SN Computer Science (2024) 5:390
https://doi.org/10.1007/s42979-024-02656-0

ORIGINAL RESEARCH

Speech Emotion Recognition Using Machine Learning: A Comparative Analysis

Sasank Nath[1] · Ashutosh Kumar Shahi[1] · Tekwo Martin[1] · Nupur Choudhury[1] · Rupesh Mandal[1]

Received: 2 June 2023 / Accepted: 26 January 2024

Abstract

It is possible to identify emotions based on a person's speech. The field of research focusing on expressing emotions through voice is continuously evolving. This study utilizes the SAVEE and IEMOCAP datasets to explore Speech Emotion Recognition. The SAVEE dataset consists of seven emotions, while 4 out of 11 emotions are considered from the IEMOCAP dataset. The features ZCR, MFCC, F0, and RMS are extracted from the raw audio files, and their means are calculated which are fed as input for training the models. The study presents a comparative analysis of emotion detection on both datasets, employing the models RNN, LSTM, Bi-LSTM, RF, Rotation Forest, and Fuzzy. The RF and Bi-LSTM models achieve highest accuracies of 76 and 72%, respectively, on the SAVEE dataset, when compared to other trained models. The fuzzy and Rotation Forest models are implemented which can be improvised with further optimization techniques. Additionally, a diagnostic User Interface is developed for analyzing audio, loading datasets, extracting features, training models, and classifying human emotions from audio using the trained models.

Keywords SER · RF · Bi-LSTM · Fuzzy · SAVEE · IEMOCAP

Introduction

Emotions have a substantial impact on interpersonal relationships and greatly influence our cognitive processes and decision-making processes. They enable us to convey intricate emotional information by responding to objects, situations, or events. Speech Emotion Recognition (SER) intends to identify the emotional aspects of speech, independent of its semantic content. In machine learning, a typical approach involves feature extraction from raw data and employing them to train a model [1]. Deep Learning models have also been employed to tackle SER tasks, utilizing diverse features derived from raw audio data [2, 3].

In recent years, the task to identify emotions embedded in audio and video recordings has gained significant attention, driven by a variety of practical applications. The capacity to accurately detect emotions in such media addresses relevant challenges in real-world circumstances where comprehending human emotions can have profound implications. The importance of emotion detection from audio files becomes apparent in a variety of real-life scenarios. In healthcare, for example, recognizing patient emotions while they talk may aid in assessing mental well-being, allowing for early assessments of emotional disorders, or better empathetic treatments. In education, tracking student engagement and emotional states during learning sessions can inform adaptive teaching strategies. Accurately identifying emotions during interrogations is critical in criminal justice for establishing the truthfulness of statements and guaranteeing a more just investigative procedure. Similarly, recognizing the emotional tone of politicians' speech is essential for determining the

This article is part of the topical collection "SWOT to AI-embraced Communication Systems (SWOT-AI)" guest edited by Somnath Mukhopadhyay, Debashis De, Sunita Sarkar and Celia Shahnaz.

✉ Nupur Choudhury
nupur.choudhury@dbuniversity.ac.in

Sasank Nath
sn.adbu@gmail.com

Ashutosh Kumar Shahi
ashutoshshahi139@gmail.com

Tekwo Martin
tekwomartin@gmail.com

Rupesh Mandal
rupesh.mandal@dbuniversity.ac.in

[1] School of Technology, Assam Don Bosco University, Guwahati, India

Published online: 04 April 2024

authenticity of their promises during election campaigns. Moreover, in the corporate world, emotion detection contributes to refining customer service by gauging client satisfaction and sentiment. Whether in personal relationships, professional settings, or educational environments, the ability to identify emotions proves invaluable, influencing various aspects of our life and interactions. As we navigate through an increasingly digital world, the infusion of emotion detection technology into human–computer interaction becomes very important. Understanding the emotional state of users during virtual interactions can enhance user experience, providing a more personalized and meaningful experience. This paper delves into the methodologies and implications of this developing field, to solve the problem of emotion detection through voice signals.

In our study, we employed two speech datasets for emotion recognition, namely SAVEE [4] and IEMOCAP [5]. The feature extraction process utilized Librosa 0.9.2, allowing us to extract meaningful features, such as Fundamental Frequency (F0), Zero-Crossing Rate (ZCR), Root Mean Square (RMS), and Mel-frequency cepstral coefficients (MFCC) from the speech signals. Our work encompasses a range of models, including Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), Bidirectional Long Short-Term Memory (Bi-LSTM), Random Forest (RF), Rotation Forest, and Fuzzy. The aim of this paper is to provide a comprehensive study on emotion recognition using speech signals, employing machine learning and deep learning techniques to accurately classify primary emotions in humans.

Problem Statement

Emotion recognition using speech signals requires the analysis of multiple features extracted from raw audio files. These features can number in the thousands for a single audio file. Additionally, training models with time-series data that encompass different voice channels poses a significant challenge. In our study, we intend to assess the performance of emotion recognition from speech signals by employing machine learning and deep learning models and selecting specific features for analysis. Through this approach, we seek to gain insights into the effectiveness of different models and feature selection techniques for accurate emotion recognition.

Motivation and Scope

The task of detecting emotions from time-series data does not consistently yield higher accuracy rates, indicating the need for further optimization of algorithms and features to enhance accuracy. We explore and compare machine learning and deep learning techniques for emotion recognition and evaluate their performance on two different audio datasets. The structure of our paper is as follows: Sect. 1 shows Introduction which gives an overview of the work, Sect. 2 shows the Literature Review that shows relevant research for the task of speech emotion recognition, Sect. 3 explains the Methods and Methodology which describes the techniques and methodologies used in our study, Sect. 4 gives the Results and Discussion that presents the findings of our analysis and discusses the results and Sect. 6 gives the Conclusion which summarizes the key findings and conclude the paper. The major contributions of this paper are as follows:

This paper does an in-depth study on various emotions using traditional machine learning and Deep learning models which may form as a foundation for futuristic research that can contribute to early detection of depression or negative emotions by analyzing voice notes and providing relevant information about emotional disbalance which can also help a patient's treatment with emotional support.

Analysis of features is done by taking the mean of the features extracted which provides insights to how it results in model performance.

A novel GUI for signal processing has been developed where the researchers and students can use the interactive GUI for audio signal analysis, feature extraction, and training models with minimum complexity involved.

Literature Review

Machine Learning methodologies have been extensively employed over time for a wide range of classification and differentiation tasks. H. Aouani and Y. B. Ayed [6] focused on the classification of emotions using Support Vector Machines (SVM). They extracted features such as Zero-Crossing Rate (ZCR), Mel-frequency cepstral coefficients (MFCC), Teager Energy Operator (TEO), and Harmonic-to-Noise Ratio (HNR), and utilized them to fit SVM for emotion classification. M. Jawad and A. Fatlawi [7] proposed a classification method that combined SVM and KNN, utilizing two feature groups including F0, energy, ZCR, and a second group based on the Fourier Transform utilizing a variety of deep learning techniques. Z. Zhao et al. [8] simultaneously employed attention-based Bi-LSTM and Fully Convolutional Networks (FCN) architectures in parallel to enhance the processing of the input data to capture relevant contextual information to increase their model accuracy. Mustaqeem et al. [9] employed convolutional neural networks (CNN) to extract

distinctive attributes, which were subsequently inputted into a Bi-LSTM model to identify the ultimate emotional state. D. Issa et al. [10] achieved accuracies of 71.61, 86.1, 95.71, and 64.3% on the RAVDESS dataset, EMO-DB dataset with 535 samples, EMO-DB dataset with 520 samples, and IEMOCAP dataset with 4 classes, respectively, by utilizing a one-dimensional CNN on the features MFCC, Mel-scale spectrogram, chromagram, tonnetz representation, and spectral contrast feature. To develop a cross-corpus multi-lingual voice recognition system, W. Zehra et al. [11] utilized an ensemble learning approach incorporating majority voting. To enhance performance, Z. Peng et al. [12] utilized multi-scale convolutional layers along with an attention module, which were implemented on both audio and text inputs. D. Li et al. [13] utilized Bi-LSTM with directional self-attention mechanism to recognize emotions by leveraging the characteristics decoded from the LSTM.

In recent years, there have been numerous studies conducted on a wide range of techniques, including deep learning and traditional machine learning approaches. L. Kerkeni et al. [14] conducted a comparative analysis between RNN, multivariate linear regression, and SVM using audio data and also shows the application of feature selection techniques. Paper [15] employed SVM, KNN, and MLP models on separate text and speech datasets. The experimental results showed that SVM achieved a higher accuracy score compared to the other models. R. Y. Rumagit et al. [16] conducted a study where SVM, MLP, and logistic regression models were trained to classify two emotions, namely happy and sad. The results showed that logistic regression outperformed the other models in terms of overall performance. However, MLP exhibited the highest precision for happy emotion, and highest recall for sad emotion. A. A. Alnuaim et al. [17, 18] utilized MLP model trained on short-time Fourier transform (STFT) and mel spectrum features. Additionally, they also employed a one-dimensional CNN trained on features MFCC, chroma, and ZCR. B. T. Atmaja et al. [19] extracted high-level statistical features exclusively from each acoustic feature set and used them to train MLP and LSTM models. A. Rehman et al. [20] demonstrated that SVM and RF classifiers exhibited comparatively better performance than KNN and MLP classifiers. In their paper [21], a fully CNN has been trained for speech emotion recognition (SER) utilizing MFCC features. In the context of emotion recognition from speech, several research papers [22–25] utilized convolutional neural networks (CNNs) while incorporating various features, such as mel-spectrograms, MFCC, pitch, contrast, ZCR, energy, chroma, linear predictive coding (LPC), and tonnetz format. Different models were experimented in our study, including RNNs and LSTMs, Bi-LSTMs, RF, Rotation Forest, and Fuzzy for audio signal features. Spectrogram images were extracted from the SAVEE dataset (H. Huang et al. [22]) and similar to S. Padi et al. [23], they extracted spectrograms from the IEMOCAP dataset, whereas we used signal features extracted from the same datasets. A. Aggarwal et al. [26] utilized PCA to obtain the initial set of features from mel-spectrograms extracted from audio recordings, and applied the [27] pre-trained VGG-16 model. In our approach, the trained Rotation Forest algorithm uses PCA as dimensionality reduction technique for feature selection. They have used images, and we have experimented with audio signals. H.I. Attar et al. [27] conducted their research on the RAVDESS dataset, analyzing male and female samples separately. They considered five emotion classes and incorporated spectrograms, MFCC, pitch, and energy features in their analysis. We have trained RNN, LSTM and Bi-LSTM, as well as Fuzzy DNN, using the features MFCC, RMS, F0, and ZCR. U. K. Singh et al. [28] compared the accuracy of RF, MLP, and SVM classifiers and achieved accuracies of 65.8%, 66.2%, and 63.2%, respectively. K. S. Raj et al. [29] used fuzzy logic for video and text analysis. We have implemented fuzzy logic for speech signal analysis. A tabular format of the various researches done in this domain is as follows (Table 1):

Methods and Methodology

This study aims to analyze speech emotion detection by employing machine learning and deep learning algorithms including Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), Bidirectional Long Short-Term Memory (Bi-LSTM), Random Forest (RF), Rotation Forest, and Fuzzy algorithms. We develop classification models for identifying basic human emotions using features extracted from raw audio data. The two datasets used in our experiment are the SAVEE dataset, which encompasses seven distinct emotions, and the IEMOCAP dataset, from which we incorporate four emotions out of total eleven emotions in the dataset due to imbalance of the number of samples per emotion. The models are trained using both datasets, and a comparative analysis of the results is performed. The feature set comprises Fundamental Frequency (F0), Zero-Crossing Rate (ZCR), Root Mean Square (RMS), and Mel-frequency cepstral coefficient (MFCC).

Block Diagram

The block diagram shown below illustrates the workflow of the work (Fig. 1). The initial step involves importing the dataset, followed by the extraction of features from the audio files. Then, the extracted features are pre-processed to prepare them for model training. The models are then trained using the features, and the obtained results are compared to evaluate their performances. Finally, the trained models are utilized for classifying emotions on unseen audio files.

Table 1 Existing work, databases, and models in SER

YOP	Author names	Database used	Classifiers and features used
2019	Ziping Zhao et al. [8]	IEMOCAP, FAU Aibo Emotion Corpus (FAU-AEC)	Attention-based CNN + BiLSTM
2019	Leila Kerkeni et al. [14]	Berlin and Spanish database	RNN vs (MLR and SVM). MFCC and modulation spectral (MS)
2020	Hadhami Aouani and Yassine Ben Ayed [6]	Ryerson Multimedia Laboratory (RML)	SVM MFCC, ZCR, HNR and TEO
2020	Mustaqeem et al. [9]	IEMOCAP, EMO-DB, and RAVDESS	CNN + BiLSTM Spectrogram
2020	Dias Issa et al. [10]	RAVDESS, EMO-DB, IEMOCAP	CNN MFCCs, Mel-scaled spectrogram, Chromagram, Spectral contrast feature, Tonnetz representation
2021	Mohammed Jawad, Ahmed Fatlawi [7]	The German Berlin database, SAVEE	SVM, KNN Fundamental frequency (F0), energy (E), zero-crossing rate (ZCR) (FEZ), Fourier parameter (FP) model
2021	Wisha Zehra et al. [11]	SAVEE, URDU, EMO-DB, and EMOVO (English, Urdu, German, and Italian)	Ensemble learning approach through majority voting eGeMAPS, which consists of 88 features connected to energy, spectrum, frequency, Cepstral, and dynamic information
2021	Zixuan Peng et al. [12]	IEMOCAP	CNN. MFCC
2021	Dongdong Li et al. [13]	IEMOCAP Spontaneous, Scripted and complete	Bi-LSTM Bidirectional LSTM decoded features
2021	Reem Hamed Aljuhani et al. [15]	Saudi dialect semi-natural emotion speech dataset	SVM, KNN, MLP MFCC and mel spectrogram
2021	Arya Aftab et al. [21]	IEMOCAP, EMO-DB	CNN Mel-frequency cepstral coefficients (MFCC)
2021	Zhengwei Huang, et al.[22]	SAVEE, Emo-DB, DES, Mandarin Emotional Speech database	Semi-CNN Spectrogram
2021	Sarala Padi et al. [23]	IEMOCAP	ResNet model (CNN) Log-mel-spectrograms
2022	Abeer Ali Alnuaim et al. [17]	RAVDESS	MLP Short-time Fourier transform and Mel-spectrum features
2022	Abeer Ali Alnuaim et al. [18]	BAVED, ANAD, and SAVEE	1D CNN MFCC, Chroma, and ZCR
2022	Bagus Tris Atmaja et al. [19]	MSP-IMPROV	MLP and LSTM pyAudioAnalysis (pAA), ComParE, eGeMAPS, EMOBASE
2022	Abdul Rehman et al. [20]	IEMOCAP, MSP-IMPROV, RAVDESS	SVM, RF, KNN, MLP classifiers
2022	Chen Jin, A.I. et al. 24]	SAVEE, RAVDESS	CNN MFCC
2022	Kamaldeep Kaur, Parminder Singh [25]	An emotional Punjabi database has been created for the purpose of SER	CNN MFCC, ZCR, LPCC, tonnetz, formant, jitter, shimmer, entropy, duration, harmonic, Perceptual Linear Prediction (PLP), and energy
2022	Apeksha Aggarwal et al. [26]	RAVDESS	DNN, VGG-16 Mel-spectrogram
2022	Husbaan I. Attar et al. [27]	RAVDESS	LSTM, CNN, Hidden Markov Models (HMMs) and Deep Neural Networks (DNNs) MFCC, pitch and energy were considered
2022	Utkarsh Kumar Singh et al. [28]	RAVDESS, IEMOCAP	RF, Gradient Boost, SVM, Logistic Regression, MNB, and MLP. Audio-only, Text-only, Audio + Text

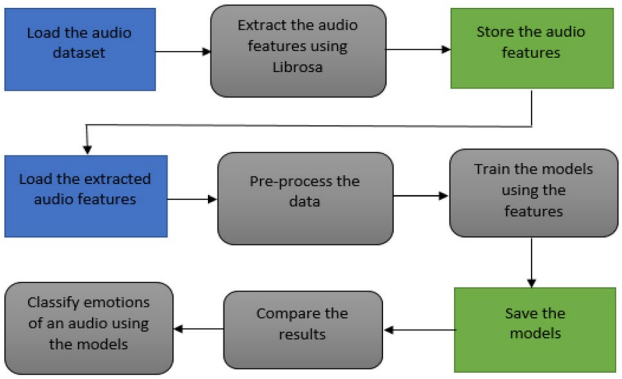


Fig.1 Block diagram of the work

Dataset Description

Surrey Audio-Visual Expressed Emotion (SAVEE): This database comprises 480 utterances in British English from recordings of four male actors showing seven distinct emotions (i.e., angry, happiness, sadness, fear, disgust, neutral, and surprise). Each emotion category in the dataset accounts for 12.5% of the total, except for the neutral which accounts for 25% of the dataset. It has been created as a need for the creation of an automatic speech recognition system. Classification systems were generated, resulting in recognition rates of 61 and 65% for audio and visual modalities, respectively, while ensuring speaker independence. The dataset was formally requested from its authors and duly granted.

The Interactive Emotional Dyadic Motion Capture (IEMOCAP): The IEMOCAP is a multi-modal dataset comprising audio, video, facial motion capture, and text transcriptions. The audio data in sentences are 10,038 files, out of which four emotions categories (anger, happiness, neutral, and sadness) are considered, resulting in 4490 files. The male and female samples are trained and tested independently resulting in 2284 and 2206 files, respectively. It is an acted, multi-speaker, and multi-modal database at USC's SAIL Lab. Access to the IEMOCAP dataset required submission of an electronic form to its authors.

Performance Metrics to Be Used

Confusion Matrix—The performance of a classification algorithm is determined by the accuracy of its predictions, in terms of true labels vs predicted labels as correctly predicted or false predicted.

Precision—Precision calculates the number of predicted positive classes that are truly positive.

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}} \quad (1)$$

Recall—Recall calculates the number of predicted positive classes out of all actual positive samples in the dataset.

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} \quad (2)$$

F1-score—F1-score is computed by taking the harmonic mean of precision and recall

$$F1 - \text{score} = 2 * \frac{\text{Precision} * \text{recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Support—The support in a classification context refers to the number of actual instances of a class present in a given dataset.

Accuracy and loss plots—Evaluate the performance of deep learning models during training to identify and address issues like overfitting or underfitting for further improvement.

Features

The voice represents various signals, which vary with emotional states. The modulation of pitch, tone, intensity, and other acoustic features in speech often correlates with an individual's emotional expression. Higher pitch and variations in tone may indicate excitement, happiness, or stress, while lower pitch and a monotonous tone could be associated with sadness or boredom. High speech rate might signify excitement, anxiety, or nervousness, whereas slow speech could be linked to sadness or contemplation. Increased loudness may convey anger or enthusiasm, while decreased loudness might suggest sadness or fatigue. Changes in prosody, including rhythm and intonation, can provide additional clues about emotional states. We have used relevant features for detecting emotions from speech, such as ZCR, MFCC, F0, and RMS which provide crucial information about the tone, pitch, and other speech elements. Researchers continue to analyze and interpret these voice signal features, aiming to accurately identify and classify various emotional states (figs. 2, 3, 4, 5, 6, 7, 8).

The features are extracted using the Librosa library with version 0.9.2. The default sample rate is set to 22,050 per second. A total of 22 features were extracted, and any NULL or infinite values were replaced with 0. The features used to train the models are defined below. The signals are extracted and ZCR features for different emotions are plotted and shown below.

This shows that there is significant difference in the signals based on their emotions where the features are considered.

Zero-Crossing Rate (ZCR)—The ZCR measures the rate at which a signal shifts from positive to zero to negative, or vice versa. It serves as an indicator of the noisiness or rapid changes in the signal. The ZCR provides insights

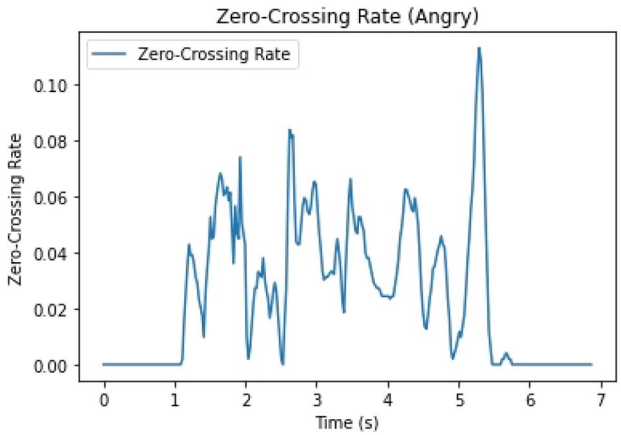


Fig.2 ZCR plot for angry

into certain aspects of vocal characteristics associated with emotions. A higher ZCR value indicates rapid changes in the signal, indicating more high-frequency components or noise. This might be associated with increased intensity or energy in speech. Higher ZCR might be indicative of a faster speech rate, which corresponds to emotions like excitement, anger, or nervousness. Whereas a lower ZCR value suggests a smoother or more constant signal with fewer changes in sign, suggesting a lower intensity or a more controlled emotional state, such as calmness or sadness. Emotional states often involve changes in pitch, intensity, and other parameters, which can contribute to a higher ZCR. Different emotional states manifest in distinct patterns of ZCR

$$\text{ZCR} = \frac{1}{T}\sum_{t=1}^{T} |s(t) - s(t-1)|, \tag{4}$$

where s(t) = 1 if signal > 0 at time t, and 0 otherwise.

Mel-Frequency Cepstrum Coefficient (MFCC)—MFCCs are a feature representation obtained by transforming a signal from the time domain to the frequency domain using a set of mel filters. MFCCs are widely used speech recognition due to their ability to capture relevant spectral information and their effectiveness in analyzing audio signals. MFCCs are coefficients that represent the short-term power spectrum of a sound, particularly emphasizing the frequencies that are most relevant to human hearing, which can be informative about emotional states expressed in speech. MFCCs capture information about the distribution of spectral energy across different frequency bands. The distribution in energy reflects variations in pitch and prosody, which are crucial components of emotional expression. Emotional states often involve changes in pitch, and these changes are captured in the MFCCs. Emotional expression in speech often involves changes in timbre, which is the quality or color of the sound. The spectral characteristics captured by MFCCs contribute to the representation of speech timbre. Different emotional states may manifest as variations in the timbral qualities of speech. Changes in emotional intensity may be reflected in the amplitude of certain MFCCs. Higher energy in specific frequency bands can indicate increased loudness or intensity in speech, which may be associated with emotional states like excitement or anger. Different emotions may have distinct spectral signatures that are captured by the MFCCs. Models trained on a dataset of labeled emotional speech can learn to associate certain patterns in MFCCs with specific emotional states.

Each MFCC represents a specific frequency component of the audio signal after a series of processing steps, and their index corresponds to the order in which they are computed. The first MFCC is often associated with the overall energy of the signal, while subsequent coefficients capture more detailed frequency information and are considered higher MFCCs. As such, MFCC[0], often referred to as the

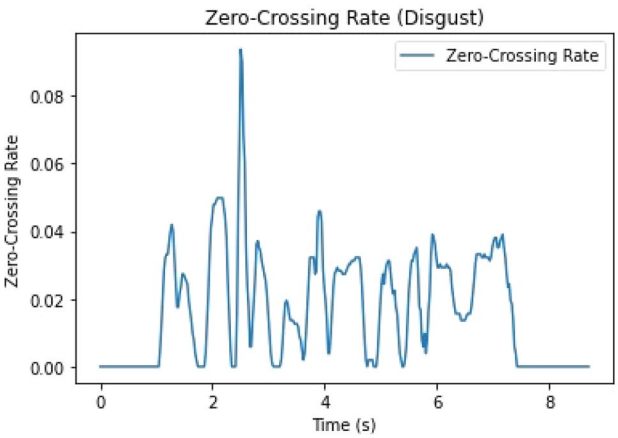


Fig.3 ZCR plot for Disgust

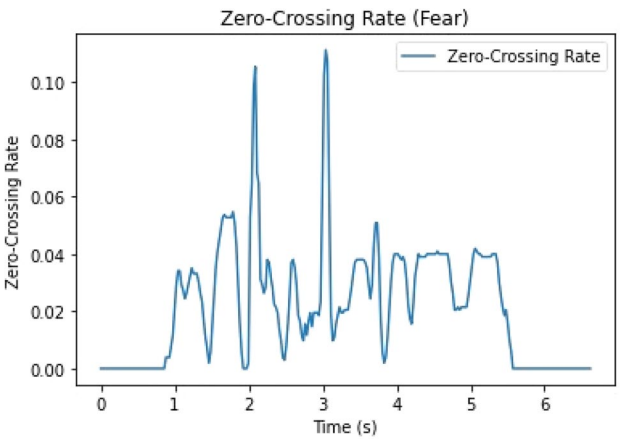


Fig.4 ZCR plot for fear

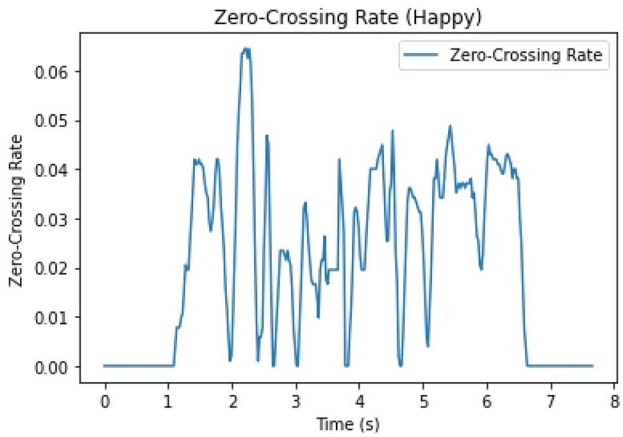


Fig.5 ZCR plot for happy

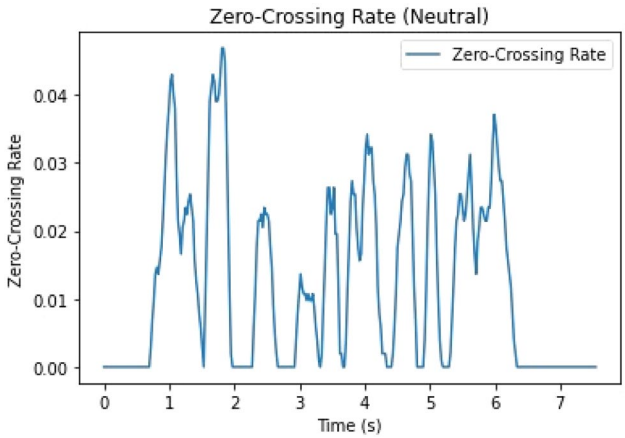


Fig.6 ZCR plot for neutral

constant or offset term, represents the overall energy or loudness of the signal, dominated by the average power of the signal. Higher MFCCs (MFCC[1], MFCC[2],…) capture detailed frequency information and spectral characteristics. Higher indexes correspond to higher frequencies and finer details in the signal, emphasizing variations in spectral shape that are important for speech and audio analysis. Changes in the higher MFCCs may be more relevant for capturing emotional nuances, as they are sensitive to variations in the spectral content of speech. In speech and audio processing, it is common to use a subset of the higher MFCCs (e.g., MFCC[1] to MFCC[12]) for various applications, including speech recognition and emotion analysis. The lower MFCC (MFCC[0]) is often used for overall loudness normalization. As for our study, we have utilized the first 20 MFCCs for training the models and improved performance

$$c[n] = \sum_{m=0}^{m-1} s[m] \cos\left(\pi n\left(m + \frac{1}{2}\right)/M\right) \; 0 \le n < M \tag{5}$$

where,

$$s[m] = \ln\left[\sum_{k=0}^{N-1} |X_a[k]|^2 H_m[k]\right], \; 0 \le m \; < M$$

Fundamental Frequency (F0)—Fundamental frequency (F0) refers to the lowest frequency component of a waveform that repeats at regular intervals. It represents the fundamental pitch of a sound and is an important parameter in speech analysis and music processing. It is a fundamental acoustic property of a sound signal that represents the perceived pitch of a sound and is associated with the rate of vibration of the vocal cords in speech. General observations state that an increase in F0 is often associated with excitement, enthusiasm, or happiness. A higher pitch may convey positive and energetic emotions. Higher pitch levels can also indicate stress or anxiety. In situations of tension, individuals may speak at a higher pitch. Whereas lower F0 values are often associated with sadness or a more subdued emotional state. A drop in pitch may convey a sense of melancholy. Lower pitch levels can also be linked to calmness or confidence. In situations where an individual feels secure or composed, the pitch tends to be lower. As emotions are multidimensional, F0 is one aspect of the acoustic features that contribute to emotional expression in speech. The integration of acoustic features helps capture the richness and complexity of emotional expression

$$f_0 = \frac{v}{2L}, \tag{6}$$

where.

v = speed of the wave, and.

L = length of the pipe.

Root Mean Square (RMS)—RMS is another acoustic feature commonly used in speech and audio signal processing. It represents the energy of a set of values. RMS is often used to quantify the amplitude or loudness of the signal. RMS provides a measure of the overall energy or loudness of a signal. Changes in emotional state often correlate with variations in speech intensity. For example, high RMS may be associated with heightened emotions such as excitement, anger, or stress, while low RMS related to more subdued emotional states like sadness or calmness. RMS captures dynamic changes in intensity and can reflect the dynamic nature of emotional expression. Quick fluctuations in RMS may correspond to rapid changes in emotional states or dynamic speech patterns. The interpretation of RMS in emotional expression is context-dependent. Factors, such as linguistic, individual differences, and speaking style, also influence the relationship between RMS and emotion. The combination of RMS with other features, such as pitch, MFCCs, and duration, improves the accuracy of emotional state prediction.

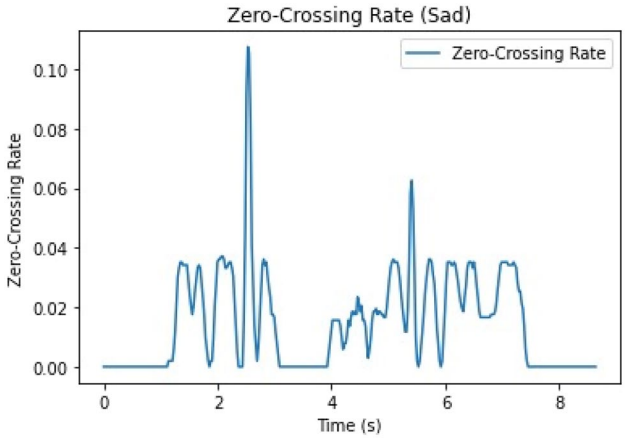


Fig.7 ZCR plot for neutral

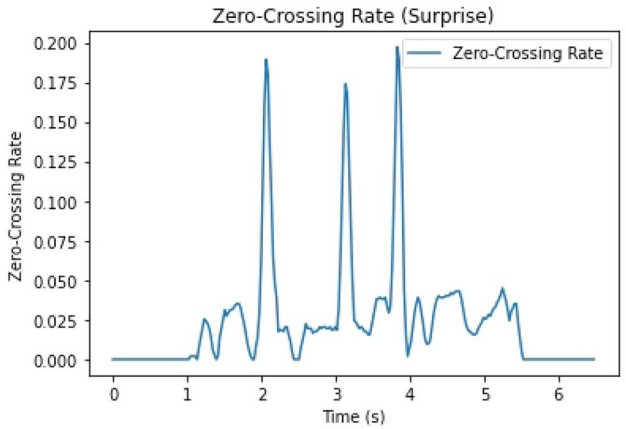


Fig.8 ZCR plot for neutral

If we have a set of n values $\{x_1, x_2, \ldots, x_n\}$, the RMS is calculated as

$$x_{\text{RMS}} = \sqrt{\frac{1}{n}(x_1^2 + x_2^2 + \ldots + x_n^2} \tag{7}$$

Data Preprocessing

Emotion labels and features are individually extracted from the datasets. The IEMOCAP dataset is pre-processed to have distinct samples of male and female actors, so as to see if the recognition varies according to male and female voices, and it further aided in comparison with the SAVEE dataset, as SAVEE contains only male actors. Also, 4 emotions (anger, happiness, neutral, and sadness) out of a total 11 emotions are considered from the IEMOCAP dataset due to imbalance of the emotion categories in the dataset. Mean of the features ZCR, F0, MFCC, and RMS are calculated for each sample. We extracted 22 features altogether and the NULL and infinite value were replaced with 0. For both datasets, a train and test split was performed using stratified sampling, with 80% for training and 20% for testing. This sampling approach ensured that the distribution of actors and emotions remained proportional in both the training and testing datasets. Standard scaler is used to normalize the features, such that the mean is 0 and the standard deviation is 1 of the distribution.

Models

Recurrent Neural Network (RNN)—RNNs utilize their internal memory to process input sequences. However, complex architectures can sometimes lead to poor performance on certain datasets. In this study, a simpler architecture is adopted, consisting of a single SimpleRNN layer, followed by two dense layers with ReLU activation functions (Table 2). The output dense layer is defined with the softmax activation function.

Long short-term memory (LSTM)—LSTMs include cells that can retain and store information over extended periods. Similar architecture as RNN is defined here also (Table 3).

Bidirectional Long Short-Term Memory (Bi-LSTM)—Bi-LSTMs are also a type of RNNs consisting of two LSTMs in a single layer where one passes information forward and

Table 2 RNN model summary

Layer (type)	Output shape	Parameters
simple_rnn (SimpleRNN)	(None, 2048)	4,198,400
dense (Dense)	(None, 512)	1,049,088
dense_1 (Dense)	(None, 128)	65,664
dense_2 (Dense)	(None, 7)	903

Total parameters 5,314,055
Trainable parameters 5,314,055
Non-trainable parameters 0

Table 3 LSTM model summary

Layer (type)	Output shape	Parameters
lstm (LSTM)	(None, 2048)	16,793,600
dense (Dense)	(None, 512)	1,049,088
dense_1 (Dense)	(None, 128)	65,664
dense_2 (Dense)	(None, 7)	903

Total parameters 17,909,255
Trainable parameters 17,909,255
Non-trainable parameters 0

the other backward. This allows the network to capture both past and future dependencies in the input sequence. The architecture is similar to other models, as shown below (Table 4).

Random Forest (RF)—Random Forest classification employs a majority voting approach to determine the final class. During training, multiple decision trees are generated, with each tree providing its own class prediction. The final class is then determined by selecting the class with the highest frequency among all the predicted classes. In our study, we incrementally increase the number of trees from one in each iteration. For each random forest, consisting of a specific number of trees, we conduct training and testing. We repeat this process until we reach a total of one hundred trees. From the collected accuracies, we select the model with highest accuracy as the best performing model (Fig. 9).

Rotation Forest [30]—In our study, we incorporate Rotation Forest for Speech Emotion Recognition (SER), following a similar approach to random forest. Rotation Forest is an ensemble method that utilizes a collection of base classifiers, each trained on a specific subset of features. The idea behind Rotation Forest is to enhance the diversity and individual performance of the base classifiers by applying feature rotations. These rotations involve transforming the feature space to different perspectives, which helps capture different aspects of the data and improve the overall classification accuracy.

Table 4 Bi-LSTM model summary

Layer (type)	Output Shape	Parameters
bidirectional (Bidirectional)	(None, 4096)	33,587,200
dense (Dense)	(None, 512)	2,097,664
dense_1 (Dense)	(None, 128)	65,664
dense_2 (Dense)	(None, 7)	903

Total parameters 35,751,431
Trainable parameters 35,751,431
Non-trainable parameters 0

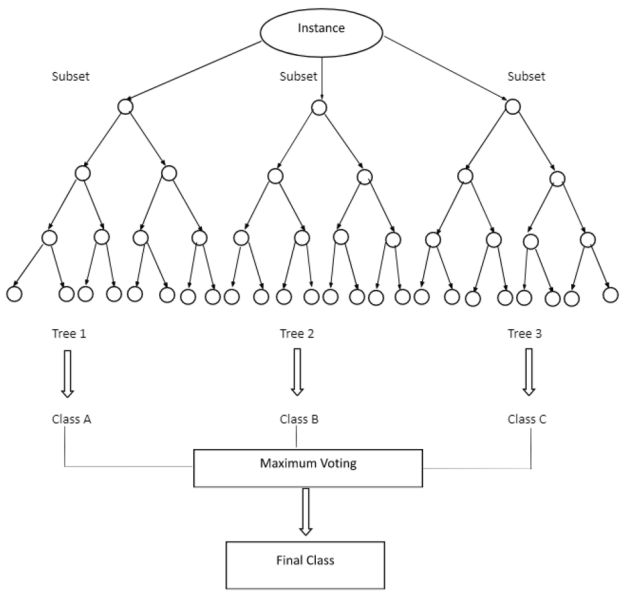


Fig.9 Random Forest architecture for classification

Table 5 Fuzzy model summary

Layer (type)	Output Shape	Parameters
fuzzy_layer (FuzzyLayer)	(None, 23, 64)	128
fuzzy_layer_1 (FuzzyLayer)	(None, 23, 64)	8192
dense (Dense)	(None, 23, 64)	4160
global_max_pooling1d (GlobalMaxPooling1D)	(None, 64)	0
dense_1 (Dense)	(None, 7)	455

Total parameters 12,935
Trainable parameters 12,935
Non-trainable parameters 0

Fuzzy—Fuzzy logic is a form of multi-valued logic that assigns truth values between 0 and 1, allowing for the representation of partial truth. It is a framework that can handle uncertainty and vagueness in decision-making processes. In our study, we incorporate fuzzy logic algorithms and utilize neural networks for training. The fuzzy algorithm is integrated with neural networks to enhance the learning and decision-making capabilities of the model. Specifically, fuzzy layers are defined and imported as part of the network architecture. The network architecture includes two fuzzy layers, followed by a dense layer with ReLU activation function (Table 5). A global max pooling layer is employed to capture the maximum value across all features. The output is then passed through a dense layer to produce the final classification. By combining fuzzy logic with neural networks, we aim to leverage the strengths of both approaches and improve the model's ability to handle uncertainty and capture complex patterns in speech data.

Results and Discussion

Classification report of the emotions on the SAVEE dataset based on the trained models is shown below (Tables 6, 7, 8, 9) (Figs. 10, 11, 12, 13).

Classification reports of the emotions on the IEMOCAP dataset based on the trained models are shown below (Tables 10, 11, 12, 13) (Figs. 14, 15, 16, 17).

On the SAVEE dataset, the emotions with the highest precision are sad and disgust; however, the neutral emotion shows the highest recall and f1-score. On the IEMOCAP

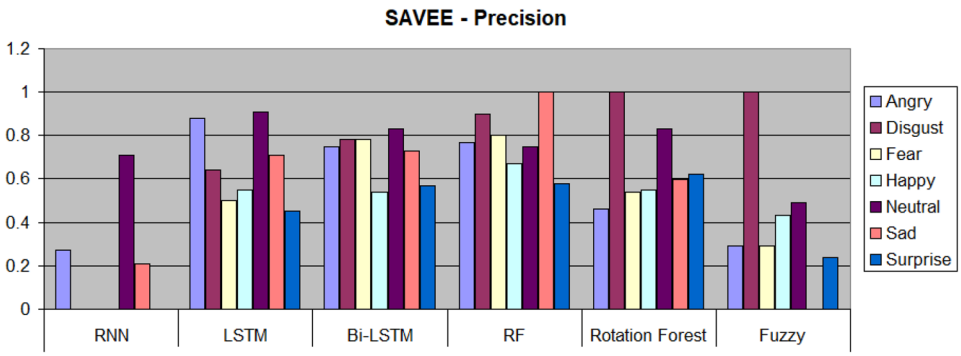


Fig.10 SAVEE—Precision

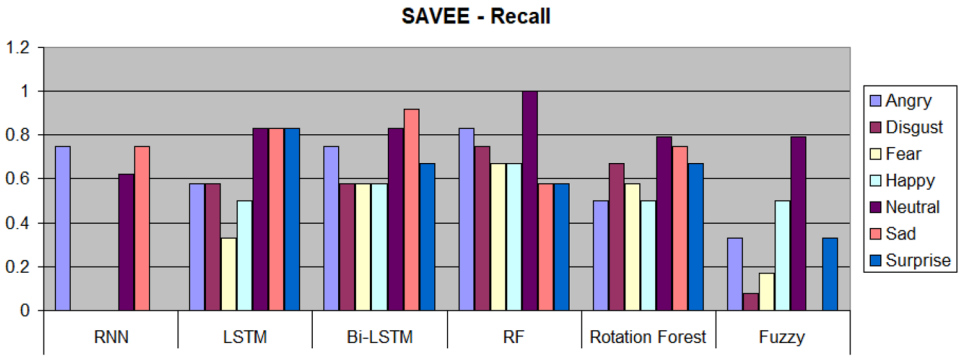


Fig.11 SAVEE—Recall

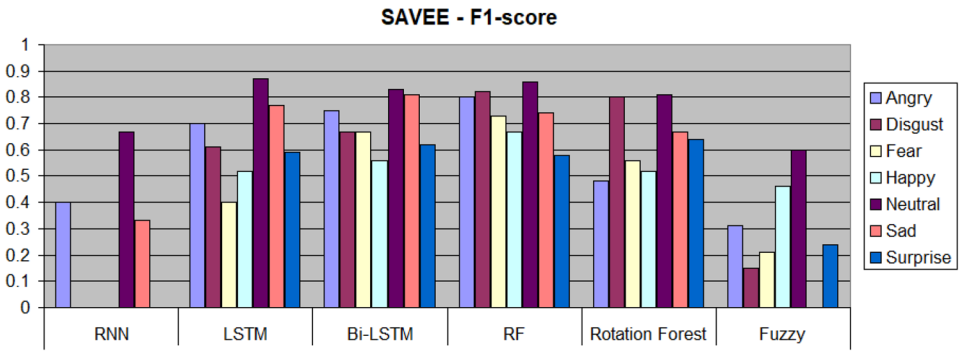


Fig.12 SAVEE—F1-score

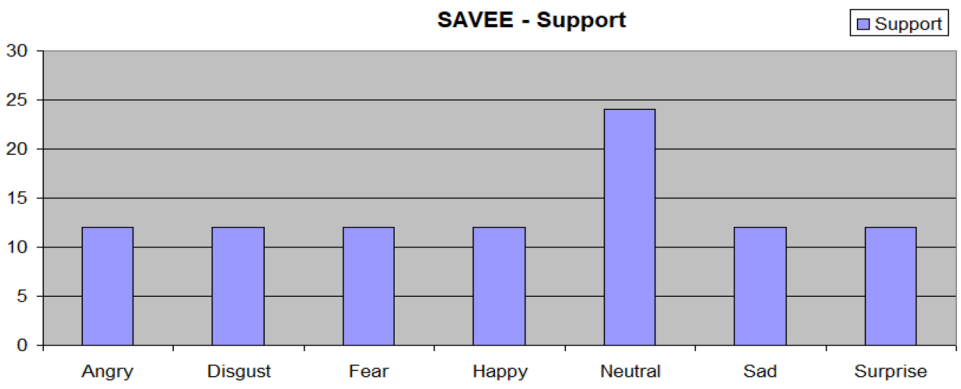


Fig.13 SAVEE—Support

Table 6 SAVEE—Precision

Precision						
	RNN	LSTM	Bi-LSTM	RF	Rotation Forest	Fuzzy
Angry	0.27	0.88	0.75	0.77	0.46	0.29
Disgust	0	0.64	0.78	0.9	1	1
Fear	0	0.5	0.78	0.8	0.54	0.29
Happy	0	0.55	0.54	0.67	0.55	0.43
Neutral	0.71	0.91	0.83	0.75	0.83	0.49
Sad	0.21	0.71	0.73	1	0.6	0
Surprise	0	0.45	0.57	0.58	0.62	0.24

Table 7 SAVEE—Recall

	Recall					
	RNN	LSTM	Bi-LSTM	RF	Rotation Forest	Fuzzy
Angry	0.75	0.58	0.75	0.83	0.5	0.33
Disgust	0	0.58	0.58	0.75	0.67	0.08
Fear	0	0.33	0.58	0.67	0.58	0.17
Happy	0	0.5	0.58	0.67	0.5	0.5
Neutral	0.62	0.83	0.83	1	0.79	0.79
Sad	0.75	0.83	0.92	0.58	0.75	0
Surprise	0	0.83	0.67	0.58	0.67	0.33

Table 8 SAVEE—F1-score

	F1-score					
	RNN	LSTM	Bi-LSTM	RF	Rotation Forest	Fuzzy
Angry	0.4	0.7	0.75	0.8	0.48	0.31
Disgust	0	0.61	0.67	0.82	0.8	0.15
Fear	0	0.4	0.67	0.73	0.56	0.21
Happy	0	0.52	0.56	0.67	0.52	0.46
Neutral	0.67	0.87	0.83	0.86	0.81	0.6
Sad	0.33	0.77	0.81	0.74	0.67	0
Surprise	0	0.59	0.62	0.58	0.64	0.24

dataset, the angry emotion achieved the highest precision and f1-score, while the neutral has the highest recall. The emotions in the SAVEE dataset are evenly distributed except for neutral which has twice the samples compared to other emotions, and there is a slight imbalance of samples per emotion in the IEMOCAP dataset.

Table 9 SAVEE—Support

Emotions	Support
Angry	12
Disgust	12
Fear	12
Happy	12
Neutral	24
Sad	12
Surprise	12

Accuracies of the models on the SAVEE and IEMOCAP datasets are shown below (Table 14) (Fig. 18).

The RF achieved the highest accuracy of 76% on the SAVEE dataset. Accuracies achieved by RF, Bi-LSTM, LSTM, Rotation Forest, Fuzzy, and RNN are 76%, 72%, 67%, 66%, 38%, and 34%, respectively.

Accuracies achieved on male samples are 68%, 64%, 61%, 53%, 48%, and 47% by RF, Bi-LSTM, LSTM, Rotation Forest, RNN, and Fuzzy, respectively on the IEMOCAP dataset. The accuracies achieved on the female samples of the IEMOCAP dataset are 67%, 63%, 62%, 54%, 46%, and 42% by RF, Bi-LSTM, LSTM, Rotation Forest, RNN, and Fuzzy, respectively.

The confusion matrix of the RF and Bi-LSTM model and accuracy vs loss plot of the RNN model is shown (Figs. 19, 20). RF gave the highest accuracy followed by

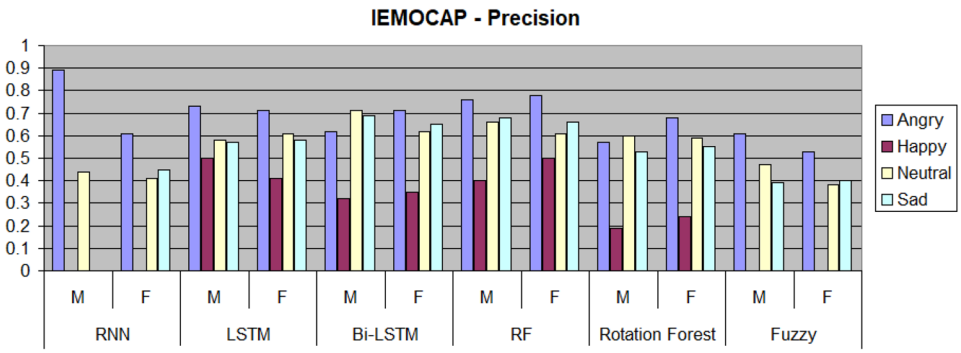


Fig.14 IEMOCAP—Precision

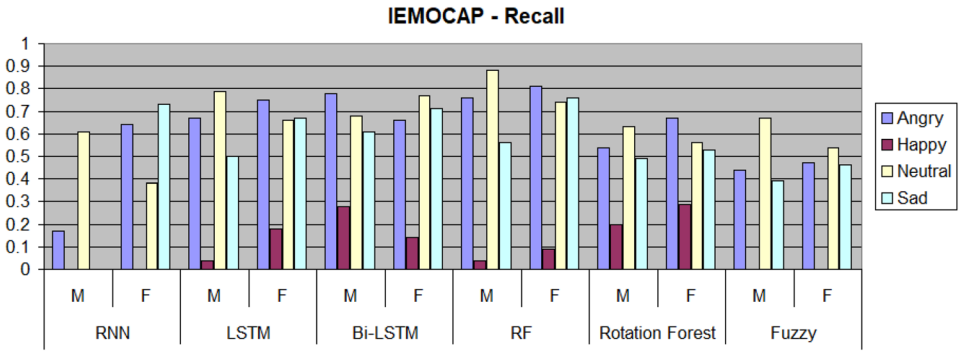


Fig.15 IEMOCAP—Recall

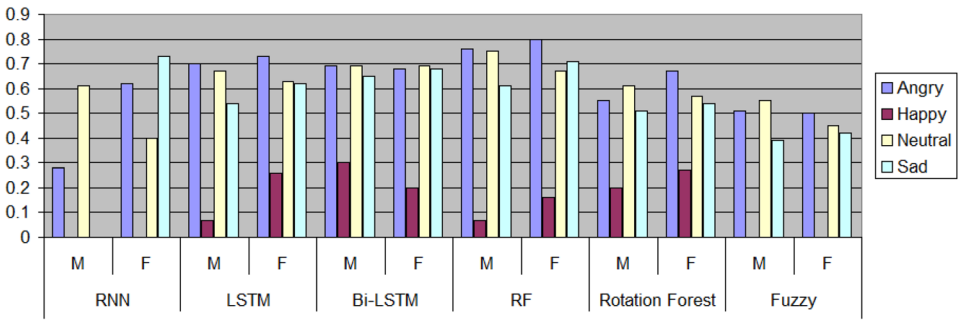


Fig.16 IEMOCAP—F1-score

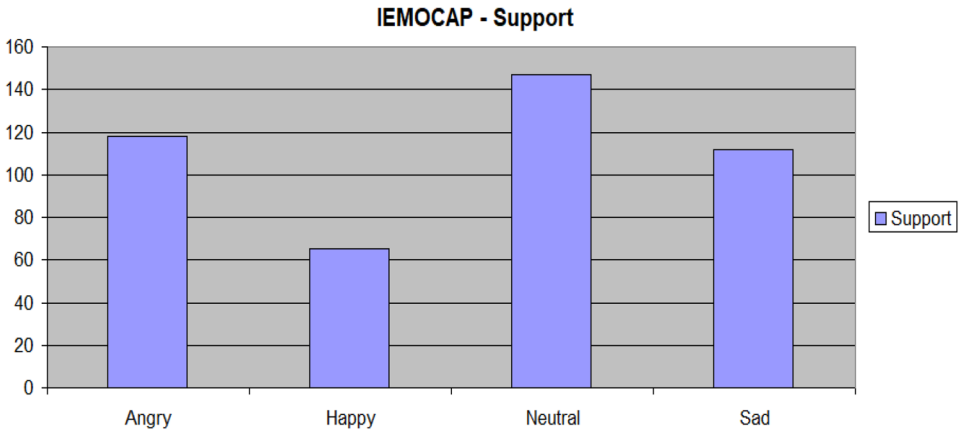


Fig.17 IEMOCAP—Support

Table 10 IEMOCAP—Precision

	Precision											
	RNN		LSTM		Bi-LSTM		RF		Rotation Forest		Fuzzy	
	M	F	M	F	M	F	M	F	M	F	M	F
Angry	0.89	0.61	0.73	0.71	0.62	0.71	0.76	0.78	0.57	0.68	0.61	0.53
Happy	0	0	0.5	0.41	0.32	0.35	0.4	0.5	0.19	0.24	0	0
Neutral	0.44	0.41	0.58	0.61	0.71	0.62	0.66	0.61	0.6	0.59	0.47	0.38
Sad	0	0.45	0.57	0.58	0.69	0.65	0.68	0.66	0.53	0.55	0.39	0.4

Table 11 IEMOCAP—Recall

	Recall											
	RNN		LSTM		Bi-LSTM		RF		Rotation Forest		Fuzzy	
	M	F	M	F	M	F	M	F	M	F	M	F
Angry	0.17	0.64	0.67	0.75	0.78	0.66	0.76	0.81	0.54	0.67	0.44	0.47
Happy	0	0	0.04	0.18	0.28	0.14	0.04	0.09	0.2	0.29	0	0
Neutral	0.61	0.38	0.79	0.66	0.68	0.77	0.88	0.74	0.63	0.56	0.67	0.54
Sad	0	0.73	0.5	0.67	0.61	0.71	0.56	0.76	0.49	0.53	0.39	0.46

Table 12 IEMOCAP—F1-score

	F1-score											
	RNN		LSTM		Bi-LSTM		RF		Rotation Forest		Fuzzy	
	M	F	M	F	M	F	M	F	M	F	M	F
Angry	0.28	0.62	0.7	0.73	0.69	0.68	0.76	0.8	0.55	0.67	0.51	0.5
Happy	0	0	0.07	0.26	0.3	0.2	0.07	0.16	0.2	0.27	0	0
Neutral	0.61	0.4	0.67	0.63	0.69	0.69	0.75	0.67	0.61	0.57	0.55	0.45
Sad	0	0.73	0.54	0.62	0.65	0.68	0.61	0.71	0.51	0.54	0.39	0.42

Table 13 IEMOCAP—Support

Emotions	Support
Angry	118
Happy	65
Neutral	147
Sad	112

Bi-LSTM, and RNN achieved the lowest accuracy on the SAVEE dataset.

Confusion matrix of RF and Bi-LSTM on IEMOCAP (Male) and IEMOCAP (Female), Accuracy and loss plots of the fuzzy model are shown (Figs. 21, 22, 23). RF achieved the highest accuracy, whereas fuzzy achieved the lowest accuracy. However, the accuracy of the fuzzy model improved from 38 to 47% in the SAVEE to IEMOCAP dataset.

The models trained using IEMOCAP dataset are tested on SAVEE dataset and vice versa (Tables 15, 16).

The models trained using the IEMOCAP dataset performed relatively better when tested on the SAVEE dataset, whereas the models trained using the SAVEE dataset gave lower accuracies when tested on the IEMOCAP dataset. The inconsistent number of unique emotions of the datasets was mapped to have the same number of emotions as they were trained, which caused a difference in performance of the models.

We have compared the accuracies of our experimentation with some papers in the past which also included the same models in theirs (Table 17).

Diagnostic Tool for SER –

A diagnostic tool has been designed for the purpose of analyzing audio data, extracting audio features, training, and analyzing the implemented machine learning models and deep learning models used in our study (Fig. 24). The tool is inspired from the WEKA [31, 32] tool, which majorly

Fig.18 Model accuracies

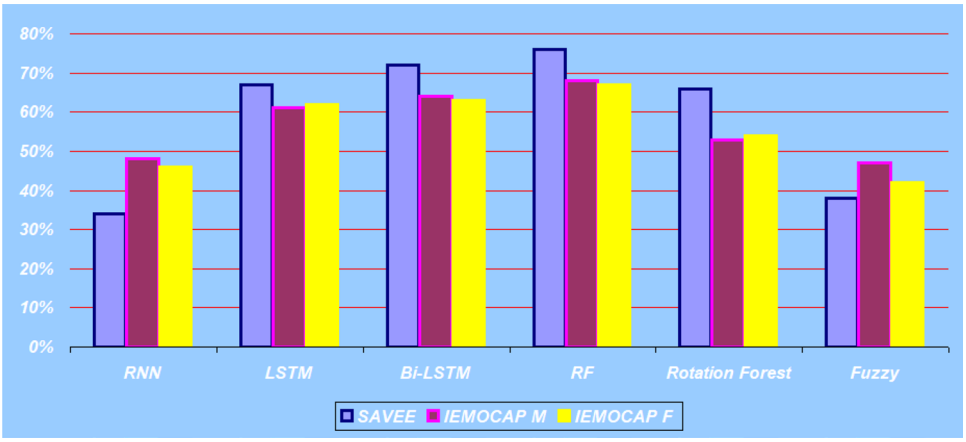


Table 14 Model accuracies

Model/Dataset		RNN	LSTM	Bi-LSTM	RF	Rotation Forest	Fuzzy
SAVEE		34%	67%	72%	76%	66%	38%
IEMOCAP	M	48%	61%	64%	68%	53%	47%
	F	46%	62%	63%	67%	54%	42%

Fig.19 Confusion matrix of RF and Bi-LSTM (SAVEE)

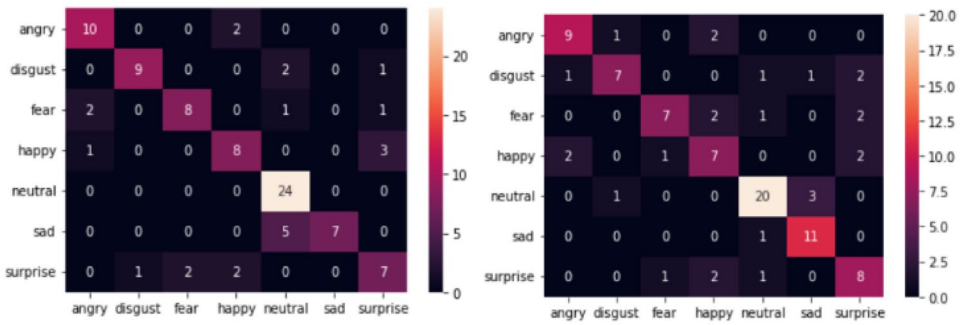
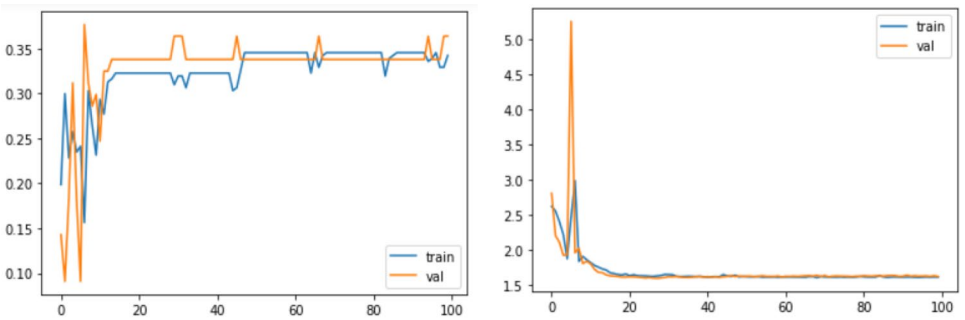


Fig.20 Accuracy vs loss of RNN (SAVEE)



serves the purpose of data mining. The UI is made using Tkinter, which is the standard GUI library for Python. The GUI enables the user to load, extract features, as well as select ML models for classification. In addition, the user is also enabled with the flexibility to choose individual or combination-based features to train the models. The tool has enhanced features of selecting and splitting the dataset as well as choosing appropriated cross validation models for carrying out the research works effectively.

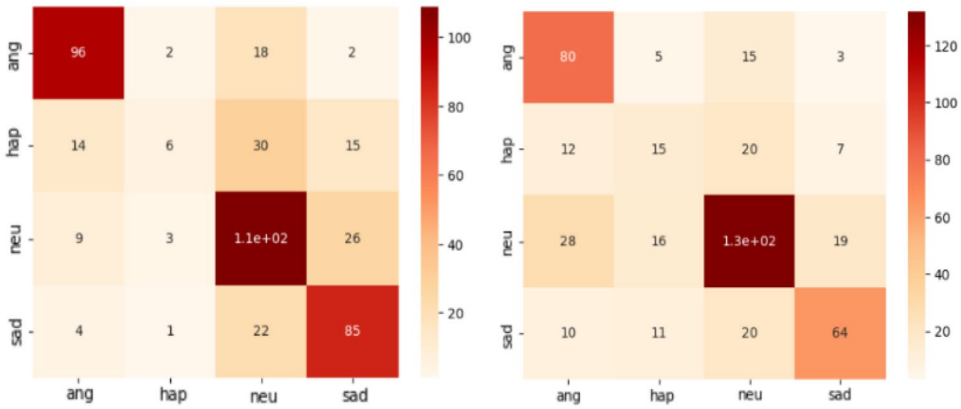


Fig.21 Confusion matrix of RF (Female) and Bi-LSTM (Male) (IEMOCAP)

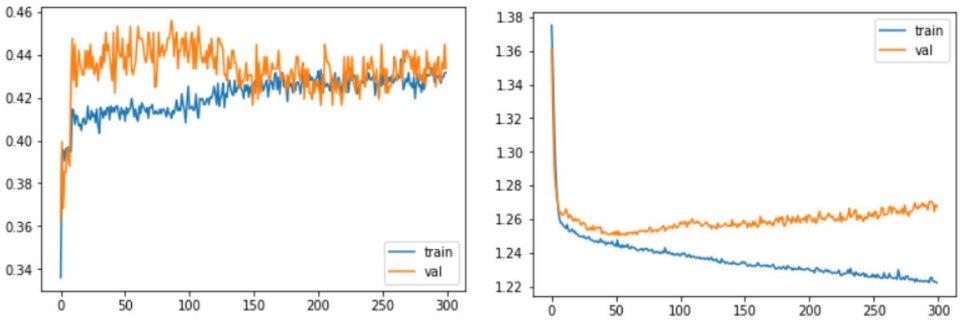


Fig.22 Accuracy vs loss of fuzzy (IEMOCAP Female)

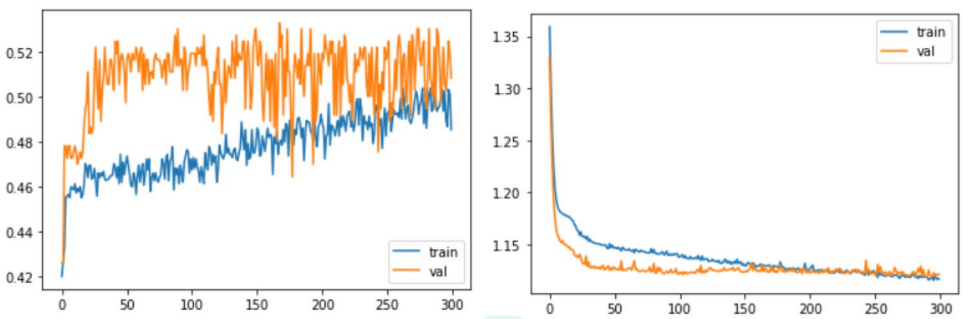


Fig.23 Accuracy vs loss of fuzzy (IEMOCAP Male)

Table 15 IEMOCAP model accuracies on SAVEE dataset

Models	Accuracies
Random Forest	93%
Rotation Forest	88%
Fuzzy	28%
RNN	57%
LSTM	40%
Bi-LSTM	46%

Table 16 SAVEE model accuracies on IEMOCAP dataset

Models	Accuracies
Random Forest	66%
Rotation Forest	40%
Fuzzy	25%
RNN	53%
LSTM	40%
Bi-LSTM	50%

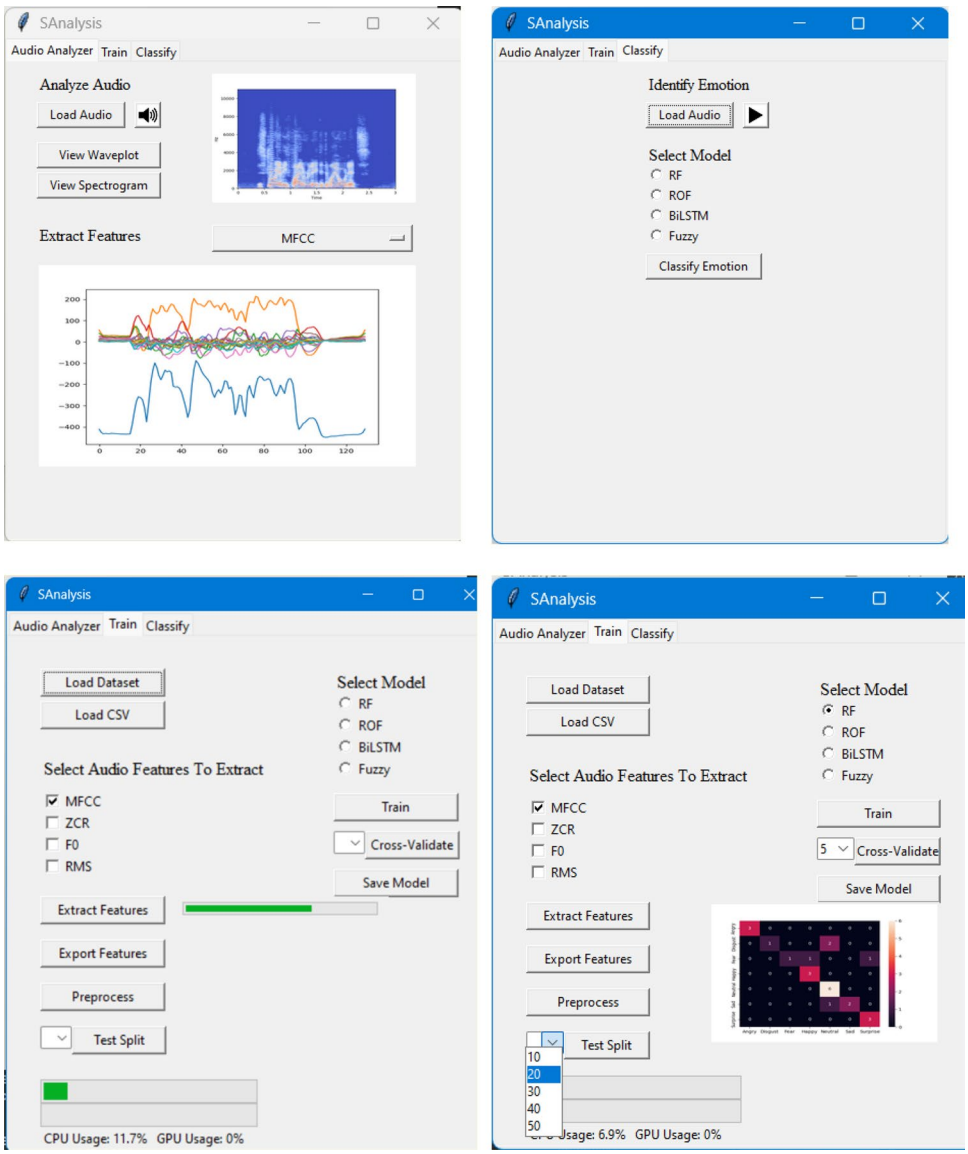
Conclusion

The field of emotion recognition from speech involves extracting features of the speech signal and analyzing it. Emotion recognition can help in various preliminary assessments involving healthcare, education, interrogations, etc. as well as contributes a lot to Human–Computer Interaction. In our study, four distinct audio features, namely Fundamental Frequency (F0), Zero-Crossing Rate (ZCR), Root Mean Square (RMS), and Mel-frequency

Table 17 Comparison with previous work

YOP	Author names	Database Used	Classifiers Used	Accuracy Achieved	Our experimentation
2022	Utkarsh Kumar Singh et al. [28]	IEMOCAP	RF	56.9%	76%
2022	Nasehatul Mustakim et al. [33]	Task-a Tamil Dataset	LSTM Bi-LSTM	35% 31%	67% 72%
2021	Dongdong Li et al. [13]	IEMOCAP	LSTM Bi-LSTM	59% 61%	67% 72%
2021	Rabia Qayyum, et al. [34]	FER2013	RNN	41%	48%
2021	Zixuan Peng et al. [12]	IEMOCAP	LSTM	65%	67%
2021	Kiran S Raj, Priyanka Kumar [29]	FER2013	Fuzzy Logic (Video) Fuzzy Logic (Text)	71% 69%	Fuzzy Logic (Audio): 47%

Fig.24 Diagnostic tool



cepstral coefficients (MFCC), were extracted from the audio data and utilized to train the machine learning and deep learning models. The seven emotions from the SAVEE dataset, and four emotions from the IEMOCAP dataset are considered. The models are trained and tested on the male and female samples independently, as well

as models trained using the SAVEE dataset are tested on IEMOCAP dataset and vice versa. The accuracies obtained by the models, namely, RF, Bi-LSTM, LSTM, Rotation Forest, Fuzzy, and RNN, on the SAVEE dataset are 76%, 72%, 67%, 66%, 38%, and 34%, respectively. On the IEMOCAP dataset, accuracies achieved on male samples are 68%, 64%, 61%, 53%, 48%, and 47% by RF, Bi-LSTM, LSTM, Rotation Forest, RNN, and Fuzzy, respectively. And the accuracies achieved on the female samples are 67%, 63%, 62%, 54%, 46%, and 42% by RF, Bi-LSTM, LSTM, Rotation Forest, RNN, and Fuzzy, respectively. The interactive GUI allows users to analyze audio, load audio datasets, extract and plot features, train machine learning and deep learning models, and classify emotion of an audio file.

Future work includes experimentation with different feature combinations that can lead to more robust representations which may result in better model performance. Combinations of audio with other modalities, such as text, videos, as well as real-time detection in live audio streams, can also be a potential field for exploration.

Data availability SAVEE dataset: http://kahlan.eps.surrey.ac.uk/savee/, IEMOCAP dataset: https://sail.usc.edu/iemocap/index.html.

Declaration

Conflict of interest The authors declare that there is no conflict of interest in this work.

References

1. Swain M, Routray A, Kabisatpathy P. Databases, features and classifiers for speech emotion recognition: a review. Int J Speech Technol. 2018;21(1):93–120.
2. Fayek HM, Lech M, Cavedon L. Evaluating deep learning architectures for speech emotion recognition. Neural Netw. 2017;92:60–8.
3. Abbaschian BJ, Sierra-Sosa D, Elmaghraby A. Deep learning techniques for speech emotion recognition, from databases to models. Sensors. 2021;21(4):1249.
4. Surrey Audio-visual expressed emotion (SAVEE) database. (n.d.). Retrieved November 15, 2022, from http://kahlan.eps.surrey.ac.uk/savee/.
5. IEMOCAP- home. (n.d.). Retrieved November 15, 2022, from https://sail.usc.edu/iemocap/.
6. Aouani H, Ayed YB. Speech emotion recognition with deep learning. Procedia Comput Sci. 2020;176:251–60.
7. Al Dujaili MJ, Ebrahimi-Moghadam A, Fatlawi A. Speech emotion recognition based on SVM and KNN classifications fusion. Intern J Electr Comput Eng. 2021;11(2):1259.
8. Zhao Z, Bao Z, Zhao Y, Zhang Z, Cummins N, Ren Z, Schuller B. Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition. IEEE Access. 2019;7:97515–25.
9. Sajjad M, Kwon S. Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. IEEE Access. 2020;8:79861–75.
10. Issa D, Demirci MF, Yazici A. Speech emotion recognition with deep convolutional neural networks. Biomed Signal Process Control. 2020;59:101894.
11. Zehra W, Javed AR, Jalil Z, Khan HU, Gadekallu TR. Cross corpus multi-lingual speech emotion recognition using ensemble learning. Complex Intell Syst. 2021. https://doi.org/10.1007/s40747-020-00250-4.
12. Peng Z, Lu Y, Pan S & Liu Y. Efficient speech emotion recognition using multi-scale cnn and attention. In ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 3020–3024. IEEE; 2021.
13. Li D, Liu J, Yang Z, Sun L, Wang Z. Speech emotion recognition using recurrent neural networks with directional self-attention. Expert Syst Appl. 2021;173:114683.
14. Kerkeni L, Serrestou Y, Mbarki M, Raoof K, Mahjoub MA & Cleder C. Automatic speech emotion recognition using machine learning. In Social media and machine learning. IntechOpen; 2019.
15. Aljuhani RH, Alshutayri A, Alahdal S. Arabic speech emotion recognition from saudi dialect corpus. IEEE Access. 2021;9:127081–5.
16. Rumagit RY, Alexander G, Saputra IF. Model comparison in speech emotion recognition for Indonesian language. Procedia Comput Sci. 2021;179:789–97.
17. Alnuaim AA, Zakariah M, Shukla PK, Alhadlaq A, Hatamleh WA, Tarazi H, Ratna R. Human-computer interaction for recognizing speech emotions using multilayer perceptron classifier. J Healthcare Eng. 2022;2022:1–12.
18. Alnuaim AA, Zakariah M, Alhadlaq A, Shashidhar C, Hatamleh WA, Tarazi H, Ratna R. Human-computer interaction with detection of speaker emotions using convolution neural networks. Comput Intell Neurosci. 2022;2022:1–16.
19. Atmaja BT, Sasou A, Akagi M. Speech emotion and naturalness recognitions with multitask and single-task learnings. IEEE Access. 2022;10:72381–7.
20. Rehman A, Liu ZT, Wu M, Cao WH & Jia CS. Real-time speech emotion recognition based on syllable-level feature extraction. arXiv preprint arXiv:2204.11382. 2022.
21. Aftab A, Morsali A, Ghaemmaghami S & Champagne B. LightSERNet: A lightweight fully convolutional neural network for speech emotion recognition. In ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6912–6916). IEEE 2022.
22. Huang Z, Dong M, Mao Q & Zhan Y. Speech emotion recognition using CNN. Proceedings of the 22nd ACM International Conference on Multimedia. 2014. https://doi.org/10.1145/2647868.2654984
23. Padi S, Sadjadi SO, Sriram RD & Manocha D. Improved speech emotion recognition using transfer learning and spectrogram augmentation. In Proceedings of the 2021 International Conference on Multimodal Interaction (pp. 645–652) 2021.
24. Jin C, Sherstneva AI & Botygin IA (n.d.). Speech emotion recognition based on deep residual convolutional neural network. Retrieved November 15, 2022, from https://journalpro.ru/articles/speech-emotion-recognition-based-on-deep-residual-convolutional-neural-network/
25. Kaur K, Singh P. Punjabi emotional speech database: design, recording and verification. Intern J Intell Syst Appl Eng. 2021;9(4):205–8.
26. Aggarwal A, Srivastava A, Agarwal A, Chahal N, Singh D, Alnuaim AA, Alhadlaq A, Lee HN. Two-way feature extraction for speech emotion recognition using deep learning. Sensors. 2022;22(6):2378. https://doi.org/10.3390/s22062378.

27. Attar HI, Kadole NK, Karanjekar OG, Nagarkar DR & Sujeet. Speech emotion recognition system using machine learning. Retrieved October 20, 2022, from https://ijrpr.com/uploads/V3ISSUE5/IJRPR4210.pdf
28. Kumar Singh U, Singh S, Khanna S, Shyam R. Speech emotion recognition using machine learning and deep learning. Intern J Eng Appl Sci Techno. 2022;6(11):181–4.
29. Raj KS & Kumar P. Automated human emotion recognition and analysis using machine learning. In 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1–9). IEEE 2021.
30. Rodriguez JJ, Kuncheva LI, Alonso CJ. Rotation forest: a new classifier ensemble method. IEEE Trans Pattern Anal Mach Intell. 2006;28(10):1619–30.
31. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD Explor Newsl. 2009;11(1):10–8.
32. Witten IH, Frank E, Hall MA, Pal CJ & DATA M. Practical machine learning tools and techniques. In Data Mining. 2005 2, 4.
33. Mustakim N, Rabu R, Mursalin GM, Hossain E, Sharif O & Hoque MM. CUET-NLP@ TamilNLP-ACL2022: Multi-class textual emotion detection from social media using transformer. In Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages (pp. 199–206). 2022.
34. Qayyum R, Akre V, Hafeez T, Khattak HA, Nawaz A, Ahmed S & ur Rahman K. Android based Emotion Detection Using Convolutions Neural Networks. In 2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) (pp. 360–365). IEEE 2021.