Multimedia Tools and Applications (2023) 82:42763–42781
https://doi.org/10.1007/s11042-023-15275-3

Speech emotion recognition using multimodal feature fusion with machine learning approach

Sandeep Kumar Panda[1] · Ajay Kumar Jena[2] · Mohit Ranjan Panda[2] · Susmita Panda[3]

Received: 17 May 2022 / Revised: 23 July 2022 / Accepted: 6 April 2023 /
Published online: 21 April 2023

Abstract

Speech-based emotional state recognition must have a significant impact on artificial intelligence as machine learning advances. When it comes to emotion recognition, proper feature selection is critical. As a result, feature fusion technology is offered in this work as a means of achieving high prediction accuracy by emphasizing the extraction of sole features. Mel Frequency Cepstral Coefficients (MFCC), Zero Crossing Rate (ZCR), Mel Spectrogram, Short-time Fourier transform (STFT) and Root Mean Square (RMS) are extracted, and four different feature fusion techniques are used on five standard machine learning classifiers: XGBoost, Support Vector Machine (SVM), Random Forest, Decision-Tree (D-Tree), and K Nearest Neighbor (KNN). The successful use of feature fusion techniques on our suggested classifier yields a satisfactory recognition rate of 99.64% on the female only dataset (TESS), 91% on SAVEE (male only dataset) and 86% on CREMA-D (both male and female) dataset. The proposed model shows that effective feature fusion improves the accuracy and applicability of emotion detection systems.

Keywords Feature fusion (FF) · Speech emotion recognition (SER) · Mel frequency cepstral coefficients · Zero Crossing Rate · Support vector machine · XGBoost

✉ Sandeep Kumar Panda
skpanda00007@gmail.com

Ajay Kumar Jena
ajay.bbs.in@gmail.com

Mohit Ranjan Panda
mohit.pandafcs@kiit.ac.in

Susmita Panda
susmitapanda@soa.ac.in

[1] Department of Data Science and Artificial Intelligence, Faculty of Science and Technology (IcfaiTech), ICFAI Foundation for Higher Education (Deemed to Be University), Hyderabad, Telangana, India

[2] School of Computer Engineering, KIIT Deemed to Be University, Bhubaneswar, Odisha, India

[3] Department of Computer Science and Engineering, SOA (Deemed to Be University), Bhubaneswar, Odisha, India

1 Introduction

1.1 Motivation and Incitement

Speech not only carries out the syntactical and semantic meaning of sentences but also expresses the state of mind of the speaker. Different emotions, such as shock, sadness, anger and happiness etc., are used to communicate effectively. Physically recognising these feelings is simple, but recognising them through a machine is challenging. Emotion recognition is a subset of speech recognition that is growing in popularity and demand. The act of attempting to recognise human emotion and affective states from speech is known as Speech Emotion Recognition (SER). It takes advantage of the fact that tone and pitch in the voice frequently reflect underlying emotion. This is also the phenomenon used by animals such as dogs and horses to comprehend human emotion. The use of SER to correctly identify people's mental state is beneficial in medical science, education, the entertainment industry, the automobile industry, security, and other industries. Global education is gradually shifting to an online format. However, recognising students' mental states during while studying may be difficult. The emotions of students can be determined using SER, which correctly indicates whether the learner is in a state of learning or not at that time [1, 2, 9]. SER is used in call centres to categorise calls based on emotions, and it can also be used as a performance parameter for conversational analysis, identifying dissatisfied customers, customer satisfaction, and other metrics to assist businesses in improving their services. It can also be used in-car board systems based on information provided by the driver's mental state to the system to initiate his/her safety, preventing accidents.

1.2 Contribution

Extraction and proper selection of features are needed due to variations of factors related to speech because the combination of significant features helps to get fruitful outcomes of SER [11, 19].

In our proposed work 3 English databases TESS (female only), SAVEE (male only) and CREMA-D (both male and female) have been used to predict the emotions. The audio samples in these datasets are then used to extract features from various feature extractors – MFCC, ZCR, Mel Spectrogram, STFT and RMS. After feature extraction, the extracted feature vectors are now fused step by step in different variations like only MFCC, only ZCR, MFCC + ZCR + Mel and MFCC + ZCR + Mel + STFT + RMS followed by different classifiers like Random Forest, Decision Tree, KNN, SVM and XGBoost to compare the accuracy from different feature fusion techniques.

1.3 Paper Organization

Further in this paper Sect. 2 discusses related works in this field, and Sect. 3 gives a brief overview of our proposed methodology, the experimented result analysis and comparative study are explained in Sect. 4, and finally, Sect. 5 describes the conclusion and future work of our proposed research work.

2 Literature Review

Previously, various research works on the SER system based on speech features and classification techniques to determine emotions were published. Following that approach, research on speech features is briefly reviewed in this section. We have also summarized various standard current research works in the overview chart which is expressed in Table 1.

The ability to recognise emotions using significant speech elements is extremely successful. Several researchers focus on different types of features such as prosodic and spectral features separately or in combination to get significant feature vectors for determining accurate emotions. Apart from feature selection, certain research is also targeted here according to classification aspects. Hidden Markov Model HMM is among the most efficient models for speech recognition and the SER system employed on its own and Mandarin datasets to recognise the six emotions anger, sadness, disgust, fear, surprise, and pleasure with an overall accuracy of 78 percent [17].

To predict emotions, some studies have utilized various types of classifiers and compared the accuracy levels of each classifier. The five emotions anger, happiness, sadness, surprise, and neutral are classified using SVM, GMM, KNN, HMM, and ANN classifiers. Among the above classifiers, for speaker dependent recognition, 89.12% overall accuracy is achieved by using GMM and 75% accuracy is achieved for both GMM and SVM for speaker independent recognition. The accuracy of 78.77% is also achieved for best features by using GMM [9].

Anger, happiness, neutrality, and sadness are classified using MFCC, energy, and formant [23]. Using the typical EMA dataset, different kernel functions of SVM are utilised to classify emotions.

In [14], Kuchibhotla S et al. presented a method in which SVM, KNN, LDA and Regularized Discriminate Analysis (RDA) were used on both the Berlin and Spanish datasets to categorize the common six emotions [12]. RDA obtained 92.6 percent accuracy on the Berlin dataset and 90.5 percent on the Spanish dataset using MFCC, pitch with energy.

MFCC and LPC features were used in this system to identify the emotions angry, joyful, and neutral on both the RAVDEES dataset and the self-created dataset ABEG. The three classes are classified using a variety of popular classifiers such as SVM, KNN, Adaboost, Logistic Regression, and XGboost. Amongst which, Logistic Regression on the ABEG dataset achieved 92 percent test accuracy, and XGboost on the RAVDEES + ABEG dataset achieved 86 percent test accuracy [5].

In another study, three popular classification algorithm SVM, LDA and D-tree were used to classify four emotions angry, happiness, sad, and neutral on RAVDEES database [12]. Taking the MFCC, DWT, pitch, energy and ZCR features from speech signal we achieved 85% highest accuracy for D-tree among the three specified classifier.

In [15], H. Kumbhar et al. presented a method that achieved 84.81% accuracy on RAVDEES dataset. In this work, a speech emotion recognition system with the LSTM model and MFCC feature was used.

Perceptual Linear prediction (PLP) is additionally added with popular prosodic(pitch, ZCR, energy)and spectral(MFCC, LPC, Formant) features to classify eight emotions anger, boredom, disgust, fear, neutral, sadness, surprise, happiness using LDA on two dataset Berlin and PDREC. The overall recognition rate of 55.74% and 47.28% was achieved on PDREC database for females and males, respectively and the average recognition rate of 78.64% and 73.40% was obtained for Berlin database for females and males, respectively. [7].

In [8], N. Ho et al. presented a multimodal approach for speech emotion recognition based on Multi-Level Multi-Head Fusion Attention mechanism and recurrent neural network (RNN).

Table 1 Overview Table for related works

Database	Research work on SER	Used features for Obtaining Highest Accuracy	Classifier used	Highest Accuracy (%)
Emo-DB	Kuchibhotla S et al. 2016 [14]	MFCC, pitch and energy	RDA	92.6
Spanish	Kuchibhotla S et al. 2016 [14]	MFCC, pitch and energy	RDA	90.5
RAVEDEES + ABEG	Chen L et al. 2012 [3]	MFCC, LPC	XGBoost	86
ABEG	Chen L et al. 2012 [3]	MFCC, LPC	Logistic Regression	92
RAVEDEES	Koduru A et al. 2020 [12]	MFCC, DWT, pitch, energy and ZCR	D-Tree	85
RAVEDEES	H. Kumbhar et al. 2019 [15]	MFCC	LSTM	84.81
MELD	N. Ho et al. 2020 [8]	MFCC (audio) + BERT (text)	RNN	63.26
CMU-MOSEI	N. Ho et al. 2020 [8]	MFCC (audio) + BERT (text)	RNN	99.19
PDEREC	Harimi A et al. 2014 [7]	pitch, ZCR, energy, MFCC, LPC, Formant and PLP	LDA	55.74(F) 47.28(M)
Emo-DB	Harimi A et al. 2014 [7]	pitch, ZCR, energy, MFCC, LPC, Formant and PLP	LDA	78.64(F) 73.40(M)

The proposed structure has inputs of two modalities: audio and text. For audio features, they determined the mel-frequenc cepstrum (MFCC) from raw signals using the OpenSMILE toolbox. Their experimental results on the three databases: Interactive Emotional Motion Capture (IEMOCAP), Multimodal EmotionLines Dataset (MELD), and CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI), reveal that the combination of the two modalities achieves better performance than using single models. They obtained an accuracy of 63.26% using multimodal on the MELD data and 48.84% and 61.66% for audio and text, respectively. For the CMU-MOSEI data, they achieved an accuracy of 99.19%.

Log-energy is another feature which is utilized to detect speech patterns and emotions. On the IITKGP-SESC and Berlin Database, many parameters such as LPCC, pitch, and energy are retrieved from the input audio signal for recognising eight emotions [13]. Auto associative neural networks, Gaussian mixture models, and support vector machines were used to develop emotion recognition systems with source, system, and prosodic features, respectively. From the results, it is observed that, each of the proposed speech features has contributed toward emotion recognition. The combination of features improved the emotion recognition performance, indicating the complementary nature of the features [20, 22, 26].

A three-level SER model was proposed to classify six speech emotions, including sadness, anger, surprise, fear, happiness, and disgust, in order to solve the speaker independent emotion recognition problem. Fisher rate, which is also used as an input parameter for Support Vector Machine, was used to select appropriate features from 288 candidates for each level (SVM). To evaluate the proposed system, four comparative experiments were designed using principal component analysis (PCA) for dimension reduction and artificial neural network (ANN) for classification, which include Fisher + SVM, PCA + SVM, Fisher + ANN, and PCA + ANN that achieved the average recognition rates for each level are 86.5%, 68.5% and 50.2% respectively. The experimental results demonstrated that Fisher outperforms PCA for dimension reduction and SVM outperforms ANN for speaker independent speech emotion recognition. [3].

After a thorough study, it has been discovered that there is no specific feature that contributes to the overall performance of the SER system. However, combining several features makes it easier to extract values for classification. Some researchers concentrate on feature selection, while others concentrate on classifier selection. However, in our opinion, neither adequate feature selection nor proper classifier selection is sufficient for developing a suitable SER system. It is far too difficult to construct a system in which an appropriate fusion of features with sufficient dimension is required for proper classification. The idea of feature fusion is integrated with a variety of machine learning (ML) classifiers to build a general framework that has been successfully applied to our popular datasets, TESS, SAVEE, and CREMA-D. After analyzing the valuable works, we have discovered another issue, some research works achieved a higher level of accuracy by classifying only a few common emotions and using only a portion of the dataset, whereas our primary objective is to classify all emotions and use the entire datasets to achieve a satisfactory level of accuracy as compared to the existing model by fusing the proper dimension of different features. It has been determined, for what dimensions of features should be fused in order to achieve high prediction accuracy.

3 Proposed Methodology

The schematic diagram of our proposed system is given in Fig. 1. The aim of a SER system is automatic emotional state determination from the audio samples. The emotional state of a person can be determined from various audio features which contain valuable information

from audio input signals. SER system followed by various phases including pattern recognition and matching.

The input audio signals go through the common phases of pre-processing, data augmentation and feature extraction. Data Augmentation is applied to increase the amount of data samples for better training. In feature extraction step, different audio features are extracted from the preprocessed data. After getting the individual features, the next step of feature fusion is used where the extracted feature vectors are fused step by step in different variations. This helps to select the best set of features which give the best prediction accuracy. The fused features are then trained using 5 different machine learning classifiers for predicting emotions.

3.1 Dataset and Technology Used

In our proposed work of SER, we have used three standard speech emotion datasets, which are.

3.1.1 CREMA-D—Crowd Sourced Emotional Multimodal Actors Dataset

It is an audio based acted dataset of 7442 original clips in WAV (Waveform Audio File) format. 91 actors (48 male and 43 female) ranging between the ages of 20 and 74 recorded these clips who came from a variety of races and ethnicities—African America, Asian, Caucasian, Hispanic, and Unspecified. Actors spoke from a selection of 12 sentences.

This dataset is used for classifying six basic emotions i.e., anger, disgust, fear, happy, neutral and sad. Most of the audio datasets use a limited number of speakers which may lead to leakage of information. For this fact CREMA-D is a very good dataset to use to make sure that the model does not overfit.

3.1.2 TESS—Toronto emotional speech set

It is an audio based acted dataset of 2800 audio files in WAV format. This dataset is female only. There is a set of 200 target words which were spoken in the carrier phrase "Say the word _" by two actresses aging between 26 and 64 years. This dataset is used for classifying seven basic emotions i.e., anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. Each actress has its own folder where she speaks the carrier phrase with the set of target words for each emotion. Most of the audio datasets are skewed towards male speakers and thus brings about a slight imbalance in representation. Hence, this dataset would work as a great training dataset for the emotion classifier in terms of generalization (not overfitting).

3.1.3 SAVEE—Surrey Audio-Visual Expressed Emotion

It is an audio based acted dataset of 480 audio files in WAV format. This dataset is male only. This dataset was recorded from four native English male speakers who were

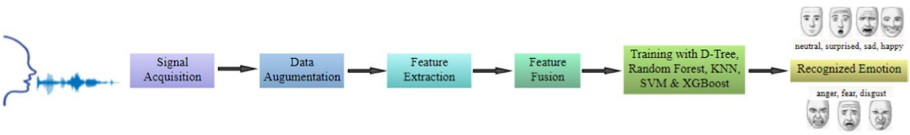

Fig. 1 Schematic diagram of the proposed speech emotion recognition system

postgraduate students and researchers at the University of Surrey, their ages ranged from 27 to 31 years.

This dataset is used for classifying seven basic emotions i.e., anger, disgust, fear, happiness, sadness and surprise and neutral. This dataset is male only, which causes a slight imbalance representation.

3.1.4 User Requirements—Hardware included 1.80 GHz CPU with 8 GB RAM

I used the library Librosa because all the features are stored as a built-in-model and also because of its ease of implementation, ease of use and ease of interoperability with other libraries.

3.2 Data Augmentation

Data augmentation (aug) is a popular strategy for increasing the quantity of training data, avoiding overfitting and improving model robustness. We can use noise injection, shifting time, time stretching, changing pitch and speed to generate syntactic data for audio. NumPy makes it simple to handle noise injection and audio range shifting, whereas librosa (library for Recognition and Organization of Speech and Audio) allows you to manipulate time stretching, pitch, and speed with just a single line of code.

In noise injection we just inject some random value into the data using NumPy. To shift audio range, we used the uniform function from numpy to uniformly distribute samples over the half-open interval [-5, 5) and then used the roll function to shift the audio range of the data over the given interval. We used librosa function of time stretch to change speed of the samples by speed factor of 1.25 to increase the speed and speed factor of 0.75 to lower the speed, we also used this function to stretch time of an audio series by a fixed rate, if rate is greater than 1, then the signal is sped up. If rate is less than 1, then the signal is slowed down. At last, we used the shift pitch librosa function to randomly change the pitch.

3.3 Feature Extraction

3.3.1 MFCC

Mel scale is a measurement unit of frequency perception of an audio signal in MFCC. In general, MFCC functions effectively in environments where the frequency range is limited. Because noise is less effective in this situation, it can be utilised to extract information from a variety of signals [18]. For the purposes of this study, 20 MFCC features are chosen. Some particular methods are taken in order to extract MFCC features from input audio signals which are Pre-emphasis, Framing, Hamming Window, Fast Fourier Transform (FFT), Mel- Frequency, Wrapping and Mel Cepstrum Coefficients [24].

3.3.2 ZCR

ZCR is determined from this signal changing rate. It is determined frame by frame. It is used to locate amplitude variations as well as the voice section of a speech stream. By comparing with unvoiced speech segment, it is determined that more ZCR have in fricative speech samples expressed in Eqs. 1

$$ZCRx(m) = \frac{1}{2N} \sum_{n=m-N-1}^{m} |\text{signx(n)} - \text{signx(n} - 1)| \quad (1)$$

3.3.3 Mel Spectrogram

The Mel scale is a scale of pitches that the listener perceives to be equal in distance from one another. For example, if the audio sources are in the same distance and atmosphere, a listener can tell the difference between 10000 Hz and 15000 Hz sounds. The Mel spectrogram is created by converting frequencies to the Mel scale. The Fourier transform can be used to convert frequencies to the Mel scale.

In Fig. 2 we can see the Mel spectrogram of a sample where the left side frequencies are in hertz and the right side has multiple Mel scale classes with colors, and how they change as the pitches change.

3.3.4 STFT (Short-time Fourier transform)

For time–frequency decomposition, STFT is commonly utilised in audio feature extraction. The short-time Fourier transform can be used to do time–frequency analysis (STFT). It is employed in the creation of representations that capture the signal's local time and frequency information. The STFT, like the Fourier transform, requires fixed-size time-shifted window functions w(n), which are expressed as Eq. 2

$$X(k, m) = \sum_{n=0}^{N-1} x(n + m)w(n)W_N^{nk}; \text{k, m} = 0, 1, 2, \ldots \ldots, \text{N} - 1 \quad (2)$$

where m is the amount of shift.

3.3.5 RMS (Root Mean Square)

The Root Mean Square (RMS) is modelled as an amplitude modulated Gaussian random process, with the RMS being related to the constant force and non-fatiguing muscle contractions. Feature extraction using the RMS approach is fairly popular because of it is computationally efficiency and speed while retaining important data.

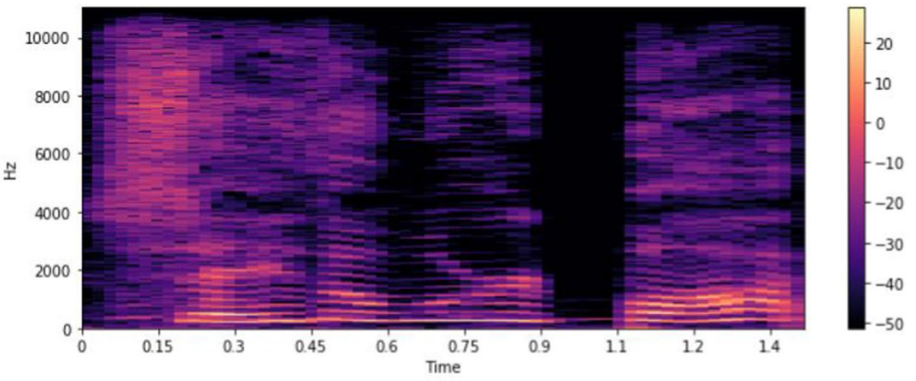

Fig. 2 Mel Spectrogram of a sample

$$RMS = \sqrt{\frac{1}{N}} \sum_{i-1}^{N} |x_i|^2 \tag{3}$$

3.4 Feature Fusion

In our proposed work 20 orders of MFCC, single ZCR value, one STFT value, single RMS value and single Mel Spectrogram feature are computed for each of the frame. Then the mean values of these features are taken for each of the audio samples. These features are fused together in several phases. Finally total 24 features are taken here. Another category of fusing features is selected by taking MFCC, ZCR and Mel Spectrogram together. Beside this only MFCC and only ZCR is taken as features individually. All of these categorical combinations are taken for comparing the prediction accuracy with the help of various classifiers which are used in this proposed work.

The pseudo code for our proposed methodology is given below.

```
load_signal(PATH)

Augmentation_1 <---noise(signal)
Augmentation_2 <---pitch(signal)
Augmentation_3 <---stretch(signal)
Augmentation_4 <---shift(signal)
Augmentation_5 <---lower_speed(signal)
Augmentation_6 <---higher_speed(signal)

Feature_1 <---MFCC(signal)
Feature_2 <---ZCR(signal)
Feature_3 <---STFT(signal)
Feature_4 <---RMS(signal)
Feature_5 <---Mel(signal)

Fused_result_1 <--- Fusion (MFCC)
Fused_result_1 <--- Fusion (ZCR)
Fused_result_1 <--- Fusion (MFCC,ZCR, Mel)
Fused_result_1 <--- Fusion (MFCC, ZCR, Mel, RMS, STFT)

XGBoost_model = Train_XGBoost(fused_result)
SVM_model = Train_SVM(fused_result)
D-Tree_model = Train_D-Tree(fused_result)
RandomForest_model = Train_RandomForest(fused_result)
KNN_model = Train_KNN(fused_result)
```

3.5 Classifier

In terms of machine learning, the model receives the trained data and classifies it with the test dataset using a specific algorithm. SVM, XGBoost, KNN, Random Forest, and Decision tree (D-tree) are five popular classifiers that have been used in this study. SVM iteratively creates hyperplane in order to minimize error. These kernel functions are required for SVM to work. Kernel functions are commonly employed to map original features to higher-dimensional space nonlinearly. Gradient boosting is a strategy for reducing errors that uses the gradient descent technique. XGBoost (Extreme Gradient Boosting) is a unique combination of software and hardware optimization techniques that generates excellent results in the shortest amount of time while using the fewest computer resources. D-tree uses the concept of analysing all possible decision outcomes and tracing each path to the conclusion. Random Forest is used to take the average of outcomes of several decision trees applied to distinct subsets of a dataset to improve its predicted accuracy. The random forest gathers estimate from all trees and forecasts the final output based on the majority vote of predictions, instead of depending on a single decision tree. The KNN algorithm takes data and uses a distance function to classify fresh data points. The KNN approaches are based on not only single closest neighbour classification but also unfamiliar sample classification based on K nearest neighbour votes. The constant k is defined here by the user. This approach uses the Euclidian distance function to find the k neighbours who are closest to the unlabeled sample from the training.

4 Experimental result

Several factors influence overall SER system performance. Some of these important aspects are used in our proposed work, such as data augmentation, audio sample quality, extracted features and classification algorithms. Three popular datasets, TESS, SAVEE, and CREMA-D, are used in our research to compare overall performance with previous work. Four types of feature combinations and six forms of data augmentation are applied to develop different feature fusion techniques which are then trained with five different classifiers and the prediction level of accuracy is analyzed to justify the efficiency of the suggested technique in the SER system. Table 1 summarizes this combination of features (Table 2).

4.1 Dataset wise experimented results

Here the result of our proposed method is presented in details with respect to the used dataset of this proposed work. For each of the dataset the bar chart shows the accuracy of the different Feature fusion techniques.

Table 2 Combination of features used in SER

Feature Fusion Techniques	Feature Combination
Feature Fusion 1 (FF1)	MFCC
Feature Fusion 2 (FF2)	ZCR
Feature Fusion 3 (FF3)	MFCC + ZCR + Mel
Feature Fusion 4 (FF4)	MFCC + ZCR + Mel + STFT + RMS
Augmentation-	Noise + Pitch + Shift + Stretch + higher speed + lower speed.

4.1.1 TESS

The different combinations of features are extracted from 2800 audio samples with sampling rate 24.414 kHz in TESS dataset. The emotions which are taken for classification from this dataset are abbreviated like anger as Ang, disgust as Disg, fear as Fear, happiness as Happy, pleasant surprise as PS, sadness as Sad and neutral as Neu. The comparative accuracy obtained for each of the different fusion of features is shown in Fig. 3 using the XG Boost classifier.

From our proposed technique it is observed that, for the TESS dataset, FF-3 is sufficient to recognize different emotions by using the XG Boost classifier. The XG Boost classifier gives more steady result as compared to another classifier for the TESS dataset. The evaluation metric for the highest accuracy obtained for TESS dataset is given in Table 3.

A confusion matrix is mainly a technique for summarising a classification algorithm's performance. The Table 4 is a Confusion matrix of highest accuracy obtained for TESS dataset which is using XGBoost model and Feature fusion-4. So, from the table we can understand that there are 468 audio samples for the emotion anger and our model has classified 467 out of 468 audio samples correctly as anger and 1 as disgusted and hence we get out precision as 0.99 using Eqs. 4 and similarly, we have presented these values for all the emotions in the confusion matrix.

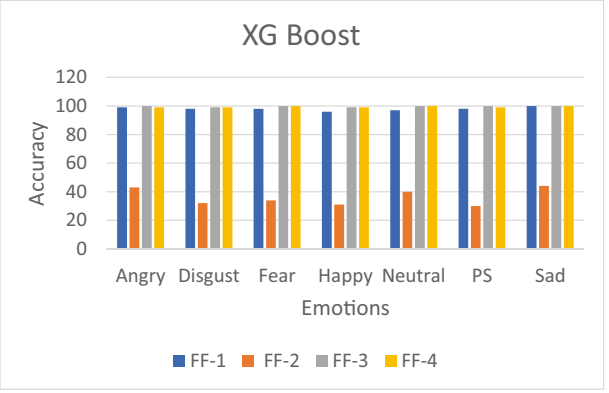

Fig. 3 Accuracy using XGBoost classifier of TESS dataset

Table 3 Evaluation metric

Class	precision	recall	F1-score	support
Angry	0.99	1.00	0.99	468
Disgust	0.99	1.00	0.99	482
Fear	1.00	1.00	1.00	490
Happy	0.99	1.00	0.99	459
Neutral	1.00	1.00	1.00	480
PS	0.99	0.99	0.99	467
Sad	1.00	1.00	1.00	514

Table 4 Confusion matrix of highest accuracy obtained for TESS dataset

	Ang	Disg	Fear	Happy	Neu	PS	Sad
Ang	467	1	0	0	0	0	0
Disg	0	481	0	0	0	1	0
Fear	2	0	488	0	0	0	0
Happy	1	0	0	457	0	1	0
Neu	0	0	0	0	479	1	0
PS	1	3	0	3	0	460	0
Sad	0	2	0	0	0	0	512

$$\text{Precision} = \frac{True\ positive}{True\ Positive\ +\ falsepositive} \tag{4}$$

4.1.2 SAVEEE

The different combinations of features are extracted from 480 audio samples with sampling rate 44.1 kHz in SAVEE dataset. The emotions which are taken for classification from this dataset are abbreviated like anger as Angry, disgust as Disgust, fear as Fear, happiness as Happy, pleasant surprise as Surprise, sadness as Sad and neutral as Neutral. The comparative accuracy obtained for each of the different fusion of features is shown in Fig. 4 using the XG Boost classifier.

From our proposed technique it is observed that, for the SAVEE dataset FF-3 is sufficient to recognize different emotions by using the XG Boost classifier. The XG Boost classifier gives more steady result as compared to another classifier for the SAVEE dataset. The confusion matrix for the highest accuracy obtained for SAVEE dataset is given in Table 5 and evaluation metric is given in Table 6.

4.1.3 CREMA-D

The different combinations of features are extracted from 7442 audio samples with sampling rate 48 kHz in CREMA-D dataset. The emotions which are taken for classification from this dataset are abbreviated like anger as Angry, disgust as Disgust, fear as Fear, happiness as

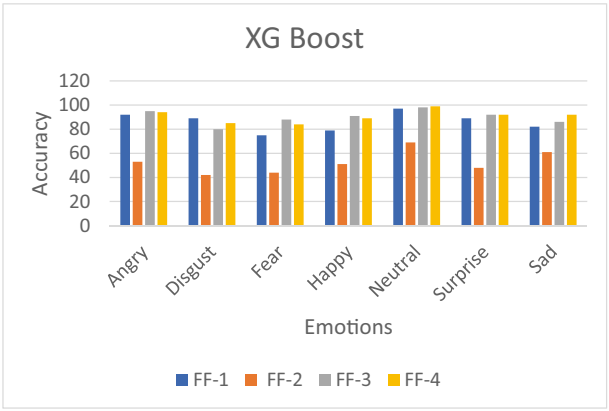

Fig. 4 Accuracy using XGBoost classifier of SAVEE dataset

Table 5 Confusion matrix of highest accuracy obtained for SAVEE dataset

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	58	0	0	2	0	0	1
Disgust	1	67	3	2	0	0	0
Fear	0	2	71	1	0	0	4
Happy	0	0	4	72	0	0	2
Neutral	1	8	2	0	130	5	0
Sad	0	2	1	0	1	61	0
Surprise	2	0	4	4	0	0	66

Table 6 Evaluation metric

Class	precision	recall	F1-score	support
Angry	0.94	0.97	0.95	60
Disgust	0.85	0.92	0.88	73
Fear	0.84	0.91	0.87	78
Happy	0.89	0.92	0.91	78
Neutral	0.99	0.89	0.94	146
PS	0.92	0.87	0.89	65
Sad	0.92	0.94	0.93	76

Happy, sadness as Sad and neutral as Neutral. The comparative accuracy obtained for each of the different fusion of features is shown in Fig. 5 using the XG Boost classifier.

From our proposed technique it is observed that, for the CREMA-D dataset FF-3 is sufficient to recognize different emotions by using the XG Boost classifier. The XG Boost classifier gives more steady result as compared to another classifier for the CREMA-D dataset. The confusion matrix for the highest accuracy obtained for CREMA-D dataset is given in Table 7 and evaluation metric is given in Table 8.

4.2 Result Analysis

Following the above-mentioned confusion matrix of various datasets with regard to classifiers and Feature fusion techniques it has been discovered that different success rates are attained for each of the four Feature fusion techniques in Table 9. The table

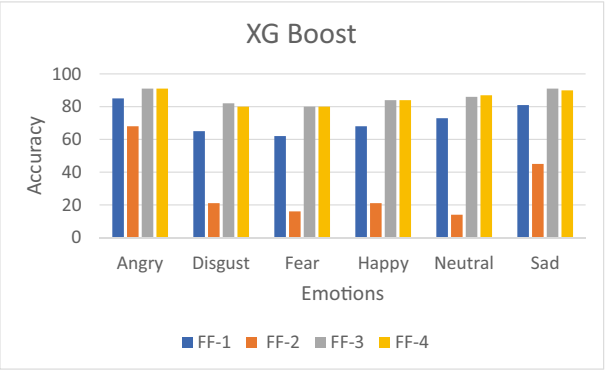

Fig. 5 Confusion matrix of highest accuracy obtained for Crema-D dataset

Table 7 Confusion matrix of highest accuracy obtained for Crema-D dataset

	Ang	Disg	Fear	Happy	Neu	Sad
Ang	1384	37	36	58	3	0
Disg	27	1231	39	53	61	41
Fear	15	39	1184	47	118	36
Happy	66	55	58	1265	37	14
Neu	18	76	47	56	1150	65
Sad	7	100	114	31	55	1408

below provides a clear picture of each classifier's recognition accuracy. In this case, classifiers are utilized to determine prediction accuracy of different Feature fusion techniques. Figures 6, 7 and 8 given below shows the accuracy of all the three datasets, TESS, SAVEE and CREMA-D using different classifiers, respectively. Our main goal is to determine a higher success rate using the appropriate model in this SER system. The data is classified into different emotions like angry, sad, happy, surprise, neutral and fear. These different classes are easily separable; hence the traditional machine learning algorithms perform well. We also used data augmentation using which we can increase the number of samples for training which helps in increasing the accuracy. Though the emphasis was not primarily on recognition accuracy. Apart from sustaining accurate thinking, the focus is on a large range of emotional considerations.

So many higher levels of accuracy are maintained in different research work by considering the minimum number of classes. Using feature fusion approaches, we aim to classify as many emotions as possible for the databases listed above. In prior studies, higher accuracy levels were reached by using fewer number of emotions, such as three to five emotions. However, in order to cover the greatest number of classes with the highest identification rate, the standard level of accuracy is attained, as shown in Table 8. The performance level is now assessed using both the dataset and the classifier: in TESS, for Random Forest 97.5%, KNN 98.57%, SVM 98.27% accuracy level is achieved by using FF-4 whereas for D-Tree 95% and XGBoost 99.64% accuracy is achieved using FF-3. In SAVEE for Random Forest80%, D-Tree85% and KNN 85% accuracy is obtained using FF-4 but for SVM 83% and XGBoost 91% accuracy is obtained using FF-3. In CREMA-D for Random Forest84%, D-Tree77% and KNN 50% accuracy is obtained using FF-4 but for SVM 45% and XGBoost 86% accuracy is obtained using FF-3.

Table 8 Evaluation metric

Class	precision	recall	F1-score	support
Angry	0.91	0.91	0.91	1518
Disgust	0.80	0.85	0.82	1452
Fear	0.80	0.88	0.84	1339
Happy	0.84	0.85	0.84	1495
Neutral	0.87	0.81	0.84	1412
Sad	0.90	0.82	0.86	1715

4.3 Comparative Study

Following the acquisition of various levels of prediction accuracy, our proposed work is compared to various standard current research work based on the database. The comparison chart is expressed in Table 9. The standard accuracy level is already attained by Choudhury et al. 2018 [5] and Slimi et al. 2020 [27] in the TESS dataset, but in our suggested work with the help of XGBoost, 99.64 percent prediction accuracy is accomplished in this dataset by employing FF-3. In the SAVEE dataset, Liu et al. 2020[16], achieved 75% accuracy, S. Kanwal et al. 2021[10], achieved 77.7% accuracy. Whereas, in our proposed study accuracy of 91% is reached by employing FF-3and taking into account all seven emotions. For the CREMA-D dataset E. Ghalebet al. 2020[6], achieved 66.5% prediction accuracy using Multimodal Emotion Recognition Metric Learning (MERML) on all 6 emotions, R. Pappagari et al. 2020[21], achieved 81.54% prediction accuracy using Fine-Tuned ResNet by concidedering only four emotions. Whereas in our proposed study accuracy of 84% is achieved using FF-3and taking account all six emotions.

Table 9 Model wise classification accuracy

Dataset	Emotions	Classifier	Classification Accuracy %			
			FF-1	FF-2	FF-3	FF-4
TESS	Angry, Disgust, Fear, Happy, PS, Neutral, Sad	Random Forest	0.93	0.56	0.97	**0.975**
		D-Tree	0.91	0.44	**0.96**	0.95
		KNN	0.97	0.43	0.97	**0.9857**
		SVM	0.91	0.21	0.9813	**0.9827**
		XG Boost	0.98	0.36	**0.9964**	0.9955
SAVEE	Angry, Disgust, Fear, Happy, Surprise, Neutral, Sad	Random Forest	0.73	0.61	0.80	**0.80**
		D-Tree	0.82	0.57	0.84	**0.85**
		KNN	0.84	0.51	0.83	**0.85**
		SVM	0.62	0.23	**0.83**	0.82
		XG Boost	0.87	0.54	**0.91**	0.90
CREMA-D	Angry, Disgust, Fear, Happy, Neutral, Sad	Random Forest	0.69	0.55	0.73	**0.84**
		D-Tree	0.73	0.33	0.77	**0.77**
		KNN	0.29	0.08	0.29	**0.5**
		SVM	0.40	0.2	**0.45**	0.42
		XG Boost	0.72	0.31	**0.86**	0.85

According to the different classifiers bold numbers shows the best results in comparison to classification accuracy

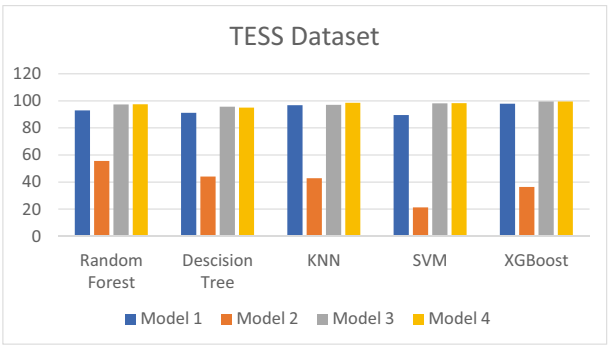

Fig. 6 Accuracy of Tess dataset using different classifiers

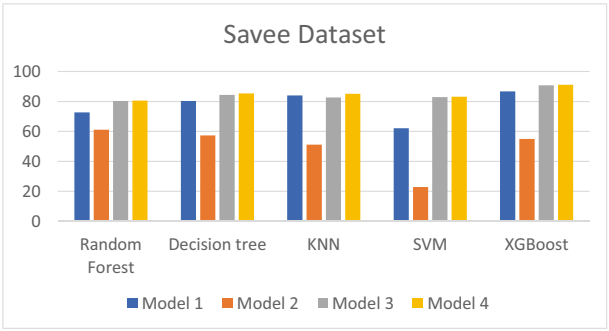

Fig. 7 Accuracy of SAVEE dataset using different classifiers

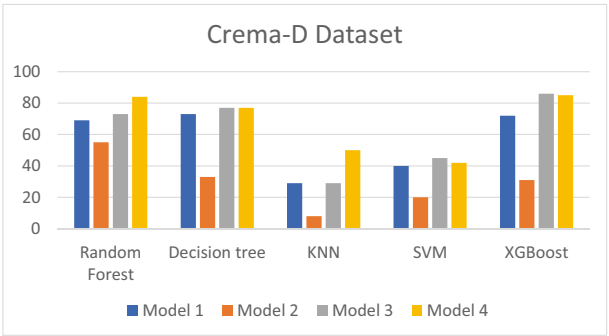

Fig. 8 Accuracy of CREMA-D dataset using different classifiers

5 Conclusion

One of the main objectives of SER research is to attain a high level of success with a specific feature set and a suitable classifier. Because the sensitivity of emotions varies by gender, features must be correctly structured to achieve a greater level of accuracy. Redundant data must be deleted using a solid feature selection method to speed up a classifier's learning process. We use both prosodic and spectrum characteristics in our research, and their combination leads in feature fusion. When fused features are compared to prosodic and spectrum features, it is discovered that feature fusion produces better results in the vast majority of cases. In our proposed work accuracy is measured depending on feature combination of MFCC, ZCR, STFT, Mel Spectogram and RMS features performance are applied on SVM, XGBoost, KNN, D-Tree and Random Forest classifiers for TESS, SAVEE, CREMA-D datasets. The results of the proposed models are also compared to those of other techniques used by researchers on the same datasets. The success rate is achieved 99.64% on TESS dataset for all feature combination on FF-3 and 99.55% on FF-4 with XGBoost. 90% accuracy is achieved for the dataset SAVEE for all feature combinations using XGBoost. For CREMA-D 86% accuracy is obtained using XGBoost. This shows that XGboost gives better performance as compared to other classifiers. As a result, XGBoost can be considered as a good classification approach and the fusion of MFCC, ZCR, and

Table 10 Comparative analysis with respect to accuracy

Database	Research work on SER	Used features for Obtaining Highest Accuracy	Classifier used	Highest Accuracy (%)
TESS	Choudhury et al. 2018 [4]	RMFCC, Epoch Location, Strength of Epoch, Slope of Strength of Epoch, Energy, Entropy of Energy, Zero-Crossing Rate, Spectral Centroid, Spectral Spread, Spectral Entropy, Spectral Flux, Spectral Rolloff	Sequential Minimal Optimization (SMO)	99.1
	Slimi et al. 2020 [25]	Spectrogram	Shallow Neural Network	99.58
	Proposed methodology	MFCC + ZCR + Mel + STFT + RMS Aug: Noise + Pitch + Shift + Stretch + higher speed + lower speed	XGBoost	99.64
SAVEE	Liu et al. 2020 [16]	F0 + MFCC + ZCR + RMS + Harmonic Noise Ratio	SVM	75
	S. Kanwal et al. 2021 [10]	INTERSPEECH 2010 feature set	SVM	77.7
	Proposed methodology	MFCC + ZCR + Mel + STFT + RMS + Aug	XGBoost	91
CREMA-D	E. Ghalebet al. 2020 [6]	Energy, Spectral and voice LLDs	MERML	66.5
	R. Pappagari et al. 2020 [21]	MFCC, aug- Noise	Fine-Tuned ResNet	81.54
	Proposed methodology	MFCC + ZCR + Mel + STFT + RMS + Aug	XGBoost	86

Mel Spectrogram features in the domain of speaker recognition for a clean speech database should be considered as a great strategy.

Human emotion is expressed not only through voice but also through other physical gestures such as facial expression or movement of body parts. As a result, emotion related to speech is frequently ambiguous due to the nature of a person. Thus, emotion recognition using machine intelligence faces numerous challenges and has a long way to go. Combinations of the given methods can be derived to improve the emotion recognition process. Also, the accuracy of the speech emotion recognition system can be improved by extracting more effective features of speech.

Furthermore, as shown in Table 10, the accuracy obtained for the tess dataset is higher than that obtained for the Crema-D and SAVEE datasets. TESS is a female only dataset whereas SAVEE is a male only dataset and crema-d has both male and female audio clips. This shows that it is easier to identify a female's emotions are they emotionally expressive and their voice has higher pitch and frequency. There is also a fairly substantial body of research demonstrating that women are the more emotionally expressive gender. This is the reason that TESS has higher accuracy. So, training a model that can handle variability in terms of expressiveness would be helpful. To achieve this, one can use CNNs as the classifier. This is because CNNs are known to be invariant to a wide range of variability and this can be beneficial in our case.

Authors' contributions The authors' contributions are summarized below. Sandeep Kumar Panda made substantial contributions to the conception and design and were involved in drafting the manuscript. Ajay Kumar Jena and Mohit Ranjan Panda acquired data and analysis and conducted the interpretation of the data. The critically important intellectual contents of this manuscript were revised by Susmita Panda. All authors read and approved the final manuscript.

Data Availability The data used to support the findings of this study are available from the corresponding author upon request.

Code availability Not applicable.

Declarations

Conflicts of interest/Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Saikat B, Jaybrata C, Arnab B et al (2017) A review on emotion recognition using speech. Int Conf Inventive Commun Comput Technol (ICICCT) (2017):109–114
2. Chavhan Y, Dhore M, Yesaware P (2010) Speech emotion recognition using support vector machine. Int J Comput Appl 1:6–9
3. Chen L, Mao X, Xue Y et al (2012) Speech emotion recognition: Features and classification models. Digit Signal Process 22:1154–1160
4. Akash RC, Anik G, Rahul P, et al. (2018) Emotion Recognition from Speech Signals using Excitation Source and Spectral Features. IEEE Applied Signal Proc (ASPCON) 257–261
5. Prashengit D, Sunanda G (2021) A System to Predict Emotion from Bengali Speech. Int J Math Sci Comput. https://doi.org/10.5815/IJMSC.2021.01.04
6. Ghaleb E, Popa M, Asteriadis S (2020) Metric Learning-Based Multimodal Audio-Visual Emotion Recognition. IEEE Multi Med 27(1):37–48. https://doi.org/10.1109/MMUL.2019.2960219

7. Harimi A, Esmaileyan Z (2014) A database for automatic Persian speech emotion recognition: collection, processing and evaluation. Int J Eng 27:79–90
8. Ngoc-Huynh H, Hyung-Jeong Y, Soo-Hyung K et al (2020) Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention-Based Recurrent Neural Network. IEEE Access 8:61672-61686
9. Ingale AB, Chaudhari D (2012) Speech emotion recognition. Int J Soft Comput Eng (IJSCE) 2:235–238
10. Kanwal S, Asghar S (2021) Speech Emotion Recognition Using Clustering Based GA- Optimized Feature Set. IEEE Access 9:125830–125842. https://doi.org/10.1109/ACCESS.2021.3111659
11. Ko T, Peddinti V, Povey D, Khudanpur S (2015) Audio Augmentation for Speech Recognition
12. Koduru A, Valiveti HB, Budati AK (2020) Feature extraction algorithms to improve the speech emotion recognition rate. Int J Speech Technol 23:45–55
13. Koolagudi SG, Rao KS (2012) Emotion recognition from speech using source, system, and prosodic features. Int J Speech Technol 15:265–289
14. Kuchibhotla S, Vankayalapati HD, Anne KR (2016) An optimal two stage feature selection for speech emotion recognition using acoustic features. Int J Speech Technol 19:657–667
15. Kumbhar H, Bhandari S (2019) Speech Emotion Recognition using MFCC features and LSTM network, IEEE International Conference On Computing, Communication, Control And Automation, pp. 1–3
16 Liu Zhen-Tao, Bao-Han Wu, Li Dan-Yun, Xiao Peng, Mao Jun-Wei (2020) Speech Emotion Recognition Based on Selective Interpolation Synthetic Minority Over-Sampling Technique in Small Sample Environment. Sens 20(8):2297
17. Nwe TL, Foo SW, De Silva LC (2003) Speech emotion recognition using hidden Markov models. Speech Commun 41:603–623
18. Ooi CS, Seng KP, Ang L-M et al (2014) A new approach of audio emotion recognition. Expert Syst Appl 41:5858–5869
19. Palo HK, Mohanty MN (2018) Comparative analysis of neural networks for speech emotion recognition. Int J Eng Technol 7:111–126
20. Yixiong P, Peipei S, Liping S et al (2012) Speech Emotion Recognition Using Support Vector Machine.
21. Pappagari, R. et al (2020) X-Vectors Meet Emotions: A Study On Dependencies Between Emotion and Speaker Recognition. Int Conf Acoust Speech Signal Process (ICASSP) 7169–7173
22. Rao KS, Kumar TP, Anusha K et al (2012) Emotion recognition from speech. Int J Comput Sci Inf Technol 3:3603–3607
23. Shah RD, Anil D, Suthar C (2016) Speech emotion recognition based on SVM using MATLAB. Int J Innov Res Comput Commun Eng 4
24. Shambhavi S, Nitnaware V (2015) Emotion speech recognition using MFCC and SVM. Int J Eng Res Technol 4:1067–1070
25. Anwer S, Mohamed H, Mounir Z et al. Emotion recognition from speech using spectrograms and shallow neural networks. Proceedings of the 18th International Conference on Advances in Mobile Computing & Multimedia (2020): n. pag.
26. Wang K, An N, Li BN et al (2015) Speech emotion recognition using Fourier parameters. IEEE Trans Affect Comput 6:69–75
27. Yang B, Lugger M (2010) Emotion recognition from speech signals using new harmony features. Signal Process 90:1415–1423