fig. 1 MFCC pipeline

Algorithm: CNN (Convolution Neural Network)

Convolution Layer: $\phi^{\ell} = \xi^{\ell-1} * \vartheta^{\ell}$ *for* ξ *and* $\vartheta: [0,\infty) \rightarrow \mathbb{R}$

Pooling Layer: $\xi_{xy}^{\ell} = max_{ij}\left(\xi_{(x+i)(y+j)}^{\ell-1}\right) \forall\ 0 \leq i\ \&j \leq n$

Fully Connected Layer: $\phi^{\ell} = \xi^{\ell-1}\vartheta^{\ell}$

ReLu: $ReLu(z_i)=max(0, z_i)$

SoftMax Layer: $softmax(Zi) = e^{zi} / \sum_j e^{zj}$

Final Prediction: : **FP**

Fig. 2 CNN algorithm

$$\phi^{\ell} = \xi^{\ell-1} * \vartheta^{\ell} \text{ for } \xi \textit{ and } \vartheta : [0,\infty) \rightarrow \mathbb{R}$$

- Pooling layer: This Layer reduces the spatial size of the previous block.

$$\xi_{xy}^{\ell} = max_{ij}\left(\xi_{(x+i)(y+j)}^{\ell-1}\right) \forall\ 0 \leq i \& j \leq n$$

- Fully connected Layer: This layer is nothing but a feed-forward layer.

$$\phi^{\ell} = \xi^{\ell-1}\vartheta^{\ell}$$

- ReLu: This is an activation function which represents a positive number as 1 and otherwise 0.

$$\text{ReLu}(z_i) = \max(0, z_i)$$

- Softmax layer: This layer uses for the classifier; this is also working for multilayer classification.

$$softmax(Zi) = e^{zi} / \sum_j e^{zj}$$

Convolutional networks are trainable, multistage architectures. It comprises multiple convolutional layers and a restricted kind of totally connected hidden layers and pooling layers during a standard multilayer neural network. CNN has a set of totally connected hidden layers. In every remote unit, restricted information is processed in all the inputs, known as the receptive field. After that convolutional kernel (filter) is constructed from the weights of such hidden units and enforced to the total input area and made feature map. There will be many filters (feature maps) in a very convolutional layer.

The Baseline Model is in the proposed framework. One dimensional convolution layer combines dropout, normalized and activation Layer—the first Layer of Convolution Layer 193 × 1 number of the array as input full of hand-pick features from raw audio files. The initial layer comprising 512 filters and kernel size 5 and stride one and batch normalization is applied and Rectified Linear Units Layer (ReLu). The dropout rate has been implemented 0.1, is used. 193 × 1 feature array involving kernel 5 which creates a new array size 189.and max-pooling is applied in the max-pooling (8) then new array size is 24 has been formed, then along with 512 convolution layer. Next batch normalization is implemented feeding its output to the max-pooling layer with a window size four next three convolution layer 512 and 128,128 respectively filter size of 5 are

located. The next Convolution Layer, followed by the ReLu activator layer and drop out layer 0.1, the final Layer of convolutional Layer with the same parameters followed by the flattening layer, is received by fully connected layer. With eight and seven as per respective predictive class after that batch normalized and softmax activation is applied. In this baseline model, we have used RMSProp optimizer with the learning rate 0.00005. We have chosen our CNN layer design and all the hyperparameter value experiment-based. We have experimented with layer and different optimizer technique in our proposed model (only Original Multilingual Database). As an example, we have chosen Adam optimizer, Stochastic Gradient Descent (SGD) and RmsProp, where we have observed 67.04%, 53.14% and 71.99% respectively. [3]

4.3 Data augmentation

Several emotional databases have been used in our analysis. However, the different databases have different sizes, like Emo-Db (German Database), with 535 corpora. Emo-Vo (Italian Database) has 588 audio files, RAVDESS Database has speech and song database, size of the corpus is 1440 and 1012 respectively. This paper has shown that emotion detection from different languages, so we have chosen only speech emotion database from RAVDESS. We applied the Convolution Neural Network model on the aforementioned emotional database. However, as we have selected 193 features, so huge data is required to get proper training.

Furthermore, no of data is less in the original database as we mentioned in the previous section. So, we augmented emotional database based on this requirement, and there are different types of augmented technique, like noise addition, time-shifting, pitch shifting, time-stretching [36]. We have used in this paper only time stretching and pitch shifting [39, 40]. Time-stretching is changing the sound speed, with some parameter value like time-stretching rate is 0.81 and 1.07 and pitch shifting −3.5 and 1 respectively. We employed librosa library. In Fig. 3, data augmentation technique is shown.

5 Proposed model and experiment

As the aim is to perform Multilingual Audio Emotion Detection from a Deep Analysis, the CNN (Convolution Neural Network) [1] model has been applied. CNN models can automatically extract features from the input, which play an essential role in this in-deep learning approach. Many researchers have used a different type of CNN which is we mention in the review.

The proposed one-dimensional convolution model consists of the convolution layer, dropout layer, batch normalization, and activation layer. It directly takes all the audio files and stores into an array then fine-tuning experimentally useful features. Mixing of five features give us the best results experimentally we have seen. Many researchers use different features, but this combination and our model design show the best work. Which we show in the literature review.

5.1 MDLM (Multiligual deep learningmodel)

Our object to detect multilingual emotion detection from the raw audio file. Firstly, we have done augmented of our multilingual database using time stretching and pitch shifting methods. We have used hand-picked features, with a label. Selection of hand-picked features chosen by experimentally. Using MFCC, Chroma, Mel-spectrogram, Tonnetz and Contrast Total 193 features has been selected, Then features set with class label chosen as an input.

In our proposed multilingual database, common emotion has been created. The Total number of the original audio file in the multilingual database is 1660 based on 5 emotions and augmented database 7785 number of audio files. The initial multilingual database contains 1660 number of audio. Which splits into two-part test and train set values are 332 and 1328 respectively. After data augmentation on the original multilingual database, it generates 7785 audio files based on time-stretching and pitch-shifting methods.

Moreover, this augmented data use for training but test on the original dataset, which is 332 for this multilingual database. All the dataset and 193 features vector input in CNN and augmented training data size is 6228number of audio files. This training augmented data breaks into two parts, and every fold changes training and testing partition. Training and test for each fold have always changed, so there is no bias for selected training and testing data. In our model, we have chosen the value of $k = 5$. Finally, we represented all K -fold value for train Val each made a summation and similarly done for test value. All the training and testing are done in a one-dimensional convolution model consisting of a convolution layer, dropout layer, batch normalization, and activation layer.

The first Layer of CNN has taken an input of 193×1 array containing full of features. The initial layer we have taken 256 filters kernel size 5. We have use activation function Relu and dropout 0.1 and max-pooling considered as 8, next layer believes of 128 kernel size same activation function Relu and dropout 0.1and repeat this three times.

Moreover, the finally same parameter flattering layer with a fully connected layer the Last Layer is a dense layer that is eight over here with 5 predictive classes. 193 Figure 3 represents the architecture of our proposed model. In Fig. 4, the topology of CNN is depicted. We have also tested our model in RAVDESS, EmoDb, Emo-Vo Database and analysis it.

Based on experiment changing of different hyperparameter value, we have design our CNN model. Initially, we have chosen a baseline model and selecting a different parameter. Optimizer selection plays a significant role in CNN performance. We have compared our Multilingual Emotion database with Baseline model between Adam Optimizer, Stochastic Gradient Descent(SGD and RMSProp and we observed 67.04%, 53.14% and 71.99% respectively and we have used RMSProp as an optimizer in our model [3].

5.2 Proposed model algorithm

In this section, we have discussed the proposed model algorithm based on the Convolution Neural Network model. In the first part, we describe our CNN model and the next part explain the K-fold multilingual deep learning model (MDLM). Both standard and augmented data are

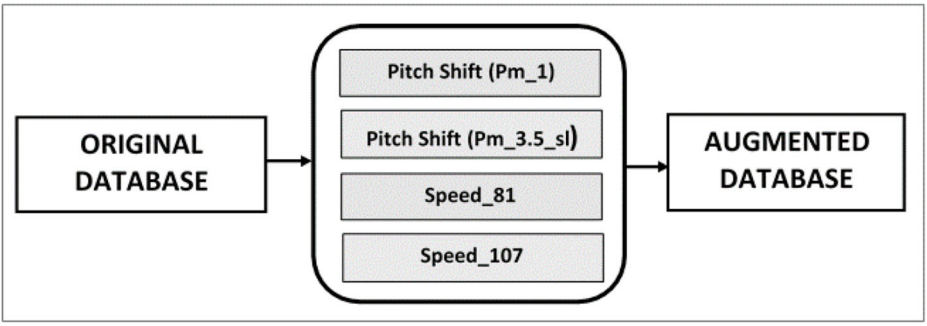


Fig. 3 Data augmentation technique

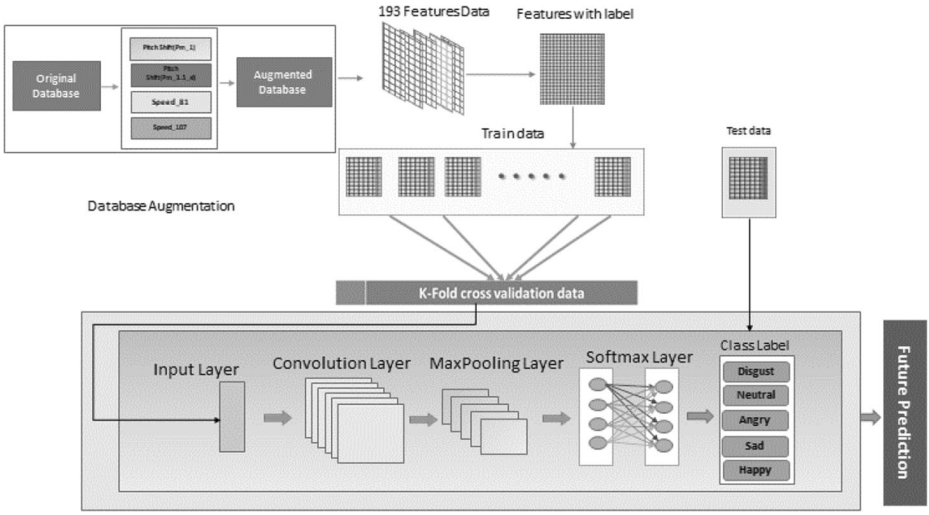


Fig. 4 The proposed model

considered and partitioned into training and testing set. We go through train data into our model with k-fold cross-validation and if it is working in expanded data, then train with original test data at the last of final prediction.

Our object to detect multilingual emotion detection from the raw audio file we also use some hand-picked features with deep features from CNN an input with label selection of this fusion features is chosen by experimentally,193 features have been selected, Then features set with a class label selected as an input Fig. 5.

$$D = \{(Xi, Yi)\}_{i=1}^{n} \text{where } Xk = (xk1, xk2, xk3, \ldots\ldots, xkd)\ Y = \{yi \mid yi \in (0\ to5)\}$$

$$\textbf{Output:}\quad \widehat{Y}\textbf{ : Multilingual Emotion Detection } \widehat{Y}\textbf{ : } (\mathbf{0}, \mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4})$$

Data augmentation is done in a different emotional database like RAVDESS, EmoDb, EmoVo Database based on pitch shift and time-stretching separately. In our proposed multilingual database, common emotion has been created the Total number of the original audio file in the multilingual database is 1660 based on 5 emotions and augmented database 7785 number of audio files.

Original Train and Test Data: Split D_{ori} into training and test set:$D_{ori} = D_{train}\ U\ D_{test}$

$$D_{oritrain} = (X_{oritrain}, Y_{oritrain}), D_{oritest} = (X_{oritest}, Y_{oritest})$$

The original multilingual database contains 1660 number of audio. Dori represents D_{ori}, and D_{ori} splits into two part $D_{train,}$ and D_{test} and values are these two 1328 and 332 respectively.

Augmented Train and Test Data: Split D into training and test set:$D_{Aug} = D_{augtrain}\ U\ D_{augtest}$

$$D_{augtrain} = (X_{augtrain}, Y_{augtrain}), D_{augtest} = (X_{augtest}, Y_{augtest})$$

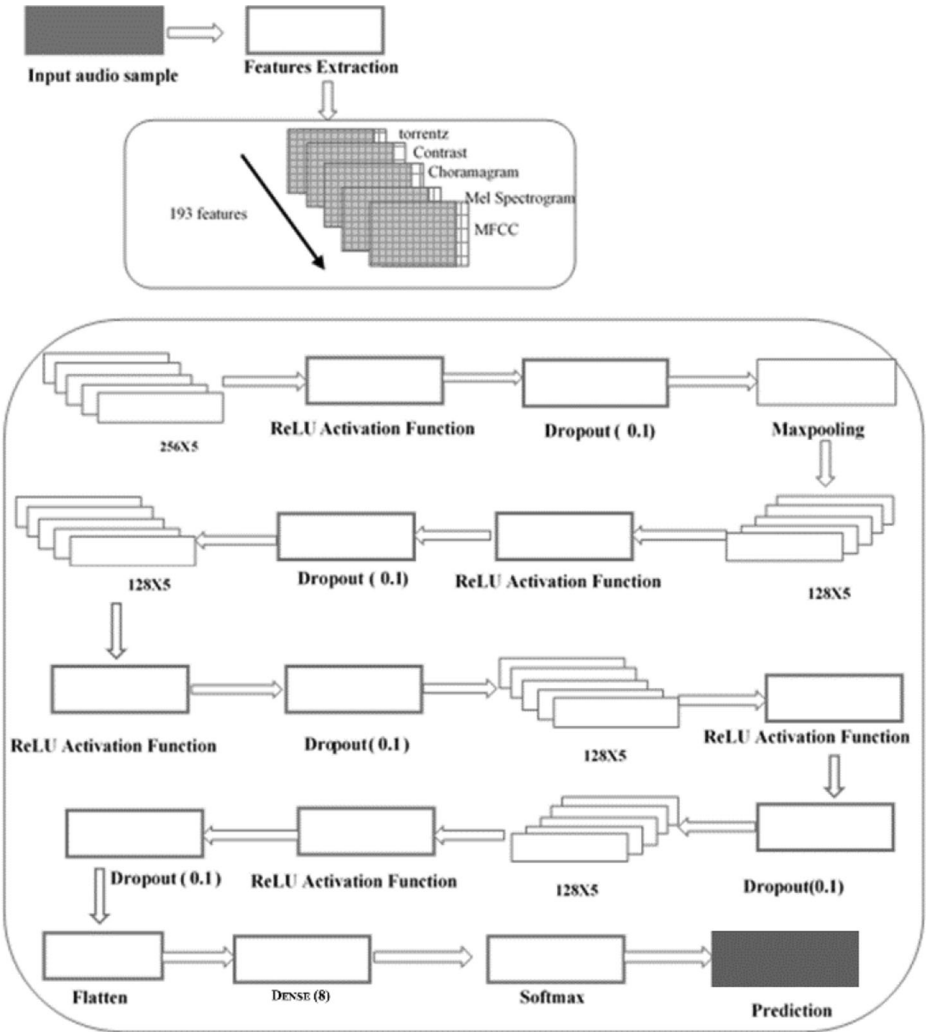


Fig. 5 Model topology of CNN

Train on Augmented Data with Original and Test on Real Data: $D_{augO}=D_{Aug}$ U D_{ori}And $X_{oritest}$

D_{augO} is a combination of Augment multilingual database ($D_{aug),}$ and Original Multilingual database ($D_{ori)}$ Total number of $_{the}$ audio file in Augmented and Original is 7785 $X_{oritest}$ is represented as 332.

All the dataset and 193 features vector input in CNN after D_{augO} as an input D_{augO} breaks into two-part $D_{augtrain}$ and $D_{augtest}$ 6228 number of audio files is in $D_{augtrain}$ and test with an original dataset which is 332 number of audio files. $D_{augtrain}$ breaks train and as well as test part

and every fold its changes train and test partition of $D_{augtrain}$ (Fig. 6)

Algorithm: CNN (Convolution Neural Network)

Convolution Layer: $\phi^{\ell} = \xi^{\ell-1} * \vartheta^{\ell}$ *for* ξ *and* $\vartheta: [0, \infty) \rightarrow \mathbb{R}$

Pooling Layer: $\xi_{xy}^{\ell} = max_{ij}\left(\xi_{(x+i)(y+j)}^{\ell-1}\right) \forall \ 0 \leq i \ \& j \leq n$

Fully Connected Layer: $\phi^{\ell} = \xi^{\ell-1}\vartheta^{\ell}$

ReLu: $ReLu(z_i) = max(0, z_i)$

SoftMax Layer: $softmax(Zi) = e^{zi} / \sum_j e^{zj}$

Final Prediction: : **FP**

Algorithm: K-Fold Multilingual Deep Learning Model **(CNN): MDLM**

Input: $D = \{(X_i, Y_i)\}_{i=1}^{n}$

where $X_k = (x_{k1}, x_{k2}, x_{k3}, \ldots\ldots, x_{kd})$ $Y = \{ y_i \mid y_i \in (0 \text{ to } 5)\}$

Output: $\hat{Y}$: Multi Lingual Emotion Detection

$\hat{Y}$: (0,1,2,3,4)

Original Train and Test Data: Split D_{ori} into training and test set: $D_{ori} = D_{train}\ U\ D_{test}$

$D_{oritrain} = (X_{oritrain}, Y_{oritrain})$, $D_{oritest} = (X_{oritest}, Y_{oritest})$

Augmented Train and Test Data: Split D into training and test set: $D_{Aug} = D_{augtrain}\ U\ D_{augtest}$

$D_{augtrain} = (X_{augtrain}, Y_{augtrain})$, $D_{augtest} = (X_{augtest}, Y_{augtest})$

Train on Augmented Data with Original and Test on Real Data: $D_{augO} = D_{Aug} \cup D_{ori}$

And $X_{oritest}$

MDLM $(D_{augO}, X_{oritest})$***:*** *K-Fold Multilingual Deep Learning Model*

Split $D_{augtrain}$ into K subsets/fold: $D_{augtrain} = \bigcup_{i=1}^{k} D_{augtrain}^{i}$

For model $CNN_i(D_{augtrain}, X_{augtest})$***:***

For each $j = 1, 2 \ldots, k$: (*For each fold*)

Training data:

$X_{train} = CNN_{j \neq i, i=1}^{k} X_{augtrain}^{i}$ $\quad Y_{train} = CNN_{j \neq i, i=1}^{k} Y_{augtrain}^{i}$

Test data:

$X_{val} = X_{augtest}^{j}$ $\quad Y_{val} = Y_{augtest}^{j}$

$\hat{Y}_{val}^{i,j} = \text{CNN}_i(X_{train}, Y_{train}, X_{val})$

$\hat{Y}_{test}^{i,j} = \text{CNN}_i(X_{train}, Y_{train}, X_{test})$

$\hat{Y}_{val}^{i} = \frac{1}{k}\sum_{i=1}^{k} \hat{Y}_{val}^{i,j}$

$\hat{Y}_{test}^{i} = \frac{1}{k}\sum_{i=1}^{k} \hat{Y}_{test}^{i,j}$

Final prediction: $\hat{Y}_{test} = FP(X_{train}, Y_{train}, X_{oritest})$

Fig. 6 K-Fold multilingual deep learning model

Algorithm: K-Fold Multilingual Deep Learning Model **(CNN): MDLM**

Input: $D = \{(X_i, Y_i)\}_{i=1}^{n}$

where $X_k = (x_{k1}, x_{k2}, x_{k3}, \ldots\ldots, x_{kd})$ $Y = \{ y_i \mid y_i \in (0 \text{ to } 5)\}$

Output: $\hat{Y}$: Multi Lingual Emotion Detection

$\hat{Y}$: (0,1,2,3,4)

Original Train and Test Data: Split D_{ori} into training and test set: $D_{ori} = D_{train}\ U\ D_{test}$

$D_{oritrain} = (X_{oritrain}, Y_{oritrain})$, $D_{oritest} = (X_{oritest}, Y_{oritest})$

Augmented Train and Test Data: Split D into training and test set: $D_{Aug} = D_{augtrain}\ U\ D_{augtest}$

$D_{augtrain} = (X_{augtrain}, Y_{augtrain})$, $D_{augtest} = (X_{augtest}, Y_{augtest})$

Train on Augmented Data with Original and Test on Real Data: $D_{augO} = D_{Aug} \cup D_{ori}$

And $X_{oritest}$

MDLM $(D_{augO}, X_{oritest})$: *K-Fold Multilingual Deep Learning Model*

Split $D_{augtrain}$ into K subsets/fold: $D_{augtrain} = \bigcup_{i=1}^{k} D_{augtrain}^{i}$

For model $CNN_i(D_{augtrain}, X_{augtest})$:

For each $j = 1, 2\ldots., k$: (*For each fold*)

Training data:

$X_{train} = CNN_{j \neq i, i=1}^{k} X_{augtrain}^{i}$ $Y_{train} = CNN_{j \neq i, i=1}^{k} Y_{augtrain}^{i}$

Test data:

$X_{val} = X_{augtest}^{j}$ $Y_{val} = Y_{augtest}^{j}$

$\hat{Y}_{val}^{i,j} = \mathrm{CNN_i}(X_{train}, Y_{train}, X_{val})$

$\hat{Y}_{test}^{i,j} = \mathrm{CNN_i}(X_{train}, Y_{train}, X_{test})$

$\hat{Y}_{val}^{i} = \frac{1}{k} \sum_{i=1}^{k} \hat{Y}_{val}^{i,j}$

$\hat{Y}_{test}^{i} = \frac{1}{k} \sum_{i=1}^{k} \hat{Y}_{test}^{i,j}$

Final prediction: $\hat{Y}_{test} = FP(X_{train}, Y_{train}, X_{oritest})$

5.3 RAVDESS model

RAVDESS database has two partition Song database and Speech database. In the first experiment, we have worked on Song and Speech database combined. Frist implemented model describe in section 4; eight different emotions is in RAVDESS database. We first applied the traditional machine learning approach like decision tree, SVM, random forest using Gini index, and entropy. We also applied Convolution Neural Network and data partition randomly speaker independently test data and train data 80% and 20% respectively. After 500

epochs, the emotion classification performance of our model is 91.93% with all 193 features. This paper has concentrated on the multilingual speech database—our experiment in only speech database with 1440 audio files. Our model produces 70.46% test and train data (80% and 20%) respectively with 300 epochs, we also applied k-fold cross validation where k value is five and results in 77.60% with 300 epochs. After this experiment to get better results, we create an augmented RAVDESS database where total 5760 number of audio file is generated by time starching and pitch-shifting method. Then we split with original as 80% and 20% as train and test data respectively and get 1152 train data and 288 test data from the initial database. Finally, 6912 audio files are selected as an input with 193 features in our proposed model and test with 288 audio files from the original database, and we got 96.53% of accuracy. A confusion matrix is given below, where it identifies all the emotions. Figures 7 and 8

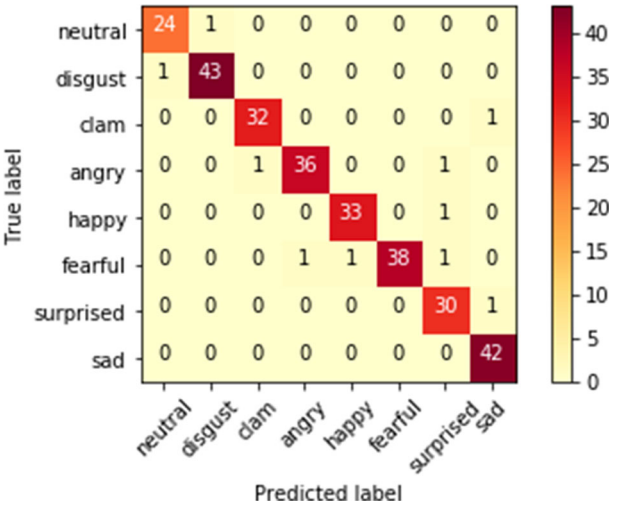


Fig. 7 Confusion matrix of k-fold cross-validation on augmented RAVDESS database

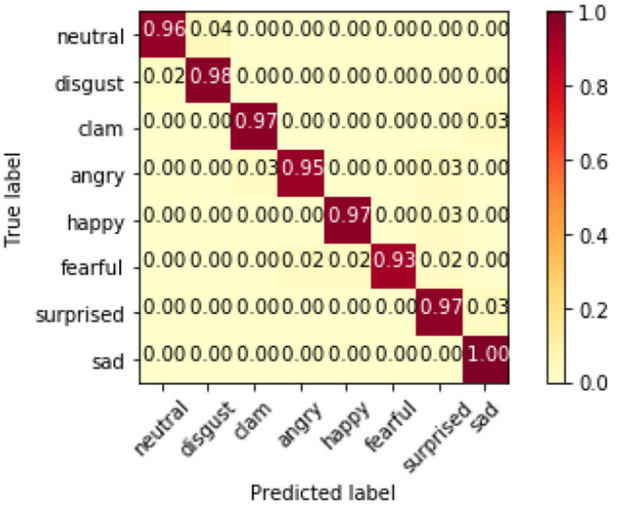


Fig. 8 Normalized confusion matrix of k-fold cross-validation on augmented RAVDESS database

represents normalized and unnormalized Confusion matrix of k-fold cross-validation on Augmented RAVDESS Database, and Table 5 represents the Classification Report of RAVDESS Augmented K-fold cross-validation.

5.4 Emo-Vo model

Our model experiment has also gone through Emo-Vo (Italian Database), where a total number of audio file is 588. We have experimented with our CNN model on 588 audio files directly and train and test randomly partition 80% and 20% respectively. Moreover, get 73.47% accuracy, and then we generate some augmented audio through the time-stretching and pitch-shifting method and create a new augmented database size is 2821. Then through the model experiment, we got 90.87% through only random partitioning train and test data 80% and 20%. Finally, we use k-fold cross-validation on augmented data. The test has been done on original dataset no of the train set 2256 and test 118 utterance from the initial file database. We got the accuracy 96.11% Figs. 9 and 10 represented both normalized and without normalized

Table 5 Classification report of RAVDESS augmented K-fold cross-validation

	precision	recall	f1-score	Support
0	0.96	0.96	0.96	25
1	0.98	0.98	0.98	44
2	0.97	0.97	0.97	33
3	0.97	0.95	0.96	38
4	0.97	0.97	0.97	34
5	1.00	0.93	0.96	41
6	0.91	0.97	0.94	31
7	0.95	1.00	0.98	42
avg / total	0.97	0.97	0.97	288

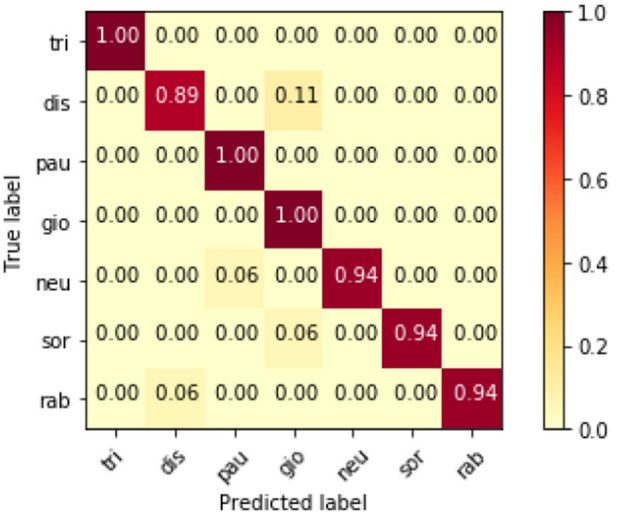


Fig. 9 Normalized confusion matrix Emo-Vo augmented data with k-fold cross-validation

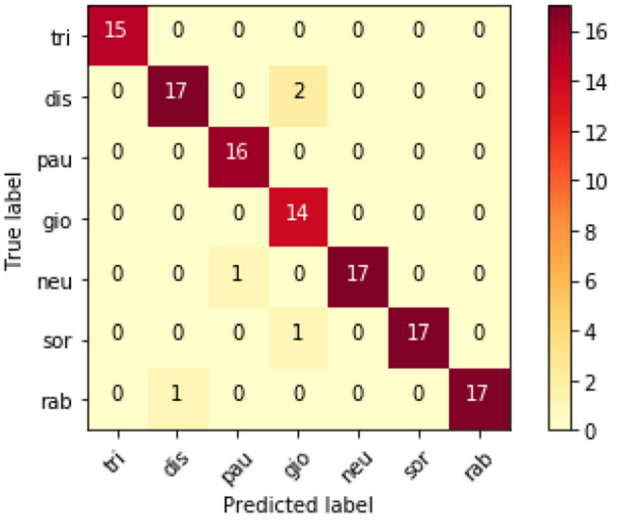


Fig. 10 Confusion matrix Emo-Vo augmented data with k-fold cross-validation

confusion matrix. Moreover, Table 6 illustrates a classification report of Emo-Vo Augmented K-Fold cross-validation. Here it firmly classifies all the emotions.

All the classification is based on speaker-independent classification technique.

5.5 Emo-Db model

Our model experiment has also gone through Emo-Db (German Database), where, the total number of audio file is 535. We have to go through our CNN model on 535 audio files directly. After that, train and test randomly partition 80% and 20% respectively with 70.37% accuracy. We generate some augmented audio through the time-stretching and pitch-shifting methods and create a new augmented database size of 2140. We have taken expanded and original audio file both in new augmented database. The size is 2568, through a model experiment, we got 92.76% through only random partitioning train and test data. Finally, we use k-fold cross-validation on augmented data and test has been done on actual dataset test 109 utterance from original file database. We got the accuracy of 97.12%. Here it firmly classifies all the emotion. Figures 11 and 12 represented Confusion Matrix Emo-Db Augmented data with k-fold cross-validation without normalized and normalized. All the classification is based

Table 6 Classification report of Emo-Vo augmented K-fold cross-validation

	precision	recall	f1-score	Support
0	1.00	1.00	1.00	15
2	0.94	0.89	0.92	19
3	0.94	1.00	0.97	16
4	0.82	1.00	0.90	14
5	1.00	0.94	0.97	18
6	1.00	0.94	0.97	18
7	1.00	0.94	0.97	18
avg / total	0.96	0.96	0.96	118

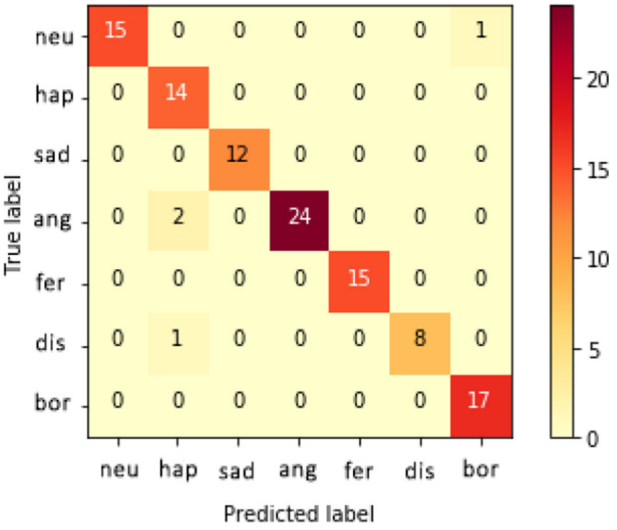


Fig. 11 Confusion matrix Emo-Db augmented data with k-fold cross-validation

on speaker-independent classification technique. Table 7 represents the Classification Report of Emo-Db Augmented K-fold cross-validation

5.6 Multilingual database

A total number of audio file mixed 2563. Emo-Vo-588, Emo-Db-535, RAVDESS database 1440 audio files. Five common emotion from different linguistic corpus like disgust, neutral, angry, sad and happy. We have to discard calm, fear and boredom. We have only chosen five common emotions which is present in different linguistic corpus.

After analysis, we got 72.59% in the original database and using K-fold cross-validation. We got 71.99% over 332 emotions from all databases; Table 8 represents the Classification

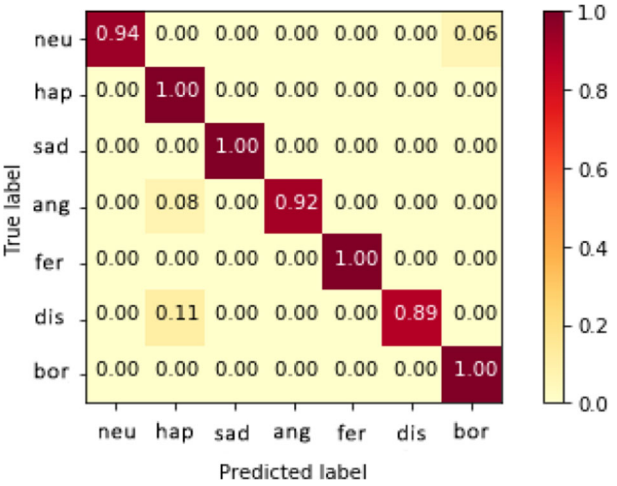


Fig. 12 Normalized confusion matrix Emo-Db augmented data with k-fold cross-validation

Table 7 Classification report of Emo-Db augmented K-fold cross-validation

	precision	recall	f1-score	Support
0	1.00	0.94	0.97	16
2	0.82	1.00	0.90	14
3	1.00	1.00	1.00	12
4	1.00	0.92	0.96	26
5	1.00	1.00	1.00	15
6	1.00	0.89	0.94	9
7	0.94	1.00	0.97	17
avg / total	0.97	0.96	0.96	109

Table 8 Classification report of all Original database with K-fold CNN

	precision	recall	f1-score	support
0	0.81	0.79	0.80	53
2	0.73	0.56	0.64	87
3	0.70	0.80	0.75	56
4	0.75	0.77	0.76	83
6	0.61	0.74	0.67	53
avg / total	0.72	0.72	0.72	332

report of all original databases on the multilingual database with K-fold CNN, Figs. 13 and 14 representing the same classification report, on Original Multilinguistic database with k-Fold CNN with normalized and without normalized Confusion Matrix respectively.

Then we did our research on augmented data, which is generated to overcome the overfitting problem. We have done two types of test. We have done train on expanded data with original data then after that test with initial data. As a result, we got an accuracy of 98.19% and test with expanded data in K-fold cross-validation we got accuracy 92.87% test on

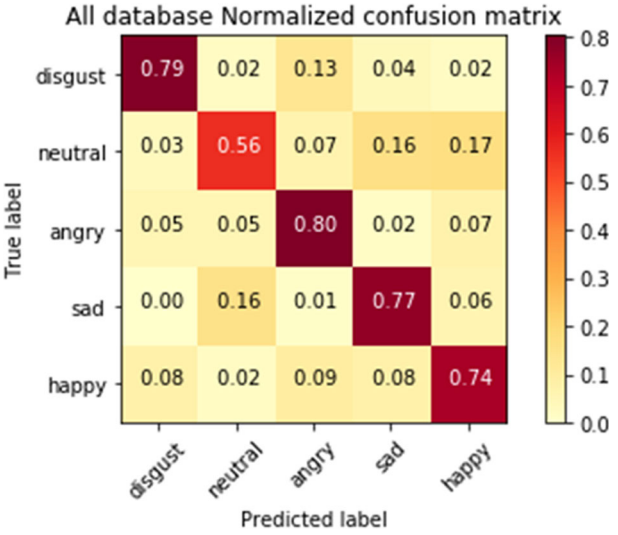


Fig. 13 Multilinguistic original database with k-Fold CNN normalized confusion matrix

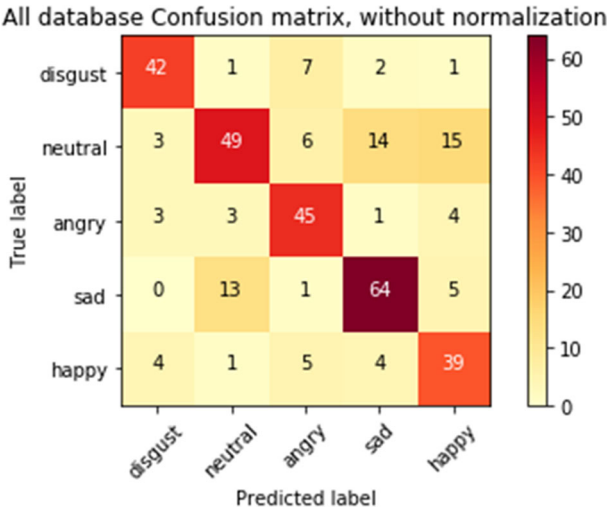


Fig. 14 Multilinguistic original database with k-Fold CNN without normalized confusion matrix

1557 number of audio files. The original database test gave us a 97.89% test on 332 audio files from the initial database. Table 9 represents the Classification Report K-fold CNN with Augmented data on the multilinguistic database, Figs. 15 and 16 represents Confusion Matrix with normalized and without normalization with K-fold CNN with Augmented data.

6 Results and analysis

6.1 Performance on different traditional machine learning algorithms

Initially, we have started analysis with a traditional machine learning algorithm [37], we have begun with Decision Tree (DT), K-nearest neighbour (KNN), Random Forest (RF), AdaBoost, Gradient Boosting etc. machine learning algorithm. We experimented with the traditional Machine Learning model in the Multilingustic database, building a decision tree model to split test database into 20% randomly. The maximum depth is default 0, and 1 or 42 and criterion has been chosen Gini Index, and secure accuracy 51.86% and similarly. We have continued with our experiment with another very robust algorithm SVM. In support vector machine(SVM) we used Gaussian kernel, RBF (radial base function) and applied cross-validation technique for splitting in training and testing set data (20%)and got accuracy 54.35%.In Random forest criterion has been chosen Gini index and maximum depth = 10,

Table 9 Classification report K-fold CNN with augmented data

	precision	recall	f1-score	support
0	0.98	1.00	0.99	53
2	0.97	0.98	0.97	87
3	1.00	0.98	0.99	56
4	0.99	0.99	0.99	83
6	0.96	0.94	0.95	53
avg / total	0.98	0.98	0.98	332

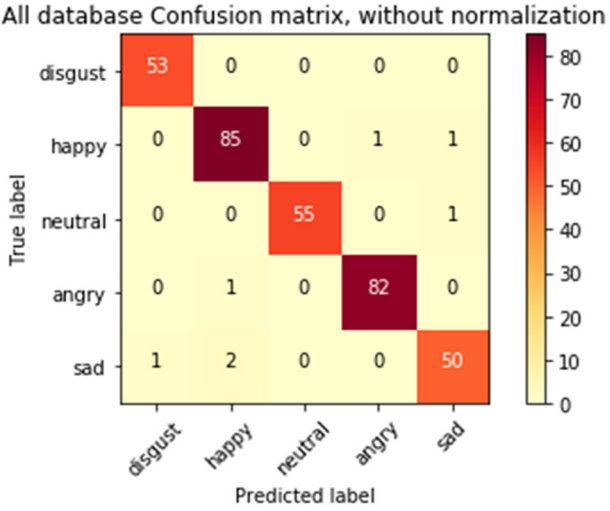


Fig. 15 Confusion matrix without normalization with K-fold CNN with augmented data

maximum leaf node =100 and accuracy 65.53%, In Gradient Boosting we have use estimator as 100, and we have select learning rate = 1, and we got accuracy 61.49%. In Adaboost Algorithm and K Nearest Neighbour, we got accuracy 41.92%, 59.62%, respectively. We also compare with our CNN model, where it reflects 71.99%. We also run the experiment in other three emotion database (RAVDESS, Emo-Vo and Emo-Db). Another observation found in Emo-Db database best accuracy found in Gradient Boosting accuracy 85.04%. In Emo-Vo database best accuracy found in KNN 74.94%, In all three other databases, no single traditional machine learning algorithm found best results. Hence, we move to deep learning model for better accuracy. Table no. 11 represent details analysis. Confusion matrix figure represents all the above mention Machine-learning algorithm results. and Table 10 represent

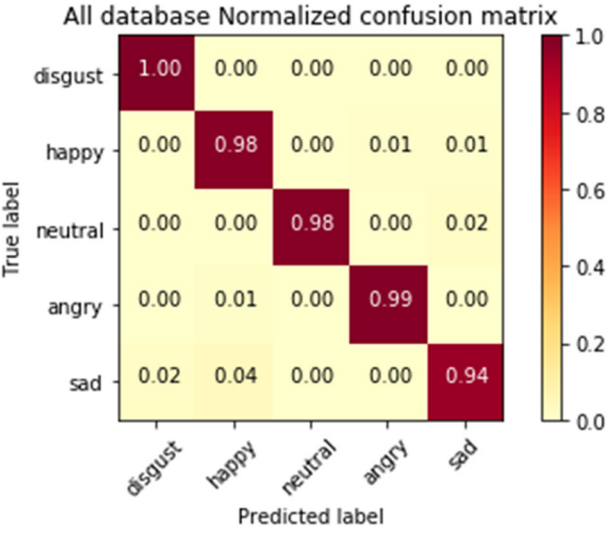


Fig. 16 Confusion matrix without normalization with K-fold CNN with augmented data

Table 10 Represent average classification details of Traditional Machine Learning algorithm in Multilingual Database

Database	Algorithm	Class	Precision	Recall	F1-score	Support
Multilingual Database	Decision Tree	5	0.54	0.52	0.52	322
	KNN		0.59	0.60	0.58	
	SVM		0.57	0.54	0.55	
	Random Forest		0.68	0.66	0.66	
	Adaboost		0.59	0.42	0.38	
	Gradient Boosting		0.64	0.61	0.60	

average classification details of Traditional Machine Learning algorithm in Multilingual Database.

6.2 Performance on CNN

Compare to other Machine Learning Algorithm CNN produce best results Table 11 represent the same. The baseline models of emotion classification use CNN, with different feature selection techniques and hand-pick features and deep features.[] We applied Data partition randomly to test data and train data 80% and 20% respectively. After 300 epochs, the emotion classification performance of our model is on Original RAVDESS database. (Only Speech database size is 1440 number of audio file) accuracy 70.46% and K-fold apply in Original database 77.60%.

Moreover, similarly, this is applied into RADVESS augmented database and got an accuracy of 91.30% and after k-fold applied, we got 96.53%. We also used Emo-Db and Emo-Vo and got accuracy in original and expanded k fold applied 71.03%,96.26%73.73%and 96.11% respectively. We also applied this to the multilingual database where Table 12

Table 11 Comparison chart of Traditional ML VS DL

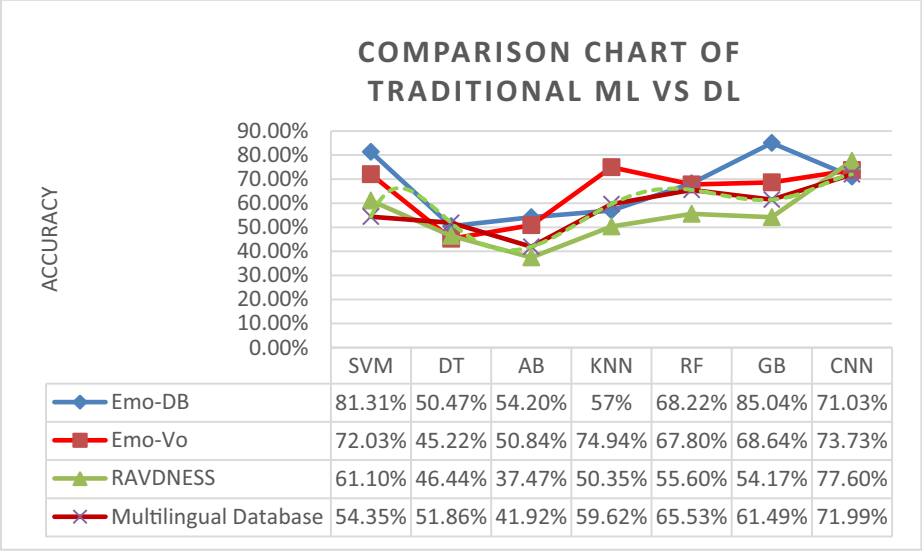
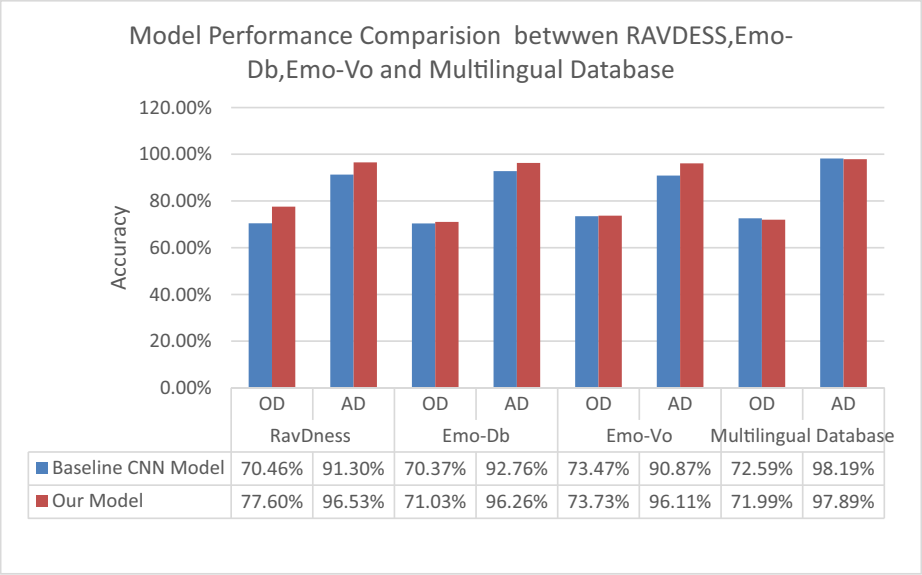


Table 12 Model performance comparision between RAVDESS, Emo-Db, Emo-Vo and multilingual database with K-fold and without k-fold



	RavDness		Emo-Db		Emo-Vo		Multilingual Database	
	OD	AD	OD	AD	OD	AD	OD	AD
Baseline CNN Model	70.46%	91.30%	70.37%	92.76%	73.47%	90.87%	72.59%	98.19%
Our Model	77.60%	96.53%	71.03%	96.26%	73.73%	96.11%	71.99%	97.89%

represents comparison with Original database (OD)and augmented database (AD). Then we have done this experiment on augmented data (Fig. 17).

Firstly, we experimented with the RAVDESS database, and the audio sample was 1440, and popova. et al. [38] got accuracy 71%. Still, they have chosen 1368 and did not mention the proper reason, and some existing approaches have been used.

Emo Db Database Result Analysis. With compare to other models in Emo-dB database our model, we got 96.26% accuracy (Fig. 18).

Emo-Vo database Result analysis: - In Emo-Vo database compare to others model our accuracy found 96.11% (Fig. 19).

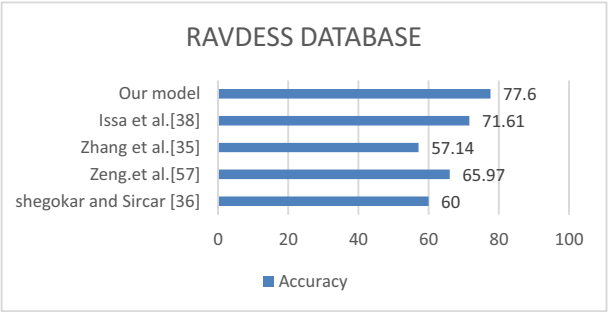


Fig. 17 Comparison of the different model on RAVDESS database

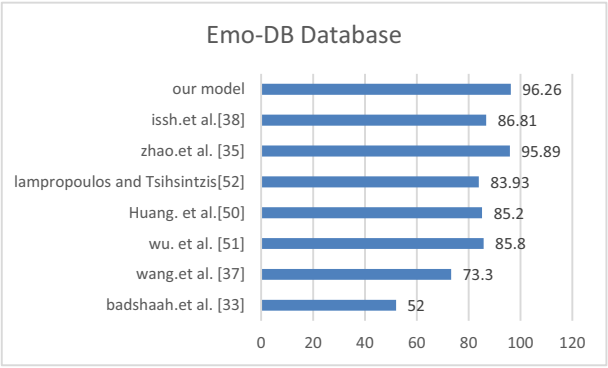


Fig. 18 Comparison of the different model on Emo-Db database

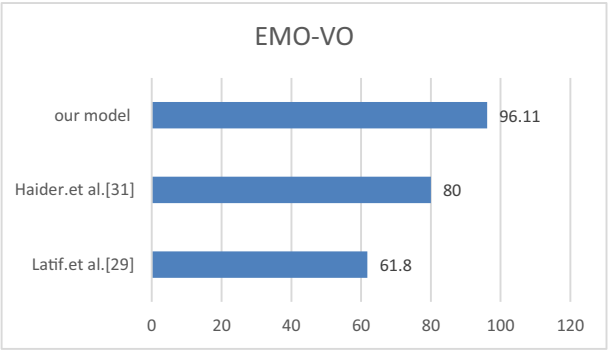


Fig. 19 Comparison of the different model on Emo-Vo database

7 Discussion & conclusion

There are two significant challenges in speech emotion detection processes, namely, feature extraction and classification. This experiment proposed a model, combining 1D CNN with five different features as input data. The investigation is conducted on three audio emotion databases, namely, RAVDESS, Emo-Db, Emo-Vo. The model is trained in RAVDESS, Emo-Db, and Emo-Vo with the original database. Avoid overfitting; an audio file is augmented applied the model into it. By achieving a perfect accuracy of all database and comparing it to Issa.D et al. [20] in the RAVDESS database model, we got better accuracy from them. The same is done for Emo-DB. However, we finally proposed a model where language-independent, speaker-independent, gender-independent emotion can be detected. Many models have presented a visual representation of the sound [18, 24, 34, 39, 41], but our model can directly work with the raw audio file. Another observation found that the baseline model performs slightly better than our model in the multilingual aspect. But our model consistently better than the baseline model for all other cases. The experiment was conducted with an Intel® Core ™ i7-7700HQ CPU at 2.80 GHz with 8 GB RAM.

The models have an accuracy of 90.86% in RAVDESS with all features and 90.45% with only MFCC on RAVDESS in CNN with audio and video (Extract audio from video) database

we have tested. The performance increase for when all features work together and only MFCC has a perfect accuracy maintained earlier.

More relevant researches can be explored on this topic, the different types of features can be applied, or an additional neural network layer can be applied to increase accuracy. Apart from this data, the comprehensive augmented technique could improve performance. In future, we are planning to select optimal features to use the best version without data augmentation.

References

1. Ahuja R, Jain D, Sachdeva D, Garg A, Rajput C (2019) Convolutional neural network based American sign language static hand gesture recognition. International Journal of Ambient Computing and Intelligence (IJACI) 10(3):60–73
2. Akçay MB, Oğuz K (2020) Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication 116:56–76
3. Ali MNY, Sarowar MG, Rahman ML, Chaki J, Dey N, Tavares JMR (2019) Adam deep learning with SOM for human sentiment classification. International Journal of Ambient Computing and Intelligence (IJACI) 10(3):92–116
4. Alsharif MH, Kelechi A, Yahya K, Chaudhry S (2020) Machine Learning Algorithms for Smart Data Analysis in Internet of Things Environment: Taxonomies and Research Trends. Symmetry 12(1):88. https://doi.org/10.3390/sym12010088
5. Atreyee K, Kumar RU (2017) Emotion recognition using prosodie and spectral features of speech and Naïve Bayes Classifier. In: 2017 international conference on wireless communications, signal processing and networking (WiSPNET). IEEE
6. Badshah AM, Ahmad J, Rahim N, Baik SW (2017) Speech emotion recognition from spectrograms with deep convolutional neural network. In: 2017 International Conference on Platform Technology and Service (PlatCon). IEEE, pp 1–5
7. Bellamkonda S, Gopalan NP (2020) An enhanced facial expression recognition model using local feature fusion of Gabor wavelets and local directionality patterns. International Journal of Ambient Computing and Intelligence (IJACI) 11(1):48–70
8. Benzebouchi NE, Azizi N, Ashour AS, Dey N, Simon Sherratt R (2019) Multi-modal classifier fusion with feature cooperation for glaucoma diagnosis. Journal of Experimental & Theoretical Artificial Intelligence 31:841–874. https://doi.org/10.1080/0952813X.2019.1653383
9. Bharati P, Pramanik A "Deep learning techniques—R-CNN to mask R-CNN: a survey". computational intelligence in pattern recognition. In: Advances in intelligent systems and computing, vol 999. Springer, Singapore. https://doi.org/10.1007/978-981-13-9042-5_56
10. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B (2005) A database of german emotional speech. In: Ninth European Conference on Speech Communication and Technology
11. Costantini M (2014) EMOVO Corpus: an Italian emotional speech database. In: Proceedings of the ninth international conference on language resources and evaluation (LREC'14). European Language Resources Association (ELRA), pp 3501–3504
12. Dangol R, Alsadoon A, Prasad PWC, Seher I, Alsadoon OH (2020) Speech emotion recognition UsingConvolutional neural network and long-short TermMemory. Multimed Tools Appl 79:1–18
13. Demircan S, Kahramanli H (2018) Application of fuzzy c-means clustering algorithm to spectral features for emotion classification from speech. Neural Comput Appl 29:59–66
14. Dey N, Ashour AS (2018) Challenges and future perspectives in speech-sources direction of arrival estimation and localization. In: Direction of arrival estimation and localization of multi-speech sources. Springer, Cham, pp 49–52
15. Dey, N., & Ashour, A. S. (2018). Direction of arrival estimation and localization of multi-speech sources. Springer International Publishing.
16. Dey N, Ashour AS (2018) Applipart is d examples and applications of localization and tracking problem of multiple speech sources. In: Direction of arrival estimation and localization of multi-speech sources. Springer, Cham, pp 35–48
17. Dey N, Ashour AS (2018) Sources localization and DOAE techniques of moving multiple sources. In: Direction of arrival estimation and localization of multi-speech sources. Springer, Cham, pp 23–34

18. Dey N, Das A, Chaudhuri SS (2012) Wavelet based normal and abnormal heart sound identification using spectrogram analysis. arXiv preprint arXiv:1209.1224
19. Dey N, Mishra G, Nandi B, Pal M, Das A, Chaudhuri SS (2012) Wavelet based watermarked normal and abnormal heart sound identification using spectrogram analysis. In: 2012 IEEE international conference on computational intelligence and computing research. IEEE, pp 1–7
20. Dias I, Fatih Demirci M, Yazici A (2020) Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control 59:101894
21. Enes Y, Hüseyin H, Cem B (2014) Automatic speech emotion recognition using auditory models with binary decision tree and SVM. In: 2014 22nd international conference on pattern recognition. IEEE
22. Haider F, Pollak S, Albert P, Luz S (2020) Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods. Computer Speech & Language 65:101119
23. Huang Z, Dong M, Mao Q, Zhan Y (2014) Speech emotion recognition using cnn. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp 801–804
24. Iker L, Eva N, Inmaculada H (2010) Feature analysis and evaluation for automatic emotion identification in speech. IEEE Trans Multimedia 12(6):490–501
25. Ingryd P, Diego S, Alexandre M, Pablo B (2018) Semi-supervised model for emotion recognition in speech. International conference on artificial neural networks. Springer, Cham
26. Jiang D-N, Lu L, Zhang H-J, Tao J-H, Cai LH (2002) Music type classification by spectral contrast feature. In: Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE international conference on, vol 1. IEEE, pp 113–116
27. Jouni P, Paavo A (2014) Multi-scale modulation filtering in automatic detection of emotions in telephone speech. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE
28. Kadiri SR, Gangamohan P, Gangashetty SV, Yegnanarayana B (2015) Analysis of excitation source features of speech for emotion recognition. In: Sixteenth annual conference of the international speech communication association
29. Kalita DJ, Singh VP, Kumar V (2020) A survey on SVM hyper-parameters optimization techniques. In: Lecture notes in networks and systems, vol 100. Springer, Singapore. https://doi.org/10.1007/978-981-15-2071-6_20
30. Lampropoulos AS, Tsihrintzis GA (2012) Evaluation of mpeg-7 descriptors for speech emotional recognition. In: 2012 Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP). IEEE, pp 98–101
31. Latif S, Qadir J, Bilal M (2019) Unsupervised adversarial domain adaptation for cross-lingual speech emotion recognition. In: 2019 8th international conference on affective computing and intelligent interaction (ACII), pp 732–737
32. Lim W, Jang D, Lee T (2016) Speech emotion recognition using convolutional and recurrent neural networks. In: Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific. IEEE, pp 1–4
33. 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:e0196391
34. Maxim S, Wolfgang M, Stanislavovich SE (2016) Speechbased emotion recognition and speaker identification: static vs. dynamic mode of speech representation. J Siberian federal Univ. Ser Math Phys 9(4):518–523
35. McFee B, Raffel C, Liang D, Ellis DP, McVicar M, Battenberg E, Nieto O (2015) Librosa: audio and music signal analysis in python. In: Proceedings of the 14th Python in Science Conference, pp 18–25
36. Mignot R, Peeters G (2019) An analysis of the effect of data augmentation methods: experiments for a musical genre classification task. Transactions of the International Society for Music Information Retrieval 2(1):97–110
37. Mirri S, Delnevo G, Roccetti M (2020) Is a COVID-19 Second Wave Possible in Emilia-Romagna (Italy)? Forecasting a Future Outbreak with Particulate Pollution and Machine Learning. Computation 8(3):74
38. Popova AS, Rassadin AG, Ponomarenko AA (2017) Emotion recognition in sound. In: International Conference on Neuroinformatics. Springer, pp 117–124
39. Röbel A (2003) Transient detection and preservation in the phase vocoder. In: International computer music conference (ICMC), pp 247–250
40. Röbel A, Rodet X (2005) Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation. In: International conference on digital audio effects, pp 30–35
41. Sen S, Dutta A, Dey N (2019) Audio indexing. Audio processing and speech recognition:1–11
42. Sen S, Dutta A, Dey N (2019) Speech processing and recognition system. In: Audio processing and speech recognition. Springer, Singapore, pp 13–43

43. Shegokar P, Sircar P (2016) Continuous wavelet transform based speech emotion recognition. In: 2016 10th International Conference on Signal Processing and Communication Systems (ICSPCS). IEEE, pp 1–8
44. Sinith MS, Aswathi E, Deepa TM, Shameema CP, Shiny R (2015) Emotion recognition from audio signals using Support Vector Machine. In: 2015 IEEE recent advances in intelligent computational systems (RAICS). IEEE
45. Wang K, An N, Li BN, Zhang Y, Li L (2015) Speech emotion recognition using fourier parameters. IEEE Trans Affect Comput 6:69–75
46. Wei\sskirchen N, Böck R, Wendemuth A (2017) Recognition of emotional speech with convolutional neural networks by means of spectral estimates. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), pp 50–55
47. Wu S, Falk TH, Chan W-Y (2011) Automatic speech emotion recognition using modulation spectral features. Speech Commun 53:768–785
48. Yang N, Jianbo Y, Yun Z, Ilker D, Zhiyao D, Wendi H et al (2017) Enhanced multiclass SVM with thresholding fusion for speechbased emotion classification. Int J Speech Technol 20(1):27–41
49. Zeng Y, Mao H, Peng D, Yi Z (2017) Spectrogram based multi-task audio classification. Multimed Tools Appl 78:1–18
50. Zhang B, Provost EM, Essi G (2016) Cross-corpus acoustic emotion recognition from singing and speaking: a multi-task learning approach. In: 2016 IEEE international conference on acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 5805–5809
51. Zhao J, Mao X, Chen L (2019) Speech emotion recognition using deep 1d & 2d cnn lstm networks. Biomed Signal Process Control 47:312–323