I.J. Information Technology and Computer Science, **2026, 1, 84-99**
Published Online on February 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2026.01.05

Ensemble Fusion Model for Enhanced Speech Emotion Recognition and Confusion Resolution

Rania Ahmed
Computer Science Department, Faculty of Computers and Information, Menoufia University, Shebin Elkom 32511, Egypt
E-mail: Rania_anwer@hotmail.com
ORCID iD: https://orcid.org/0009-0009-3450-8674
*Corresponding author

Mahmoud Hussein
Computer Science Department, Faculty of Computers and Information, Menoufia University, Shebin Elkom 32511, Egypt
E-mail: mahmoud.hussein@ci.menofia.edu.eg
Department of Computer Science, Faculty of Information Technologies, Philadelphia University, Amman 19392, Jordan
E-mail: mhussein@philadelphia.edu.jo
ORCID iD: https://orcid.org/0000-0002-3742-7548

Arabi keshk
Computer Science Department, Faculty of Computers and Information, Menoufia University, Shebin Elkom 32511, Egypt
E-mail: arabikeshk@yahoo.com
ORCID iD: https://orcid.org/0000-0000-0002-8389-7989

Received: 16 July 2025; Revised: 28 August 2025; Accepted: 19 October 2025; Published: 08 February 2026

Abstract: In the field of human-computer interaction, identifying emotion from speech and understanding the full context of spoken communication is a challenging task due to the imprecise nature of emotion, which requires detailed speech analysis. In the area of speech emotion recognition, various techniques have been employed to extract emotions from audio signals, including several well-established speech analysis and classification methods. Despite numerous advancements in recent years, many studies still fail to consider the semantic information present in speech. Our study proposes a novel approach that captures both the paralinguistic and semantic aspects of the speech signal by combining state-of-the-art machine learning techniques with carefully crafted feature extraction strategies. We address this task using feature-engineering-based techniques, which involve extracting meaningful audio features such as energy, pitch, harmonics, pauses, central momentum, chroma, zero-crossing rate, and Mel-frequency cepstral coefficients (MFCCs). These features capture important acoustic patterns that help the model learn emotional cues more effectively. This work is primarily conducted on the IEMOCAP dataset, a large and well-annotated emotional speech corpus. By framing our task as a multi-class classification problem, we extract 15 features from the audio signal and use them to train five machine learning classifiers. Additionally, we incorporate text-domain features to reduce ambiguity in emotional interpretation. We evaluate our model's performance using accuracy, precision, recall, and F-score across all experiments.

Index Terms: Speech Emotion Recognition, Machine Learning, Multimodal, Ensemble Model.

1. Introduction

Human emotions can be automatically recognized through various modalities, including Physiological Signals, facial expressions, voice, semantic and syntactic elements of text messages, and body movements. However, in scenarios where visual features are absent, such as in voice messages or contact center applications, speech becomes the only method for recognizing emotions. Speech Emotion Recognition (SER) becomes crucial in these contexts. SER [1], which involves analyzing the acoustic features of speech to determine the speaker's emotional state, can recognize emotions in two common models, Discrete emotion theory which is based on the six categories of basic emotions: sadness, happiness, fear, anger, surprise, and neutral, or Dimensional emotional model which is an alternative model that uses a small number of latent dimensions to characterize emotions such as valence, arousal, control, and power. Recently, a set of machine learning approaches has been used to predict emotion in speech. These approaches can be classified into two categories: Traditional Machine Learning Models and Ensemble Learning Models [2].

For classification tasks, traditional machine learning models rely on manually extracted features such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch, and energy. Here, only individual classifiers have been used, such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Hidden Markov Model (HMM), Artificial Neural Network (ANN), and Random Forest (RF). We concentrate on Ensemble Learning Models, which combine several base learners to get better prediction accuracy. This decision is consistent with our approach, which focuses on enhanced generalization and model robustness. Achieving high accuracy in classification requires a combination of proper data handling, model selection, and optimization techniques. In terms of data handling, the dataset should be large enough to represent all classes and variations (e.g., different speakers, accents, and noise conditions). Data augmentation techniques can be used to increase the size of the dataset by adding variations (e.g., adding noise, pitch shifting, or time-stretching in speech data), To prevent biased learning, the dataset should also be balanced to ensure that all classes are equally represented, techniques like Synthetic Minority Over-sampling (SMOTE) and random oversampling can be used to balance the dataset. Furthermore, data preprocessing, such as Feature Scaling, normalization, and standardization, is a critical step in building accurate and effective machine learning models, which means transforming the data to make it more suitable for modeling. One of the common normalization techniques is Min-Max Normalization. These techniques can assist in improving model performance, reducing the impact of outliers, and ensuring that the data is on the same scale. Finally, High dimensionality in a dataset presents a challenge since it introduces many features, which can complicate the classification process and make it harder to identify relevant data for analysis. Typically, not all features contribute meaningfully to the classifier's performance; some may be irrelevant or redundant. To improve efficiency, reduce complexity, and manage dimensionality, it is crucial to eliminate features that are unnecessary or have little impact on classification performance. Regarding model selection, results from various studies indicate that prediction accuracy can be enhanced by using an ensemble learning model, especially when combined with multiple modalities (also known as a multi-modal approach), which involves integrating and analyzing data from different sources, such as text, audio, or visual inputs, to provide a more comprehensive understanding and improve model predictions. This improvement is further supported by proper data pre-processing. Furthermore, some attention-based techniques are used as optimization techniques to improve the performance of a model. Despite all previous advancements, the primary limitation in previous SER systems is either a lack of data handling or their reliance on a single modality, particularly acoustic features, which fail to capture the full spectrum of emotions conveyed in speech, while others achieved high accuracy but predicted only a few categories of emotions. Additionally, while some studies have explored combining multiple modalities, they often do not effectively leverage the collaboration between text and audio features. As a result, current algorithms still struggle with accurately detecting emotions across diverse contexts.

In this paper, we propose a method that combines both text and audio features to improve the performance of SER systems. This approach is based on the idea that emotions are conveyed not only through the acoustic properties of speech but also through the meaning of spoken words. We employed carefully chosen features from speech segments, including parameters in time, frequency, amplitude, and spectral distribution domains. For the textual modality, we utilized word embeddings from speech transcriptions using the Bag of Words for Word representation (BoW) [3]. The BoW approach was selected for its ability to capture frequency-based word representations, aligning with our methodology that emphasizes simple yet informative textual features for further machine learning tasks. In our experiments, we address the task as a multi-class classification problem and evaluate the performance of multiple machine learning models on the IEMOCAP dataset [4].

In summary, our framework parses emotional evaluation data from the IEMOCAP dataset. We process multiple session directories, extracting start and end times, file names, emotion labels, and continuous values for valence, arousal, and dominance (Val, act, Dom). Then we processed audio files from the IEMOCAP dataset to generate truncated audio segments corresponding to emotion labels, and we stored these segments in a dictionary as audio vectors. Next, we extracted various audio features from pre-processed audio vectors, like mean, standard deviation of signal, root mean square energy, silence ratio, harmonic components, auto-correlation features, we extracted also Chroma, zero Crossing Rate, and MFCCs. Then we extracted audio file identifiers and their corresponding transcriptions from IEMOCAP transcription files for further use in the classification process. At last, we trained traditional machine learning classifiers, namely, Random Forests, Gradient Boosting, Support Vector Machines, Naive Bayes, and Logistic Regression. The models are evaluated on the IEMOCAP dataset under different modalities, namely, Audio-only, Text-only, and Audio + Text. We have also used an ensemble method to enhance SER and deduced that combining acoustic and text features gives better accuracy for each emotional class. Achieving accuracy by the proposed method is 74.8%, which outperforms the state-of-the-art research on the same dataset and similar modalities. The main contributions of this paper are:

- Extraction of the most related features in multiple domains (time, frequency, amplitude, and spectral distribution).
- A technique to solve the imbalanced and scarcity problem (oversampling technique).
- Combining audio features and corresponding text transcripts as a multi-modal.
- An ensemble model to increase the classification performance that uses RF, XGB, MLP, and LR.

The rest of the paper is as follows. Section 2 gives a brief overview of the related works. Section 3 explains the proposed method. Section 4 describes the experiments and results of our methodology, comparing them to those of

previous methods. Finally, in Section 5, we conclude the paper.

2. Related Works

Recently, several techniques have been proposed to recognize emotions from speech using machine learning or deep learning techniques [5]. Some techniques improve prediction accuracy using a single classifier or ensemble models, and some focus on handling the dataset in preprocessing to solve problems like imbalance, small data size, or high dimensionality. On the other hand, some techniques use single modalities or combine multiple modalities, and others use deep-learning-based enhancement techniques to improve the classification process. In this section, we discuss the techniques proposed in recent years for recognizing emotions from speech.

A framework for speech emotion recognition using both machine learning (RF, XGB, SVM, MNB, LR) and deep learning (MLP, LSTM) algorithms have been proposed (Gaurav Sahu, 2019) [6], it uses multimodal technique, first for audio data, it used up-sampling techniques to alleviate the issue of under-represented emotions, then extracted some handcrafted features like Pitch, Harmonics, Pause, and Central moments. For text data, the available transcriptions were first normalized to lowercase, and any special symbols were removed. It used Term Frequency-Inverse Document Frequency (TFIDF) to extract text features, and then different classification techniques were applied. The best accuracy achieved with Ensemble (RF + XGB + MLP + MNB + LR) with Audio + Text setting is 70.1%.

A Wav2vec 2.0 [7] (An enhanced version of the original wav2vec model) framework is used as a feature extractor for speech emotion recognition in Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings (Leonardo Pepino et al.,2021) [8]. It is a framework for self-supervised learning of representations from raw audio. They extracted features from two released wav2vec 2.0 models.

The model is divided into three steps.

The first step is a local encoder, which encodes the raw audio into a sequence of embeddings with a stride of 20 ms and a receptive field of 25 ms. It contains several convolutional blocks. The contextualized encoder, which is the second level, receives the local encoder representations as input. A number of transformer encoder blocks make up its architecture. In the end, a quantization module with two codebooks with 320 entries each receives the local encoder representations as input. The local encoder representations are converted into logits via a linear map. Gumbel-SoftMax [9] is applied to a sample taken from each codebook based on the logits. A quantized representation of the local encoder output is produced by concatenating the chosen codes and applying a linear transformation to the resulting vector. This representation is then utilized in the objective function. The best accuracy of this framework is 67.2%.

A WavFusion model that leverages the power of wav2vec 2.0 and incorporates textual and visual modalities to enhance the performance of audio-based emotion recognition (Feng Li et al., 2024) [10]. To create a comprehensive multimodal representation, they first use the shallow transformer layer to encode low-level audio features, then use the deep transformer layer to combine text and visual features. The redesigned transformer layer is referred to as a deep transformer, while the original transformer layer is called a shallow transformer layer. Wav2vec 2.0's incorporation of text and visual enhances emotional information in the multimodal fusion representation by detecting relevant data among the huge amount of pre-trained audio information. The best accuracy achieved with this model is 70.53%.

Deep Imbalanced Learning for Multimodal Emotion Recognition in Conversations (Tao Meng et al., 2023) [11]. This study suggests a Class Boundary Enhanced Representation Learning (CBERL) model, a multimodal emotion recognition in conversation framework. CBERL combines complementary input from several modalities and extracts contextual semantic information within modalities. Simultaneously, CBERL considers the issue of unbalanced data distribution and addresses it from three perspectives: loss-sensitivity, sampling method, and data augmentation.

CBERL model used the generative adversarial networks (GAN) data augmentation method to generate new samples, and it also added an identity and classification losses to reduce the distribution difference between the generated and original labels, respectively. In order to reduce feature dimensionality and capture complementary semantic information between various modalities, the original and newly created data were fed into a Deep Joint Variational Autoencoder (DJVAE). Then, the fused low-dimensional feature vector is input into Bi-LSTM to obtain a feature representation of richer contextual semantic information. Next, the obtained contextual feature vectors are fed into the proposed Multi-task Graph Neural Network (MGNN). The best accuracy achieved with this model is 69.365%.

Improving speech emotion recognition with unsupervised speaking style transfer (Leyuan Qu et al.,2023) [12]. This study tackles the problem of data imbalance and shortage in SER, and it presents the EmoAug model. This reconstructs speech in an unsupervised way; it consists of a speech quantization module, a semantic encoder, a paralinguistic encoder, and an attention-based decoder. In. Speech Quantization and Semantic phase, they use HuBERT representations to capture semantic speech information. These representations are acquired through self-supervised training, which eliminates the need for human annotations in ASR training. The paralinguistic encoder aims to learn utterance-level non-verbal information from input audio, which contains speaking styles, emotion states, speaker identities, and so on. It is based on a model to perform the task of speaker verification. using one Conv1D layer, a ReLU function, and Batch Normalization (BN), followed by three SE-Res2 blocks. In the decoder phase, two FC layers, an LSTM module, and FC layers are used to generate Mel-spectrograms. The generated mel-spectrograms are then transformed into waveforms. Then they implement a discriminator to differentiate between genuine and synthesized speech by differentiating pitch changes while preserving semantic content. After pre-training, speaking style transfer can

be achieved by directly replacing the paralinguistic encoder input with a target speaking style. This model (HuBERT Large + EmoAug) achieved a WA of 72.66% and a UA of 73.75%.

Table 1. Related works summary

Reference	Extracted Features	Solving Imbalance / Scarcity	Multi-Modal	Used Classifiers	Enhancement technique	Emotion classes	Highest Accuracy
(Gaurav Sahu, 2019)	Low-level features (Pitch, Harmonics, Speech Energy, Pause, Central moments) for audio & (TFIDF) for text	YES, (Random oversampling)	YES (Audio & Text)	Single and ensemble (RF, XGB, SVM, MNB, LR, MLP, LSTM) for both audio and text	NO	Happy Sad Neutral Angry Fear Surprise	70.1 %
(Leonardo Pepino et al.,2021)	Low-level features (Pitch, Jitter, Formant, Shi-mmer, Loudness, Harmonics-to-Noise Ratio (HNR), MFCC) for audio & Wav2Vec 2.0 for text	NO	YES (Audio & Text)	CNN, LSTM	YES Attention Heads	Happy Sad Neutral Angry	67.2 %
(Junghun Kim et al.,2022)	High-level acoustic features (MFCCs)	NO	NO (Audio only)	CNN, LSTM, DNN	YES Focus-Attention (FA), Calibration-Attention (CA) & Multi-head self-attention.	Happy Sad Neutral Angry	72.83 %
(Tao Meng et al., 2023)	Used the openSMILE toolkit for audio & BERT for text	YES, (Synthetic Minority Over-Sampling Technique (SMOTE)	YES (Text, Audio & Video)	Bi-LSTM, CNN, DNN	NO	Happy Sad Neutral Angry Excited Frustrated	67.78 %
(Leyuan Qu et al. ,2023)	High-level acoustic features (Mel-Spectrograms)	YES, (EmoAug)	NO (Audio only)	HuBERT, CNN, LSTM, DNN	Style Transfer & Attention	Happy Sad Neutral Angry	73.75 %
(Feng Li et al., 2024)	Used RoBERTa for text, wav2vec2.0 for audio & EfficientNet for video	NO	YES (Text, Audio &Video)	CNN, RNN	Gated Cross-Modal Attention	Positive & Negative	70.53 %

Improving Speech Emotion Recognition Through Focus and Calibration Attention Mechanisms (Junghun Kim et al.,2022) [13]. They proposed that performance in emotion recognition is significantly affected by the highest amplitude. Therefore, emotion detection performance can be enhanced by identifying the highest amplitude and learning the associated representations. They suggest a new Calibration-Attention (CA) mechanism as well as a Focus-Attention (FA) mechanism. FA assists in the detection of the segment's highest amplitude part. Because of multi-head self-attention, there are several attention alignments, and not every attention head has the right alignment. So, they assigned each attention head a different weight and developed a new CA mechanism that will increase the use of surrounding contexts and modify the flow of information. The proposed architecture is as follows. Initially, two 1D convolutional layers with a channel size of 64 and a kernel size of (3, 1) are stacked. Batch normalization and max pooling with kernel and stride size (2, 1) come after each convolutional layer. Two bidirectional LSTM layers with a hidden size of 256 are then stacked. Two FC layers are then layered after multi-head self-attention with a head size of 8 that connects the output of the LSTM layer. In the scaled dot-product process, the CA mechanism runs before the multi-head being concatenated, and the FA mechanism is inserted before the SoftMax computation. MHSA-FACA (Multi-head self-attention focus attention calibration attention) achieved 72.01% in WA and 72.83% in UA.

Summary. A summary of the recent research is shown in Table 1. All existing approaches use the Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) as the dataset for training and testing the different models (USC machine learning repository, 2008).

To overcome the data imbalance and scarcity problem, a number of techniques have been introduced, as presented in Table 1. First, a random over-sampling technique is used through balancing a dataset by increasing the samples of the minority class (Gaurav Sahu, 2019). Second, the SMOTE method has been used. This method produces samples on the line between minority class samples (Tao Meng et al., 2023). Finally, the EmoAug model is used, which addresses the challenge of data scarcity by enriching emotional speech with various prosodic attributes like stress and rhythm (Feng Li et al., 2024).

Some previous studies rely on speech information only (Junghun Kim et al.,2022; Leyuan Qu et al.,2023). However, single modality is not adequate to meet real-world demands. To address this problem, researchers utilize multi-modal information to identify emotional states. In multi-modal Emotion Recognition, the information of multiple modalities such as audio and text (Gaurav Sahu, 2019), (Leonardo Pepino et al.,2021) is used, others utilized a multi-modal of audio, text and video (Tao Meng, et al., 2023), (eng Li et al., 2024), providing additional cues to mitigate semantic and emotional ambiguities. To enhance the performance of a classifier model, some researchers used an auxiliary enhancement technique such as Attention Heads (Leonardo Pepino et al.,2021), Focus-Attention (FA), Calibration-Attention (CA) & Multi-head self-attention (Junghun Kim et al.,2022), Style Transfer & Attention (Leyuan Qu et al., 2023), and Gated Cross-Modal Attention (Feng Li et al., 2024). The best accuracy achieved is 73.75 % in predicting four emotions (Happy, Sad, Neutral, and Angry) using a single modality (audio only), HuBERT Large + EmoAug model to address data imbalance/scarcity, CNN+LSTM classifiers, and Style Transfer & Attention enhancement technique.

3. Proposed Method

To improve the prediction performance using machine-learning techniques, we considered three aspects: the data quality, modality, and classifiers used. In this section, we present a framework designed to enhance speech emotion recognition by addressing these aspects. The framework consists of four main components: (1) improving data quality through dataset balancing, (2) dimensionality reduction by selecting the most contributing features, (3) combining speech and text features via multimodal fusion, and (4) enhancing classification performance using ensemble learning through average predictions from multiple classifiers.

Since IEMOCAP is a multimodal dataset, it includes speech (audio recordings of conversations and monologues) performed by actors and text (transcriptions of the audio). The framework first starts by pre-processing the original audio and text files, followed by random oversampling to balance the dataset. Dimensionality reduction is then applied using Gradient Boosting's built-in feature importance calculation to select the most relevant features from both audio and text. The balanced dataset is randomly split into training (80%) and testing (20%) sets. In the next step, an early fusion technique combines the text and audio features at the feature level [14]. This technique involves concatenating features from different modalities into a single feature vector, which is then passed to a classifier for training. This approach leverages complementary information from both modalities. Finally, the training set is used to train the ensemble classifier. As part of our experiments, we first perform emotion recognition using speech only, text transcriptions only, and finally combine both modalities for a more in-depth analysis.

3.1. Speech-based Emotion Recognition

In this section, we discuss the speech-based emotion recognition technique.

A. Pre-processing

According to the first frequency study of the IEMOCAP dataset, it is clear that the distribution of utterances for each class is not identical, which makes the dataset unbalanced, as shown in Figure 1. Distribution of the data for each emotional category.

To address class imbalance, we merged the samples from the "happy" and "excited" classes into a single "happy" class, as the two emotions are closely related and often difficult to distinguish both semantically and acoustically. However, we acknowledge that this merging may cause a loss of granularity in emotion classification, and it may reduce the model's ability to distinguish between nuanced variations in positive emotional expressions. This trade-off was deemed acceptable considering the significant imbalance between classes. Furthermore, we excluded samples labeled as "others" due to their ambiguous nature, which made them difficult to interpret even for human annotators. To further balance the dataset, we applied random oversampling to increase the number of instances in minority classes, specifically "fear" and "surprise." These preprocessing steps resulted in a total of 9,797 samples.

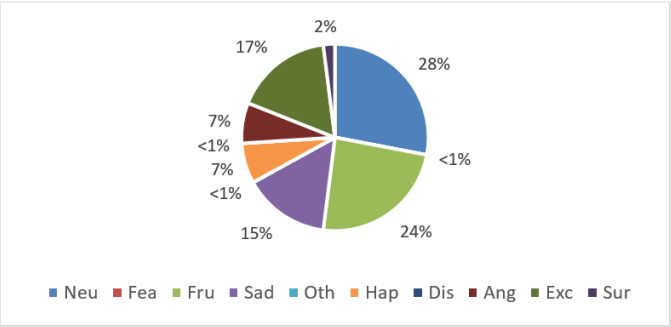

Fig.1. Distribution of the data for each emotional category. (Neu= neutral state, Hap = happiness, Sad = sadness, Sur = surprise, Ang = anger, Fea = fear, Fru = Frustration = disgust, Exc = Excited and Oth = Other)

Table 2. Final class distribution

Emotion	Happy	Angry	Sad	Fear	Surprise	Neutral	Total
Approximate Count	1636	1103	2933	1708	1177	1240	**9797**

In order to prepare the data for processing, we first iterate sessions 1 through 5 evaluation files of the IEMOCAP dataset. These files contain annotated emotion labels for recorded audio-visual conversations and provide metadata about each spoken utterance, including:

- Time Stamps: The start and end times of the utterance within the recording.
- WAV File Name: The name of the corresponding audio file.
- Emotion Label: The perceived emotion (e.g., happy, sad, angry, neutral).
- VAD Scores: Numerical values (range: 1-5) indicating emotional intensity:

  - Val: Valence score (emotion positivity/negativity).
  - Act: Arousal score (energy level of emotion).
  - Dom: Dominance score (control or intensity of emotion).

This data was saved in a CSV file, structured similarly to Table 3.

Table 3. Retrieved data from evaluation files

start_time	end_time	wav_file	emotion	val	act	dom
0.0	2.5	Ses01F.wav	happy	2.5	3.0	2.0
1.5	4.0	Ses01M.wav	sad	1.0	1.5	2.5

After retrieving the labels, we utilized the compiled CSV file to segment the original WAV files into multiple frames and construct audio vectors as follows: The speech signals in the dataset were initially processed at a 44,100 Hz sampling rate, following the Nyquist theorem for audio signals [15]. According to this theorem, the sampling frequency must be more than twice the highest frequency that needs to be reproduced. Since human hearing ranges approximately from 20 Hz to 20,000 Hz, a sampling rate exceeding 40 kHz is required. The standard CD audio rate of 44.1 kHz was selected as it meets the Nyquist criterion while ensuring compatibility with existing video equipment. Next, we processed sessions 1 to 5 of the IEMOCAP dataset. For each session, we iterated through the WAV files, extracting metadata such as start and end times, file names, and emotion labels. We then computed the start and end frames based on the time and sampling rate, truncating the audio vector accordingly. The processed audio vectors were stored in the audio_vectors dictionary, using the truncated file name as the key. Finally, we saved the dictionary as a pickle file for each session, which improves efficiency and simplifies data handling in audio processing pipelines.

B. Dataset Dimensionality Reduction

In machine learning (ML), dimensionality reduction [16] is a basic technique that simplifies datasets by reducing the quantity of input variables or features. This is crucial for improving computational efficiency and model performance. This involves keeping the essential (optimal) features while eliminating irrelevant ones to enhance the model's accuracy.

So, we started to extract some acoustic features by loading audio vectors from pickle files for each session. The acoustic characteristics of speech signal features such as:

- Speech energy: The energy of a speech signal correlates with its loudness, making it a useful feature for detecting emotions. For instance, an angry speech signal typically has higher energy, while a sad speech signal exhibits lower energy levels (as shown in Figure 2).

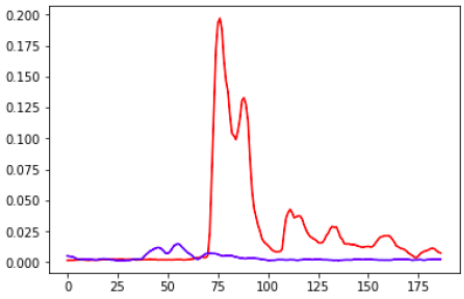

Fig.2. Audio signals: RMSE plots of angry (red) and sad (blue)

We represent speech energy using the standard Root Mean Square Energy (RMSE), calculated using Equation (1):

$$E = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y[i]^2} \tag{1}$$

where *y[i]* denotes the speech signal values in a frame, and *n* is the number of samples in that frame. RMSE is computed on a frame-by-frame basis, and both its average and standard deviation are extracted as features for analysis.

- Pitch: Pitch [17] plays a crucial role as the waveforms generated by our vocal cords vary with our emotions. Equation (2) represents a center-clipping function, commonly used in pitch detection

$$\mathrm{y_{clipped}}[n] = \begin{cases} y[n] - Cl, & if\ y[n] \geq Cl \\ 0, & if\,|y[n]| < Cl \\ y[n] + Cl\,, & if\ y[n] \leq Cl \end{cases} \tag{2}$$

y[n] is the input, which is center clipped to give a resultant signal, *yclipped[n]*. *Cl* represents the clipping threshold. Center Clipping is a technique used in autocorrelation-based pitch detection to remove weak signals, enhance dominant frequencies, and filter out background noise while preserving key voice features.

- Harmonics [18]: In the emotional state of anger or stressed speech, additional excitation signals are present beyond just pitch. These excitations appear in the spectrum as harmonics and cross-harmonics (see Figure 3).

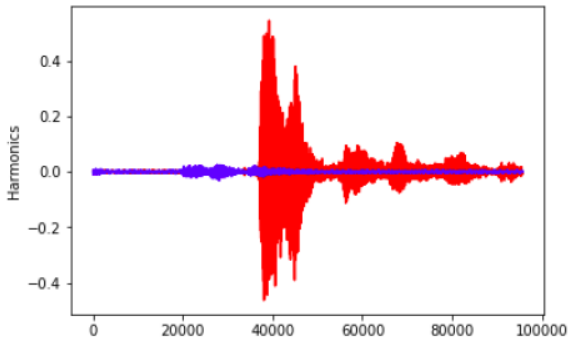

Fig.3. Harmonics of angry (red) and sad (blue) audio signals

To calculate harmonics, we use a median-based filter as described in [15]. The median filter is applied over a window of size l, defined as Equation (3):

$$y[n] = median(x[n-k : n+k]|\ k = (l-1)/2) \tag{3}$$

where *l* is an odd number. If *l* is even, the median is computed as the average of the two middle values in the sorted list.

This filter is then applied to S_h, the h^{th} frequency slice of a given spectrogram *S*, to obtain the harmonic-enhanced spectrogram frequency slice H_h, *as* we used Equation (4):

$$\mathrm{H}_i = M(S_h, l_{harm}) \tag{4}$$

where: *M* is the median filter, *i* is the i^{th}time step, and l_{harm} is the length of the harmonic filter. This process enhances harmonics in the spectrogram, making it useful for speech analysis in emotion.

- Pause: We use this feature to represent the silent portions in the audio signal. This measure is directly linked to emotions; for example, when excited (such as feeling angry or happy), we tend to speak faster, leading to a lower Pause value. The feature is computed as Equation (5):

$$Pause = Pr(y[n] < t) \tag{5}$$

where t represents a carefully chosen threshold ≈ 0.4*E, E is the RMSE.

- Central Momentum: Finally, we incorporate a summary of the input signal by computing the mean and standard deviation of its amplitude. These statistical measures help capture overall variations in the signal, providing valuable information for analysis.
- Chroma: Chroma features are a representation of the 12 pitch classes (C, C#, D, D#, …, B) in an audio signal; it focuses only on pitch classes, ignoring absolute frequency values, and is computed using Short-Time Fourier Transform (STFT) [19] or Constant-Q Transform (CQT) [20].

Chroma is computed as follows:

- Convert audio to a spectrogram (using STFT or CQT).
- Map frequencies to 12 chroma bins (one for each pitch class).
- Normalize chroma values to remove amplitude variations.

Figure 4 shows chroma representation of angry and sad speech.

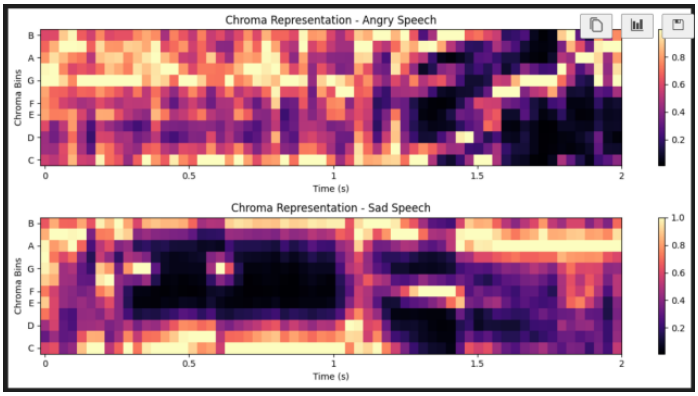

Fig.4. Chroma visualization of angry and sad

- Zero Crossing Rate (ZCR): is a measure of how frequently the signal waveform crosses the zero-amplitude axis. It is commonly used in speech processing and audio classification. ZCR can be computed using Equation (6).

$$ZCR = \frac{1}{N-1}\sum_{i=1}^{N-1} 1(s[i] \,.\, s[i-1] < 0) \tag{6}$$

Where N = Number of samples in a frame, $s[i]$ = Signal amplitude at sample i, the indicator function 1 counts when the signal changes sign (crosses zero).

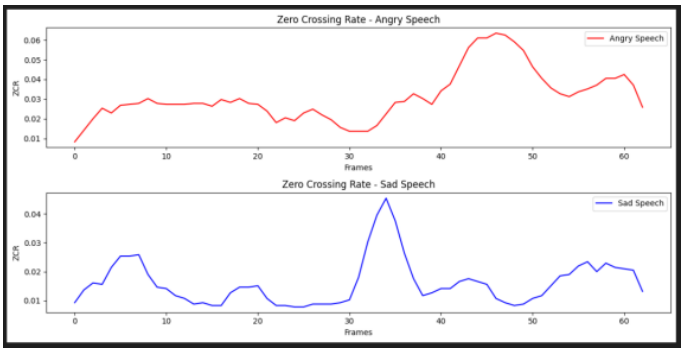

Fig.5. Zero crossing rate visualization of angry and sad speech

Figure 5 shows that angry speech has a higher ZCR (faster variations in signal), indicating rapid fluctuations in amplitude, and sad speech has a lower ZCR (smoother waveform, fewer crossings), indicating a more continuous and less

abrupt signal.

- Mel-frequency cepstral coefficients (MFCCs) [21] are features extracted from audio signals that represent the spectral shape (short-term power spectrum) of a sound. (MFCCs) are computed using Equation (7).

$$C_n = \sum_{k=1}^{k} S_k \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{k}\right] \quad (7)$$

Where:

- C_n is the n^{th} MFCC coefficient.
- K is the total number of Mel filter banks.
- S_k is the log energy from the n^{th} Mel filter bank.
- Cos is the Discrete Cosine Transform (DCT) function.
- n is the coefficient index (usually 13 to 20 coefficients are used).
- π is a Mathematical constant (≈ 3.1416).

Figure 6 shows the process of computing MFCC for a signal.

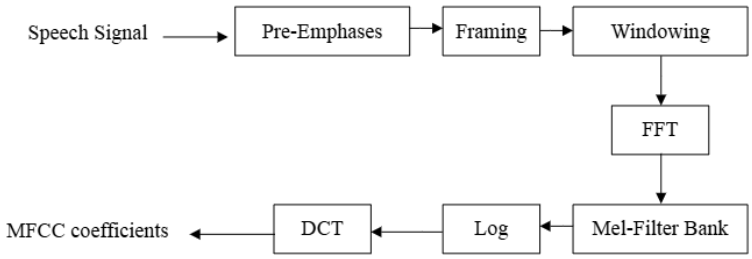

Fig.6. Block diagram of MFCC

MFCCs Function as follows:

- Pre-emphasis: Enhances high frequencies to achieve spectral balance.
- Framing & Windowing: Divides the signal into small overlapping segments to observe temporal variations.
- Fourier Transform (FFT): Transforms the signal from the time domain to the frequency domain.
- Mel-Scale Filtering: Adjusts frequencies to the Mel scale, simulating human auditory perception.
- Logarithm & Discrete Cosine Transform (DCT): Compresses spectral data into a compact set of coefficients, minimizing redundancy.
- MFCC Extraction: Typically extracts 13 to 20 coefficients (most commonly 13) as key features for analysis.

The first 13 MFCC coefficients are employed in this system's feature extraction process. Since the majority of the information, including the average power of the input signal and the distribution of spectral energy between low and high frequencies, is included in the lower order coefficients. A total of 21 features are extracted for each frame and concatenated for each utterance.

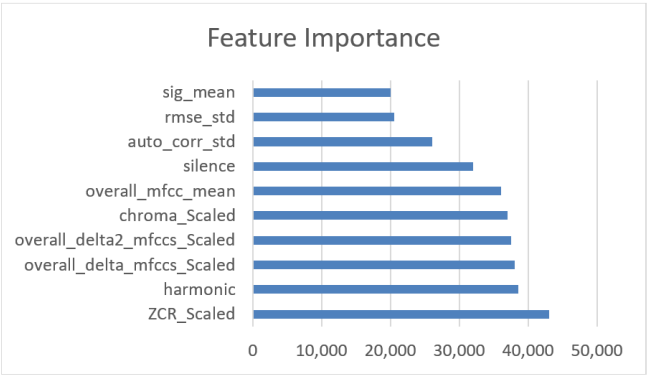

Fig.7. Importance of audio features

C. *Feature Selection and Importance*

Feature selection is the process of selecting features from extracted features to remove unused information and redundancy, and to reduce the processing time. We selected the Global features, such as the value of Min, Max, Mean,

Median, and standard deviation, to save processing time.

We explored which features in this classification task contribute the most to prediction. So, we evaluated the audio features using the XGB model.

Figure 7 'The Importance of audio features' shows that ZCR contributes the most, and it is interesting to see that "Harmonic," which is directly related to the excitation in signals, also have high contribution as scaled first and second derivatives of MFCC and chroma.

3.2. Text-based Emotion Recognition

The primary objective of this part of the study was to process and align the text transcriptions with their corresponding audio files, which are essential for emotion recognition tasks.

Data Structure:

The text files containing the transcriptions are organized by session. Each session includes several dialogue files, where each line typically represents a speaker's statement. The audio code at the beginning of each line corresponds to a unique identifier for the audio clip associated with that statement. The transcription follows the colon symbol and represents the spoken content of the audio clip.

To work with these transcriptions and pair them with their corresponding audio files, the following preprocessing steps were carried out:

- A regular expression (regex) was designed to extract the audio code from each transcription line.
- The processing pipeline iterates over each session in the IEMOCAP dataset (from session 1 to session 5), navigating through the respective directories to access the transcription files. These files are read line by line to extract both the audio code and the corresponding transcription. The extracted pairs are stored in a dictionary, with the audio code as the key and the transcription as the value.
- Each line's audio code is extracted using the regex and is associated with the transcription, which is the part of the line after the colon (:). The transcription is stripped of any unnecessary white space for cleanliness. These pairs are stored in a dictionary, which serves as the main mapping between audio codes and their respective transcriptions
- The dictionary is saved as a binary file (audiocode2text.pkl) for easy access during model training and evaluation.
- The transcriptions processed in this step serve as one modality (text) for emotion recognition. The next stage involves integrating these text-based features with audio features for a more comprehensive multimodal emotion recognition system. The combination of these two modalities can significantly enhance the system's ability to detect emotions by leveraging both the acoustic and linguistic cues present in the data.

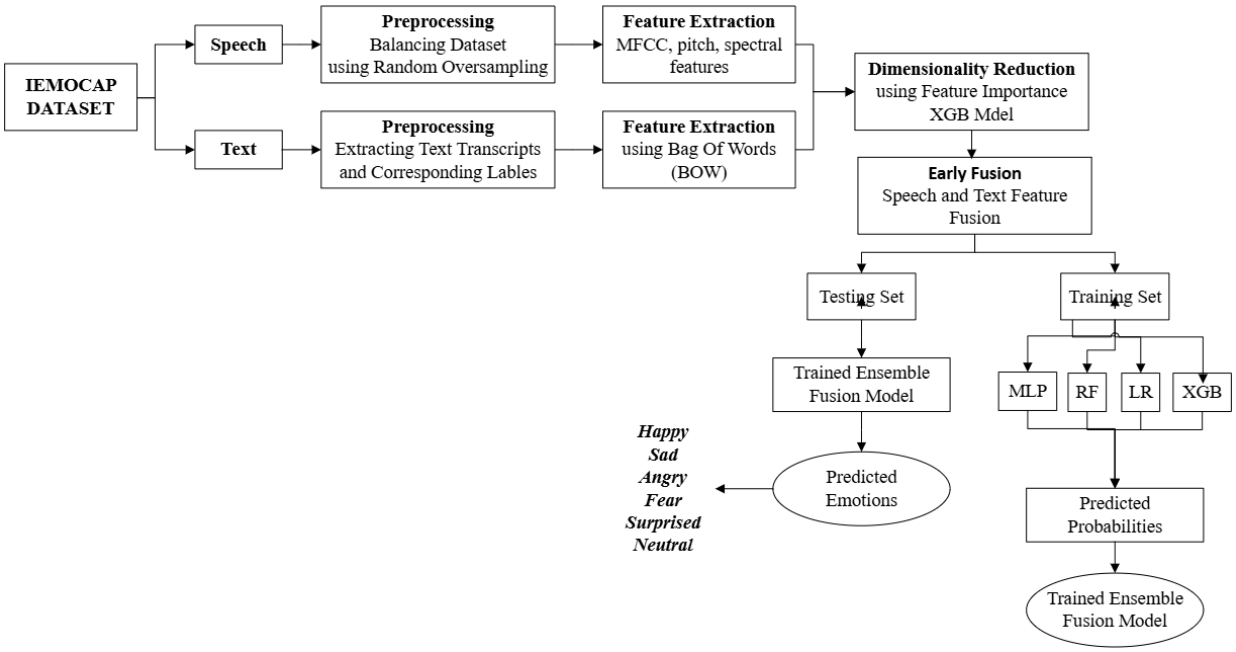

Fig.8. Block diagram of our proposed framework

To convert textual data into a numerical format suitable for machine learning, we employed the Bag of Words (BoW) [22] approach using Count Vectorizer from scikit-learn" popular open-source machine learning library for Python. BoW is a widely used text representation technique that transforms raw text into structured feature vectors while disregarding the word order. Formally, we could consider BoW as a text data feature extraction approach. It turns words in a text into a matrix representation by extracting their features. This allows us to see which word is included in the sentence and how

frequently. The BoW technique depends on breaking down a sentence into separate words or tokens, which is called tokenization. After tokenization, BoW generates a dictionary or vocabulary of distinct terms or tokens from the complete text data corpus (Vocabulary Creation). Each document is represented as a vector, where each entry corresponds to the frequency of a word appearing in that document (Word Frequency Counting).

3.3. Fusing Speech and Text Features

The paper's main proposed technique is to combine speech-based acoustic features and text-based features for categorical emotion recognition. We combine the feature vectors from the two modalities by concatenating the text and audio feature vectors to obtain the combined feature vectors. The optimal model has been established by composing many model combinations, including Multi-Layer Perceptron (MLP), Random Forest (RF), Gradient Boosting (XGB), Support Vector Machine (SVM), Logistic Regression (LR), and Multinomial Naive Bayes (MNB). We found that averaging the predicted probabilities of ensemble MLP, RF, XGB, and LR is the best-obtained result among others. Figure 8 shows the block diagram of our proposed framework.

4. Results

In this section, we present our proposed approach using the IEMOCAP database. Therefore, we start by giving a general description of the dataset before presenting the experimental setup and the results achieved.

4.1. IEMOCAP Dataset

The IEMOCAP dataset, introduced in 2008 by researchers at the University of Southern California (USC), is used in our study to evaluate the proposed approach for emotion recognition. It consists of both scripted and spontaneous acts of five recorded sessions of conversations with 10 speakers, totaling around 12 hours of audio-visual content. It represents nine categorical emotions (anger, happiness, excitement, sadness, frustration, fear, surprise, other, and neutral state). We used only 6 emotion categories, i.e., anger, happiness, sadness, neutral, surprise, and fear.

The dataset also includes dimensional labels (activation, valence, and dominance values ranging from 1 to 5), which are not utilized in our work. Each session is segmented into multiple utterances, with every utterance annotated by at least three human evaluators, and we further divide each utterance file to produce wav files for every sentence using start and finish timestamps supplied for the transcribed sentences. This produces about 10039 audio files, while we used only 9797 utterances to extract features from. The final emotion label for each utterance is determined using a majority voting scheme. Our model is trained and evaluated using these categorical emotion categories as the basis for data. IEMOCAP is used to study and analyze emotions in human interactions. It is also suitable for emotion recognition in speech, text, and multimodal settings.

Dataset Overview

General Information

- Keywords: Emotional, Multimodal, Acted, Dyadic
- Language: English
- 10 Actors: 5 male and 5 female
- Emotion Elicitation Techniques: Scripts and Improvisations

Available Modalities

- Motion Capture Face Information
- Speech (audio)
- Videos (face and head motion capture)
- Head Movement and Head Angle Information
- Dialog Transcriptions
- Word level, Syllable level, and Phoneme level alignment

Annotations

- Sessions are manually segmented into utterances
- Each utterance is annotated by at least 3 human annotators
- Categorical Attributes:
    - anger, happiness, excitement, sadness, frustration, fear, surprise, other, and neutral state
- Dimensional Attributes:
    - valence, activation, dominance

4.2. The Experimental Setup

Our experiments have been performed on an Intel Core i7, CPU 2.60 GHz, and 16.0 GB of memory. Python 3.11 has been used to implement our proposed framework.

In our implementation, a number of packages have been utilized, such as sickit learn, numpy, scipy, pandas, matplotlib, soundfile, and librosa libraries, which are commonly used for processing the audio files and extracting features from them. We randomly split our dataset into a train (80%) and test (20%) set. To ensure a fair comparison between our method and related studies, we conduct experiments on six emotion categories: Angry, Happy, Sad, Fear, Surprise, and Neutral.

4.3. The Proposed Approach Results

In this section, we present the experimental results measured by the same evaluation metrics applied to all models, i.e., accuracy, F1-score, precision, and recall, which are discussed below. In Section 3, we proposed a framework that consists of four main phases: increasing the data quality through balancing the dataset using random over sampling (phase 1), dimensionality reduction through selection of the most contributing features using XGB pretrained model (phase 2), using multimodal by combining speech features with word embeddings from speech transcriptions using bag of words (phase 3), and increasing the classification process performance by using ensemble prediction by averaging the predicted probabilities of individual classifiers of SVM, RF, XGB, MLP, and LR (phase 4). To evaluate the different aspects of our proposed framework, we have performed evaluations for each classifier individually and evaluated our proposed ensemble approach in three settings:

- Speech-based emotion recognition: In this approach, we utilized only the previously mentioned audio feature vectors to train each classifier separately, as well as in an ensemble setting.
- Text-based emotion recognition: In this approach, we utilized only the text feature vectors extracted using the term frequency inverse document frequency (TFIDF), Word2Vec, and Bag of Words (BoW) methods to train all classifiers. We selected BoW because it delivered the highest accuracy results.
- Fusion of Speech and Text Features: In this approach, we integrated feature vectors from both audio and text by directly concatenating them to form a unified representation. This experiment allows us to assess how the fusion of these modalities impacts the model's performance. For all settings, all models are evaluated in the same metric, i.e., accuracy, F1, Precision, and Recall.

Since accuracy provides a general measure of the model's correctness, it may not fully capture the model's ability to identify minority classes, especially in imbalanced datasets. Therefore, we also include recall, which quantifies the proportion of actual positive instances correctly identified by the model as computed in equation 8:

$$Recall = \frac{TP}{TP+FN} \tag{8}$$

where TP is True Positives, and FN is False Negatives. Precision measures the proportion of correct predictions among all instances predicted as positive, as computed in equation 9:

$$Precision = \frac{TP}{TP+FP} \tag{9}$$

where FP is False Positives. Accuracy measures the overall proportion of correct predictions made by the model across all classes as computed in equation 10:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} = \frac{TP+TN}{TP+TN+FP+FN} \tag{10}$$

where TN represents True Negatives. To balance the trade-off between precision and recall, we also report the F1-score, which provides a harmonic mean of precision and recall as computed in equation 11:

$$F1-Score = 2 * \frac{Precision*Recall}{Precision+Recall} \tag{11}$$

These metrics were selected to provide a thorough evaluation of the model's performance, particularly in the context of imbalanced emotion recognition tasks.

Table 4 presents the performance results of six speech models. The ensemble of RF, XGB, and MLP (E1) achieved higher accuracy compared to individual models.

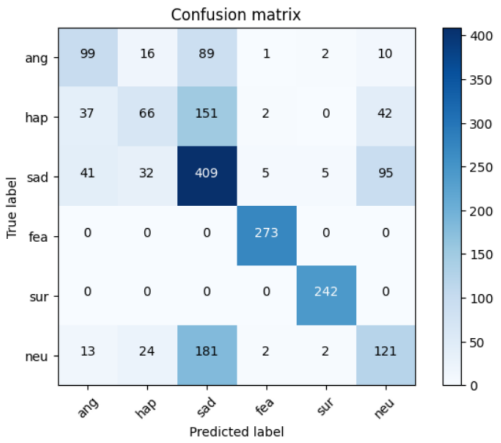

Fig.9. Confusion matrix of Ensemble speech-based settings (E1)

Table 4. Speech-based results

Model	Accuracy	F1	Precision	Recall
RF	0.623	0.623	0.648	0.621
XGB	0.606	0.619	0.631	0.617
SVM	0.543	0.541	0.576	0.547
MLP	0.640	0.613	0.615	0.623
LR	0.353	0.240	0.363	0.263
E (1)	0.619	0.621	0.645	0.621

A look at the confusion matrix (Figure 9) reveals that identifying “neutral” or differentiating between “angry”, “happy”, and “sad” is the most challenging task of the model.

For text-based settings, Table 5. presents the performance for three different text representation models: TF-IDF, Word2Vec, and Bag of Words (BoW). Bag of Words (BoW) performs best across all metrics, with the highest accuracy (0.711), F1-score (0.720), precision (0.744), and recall (0.708). This suggests that BoW captures useful features for classification in the emotion recognition task.

Table 5. Performance of TFIDF, Word2vec, and BOW models

Model	Accuracy	F1	Precision	Recall
TFIDF	0.631	0.633	0.671	0.616
Word2vec	0.585	0.563	0.622	0.572
BoW	0.711	0.720	0.744	0.708

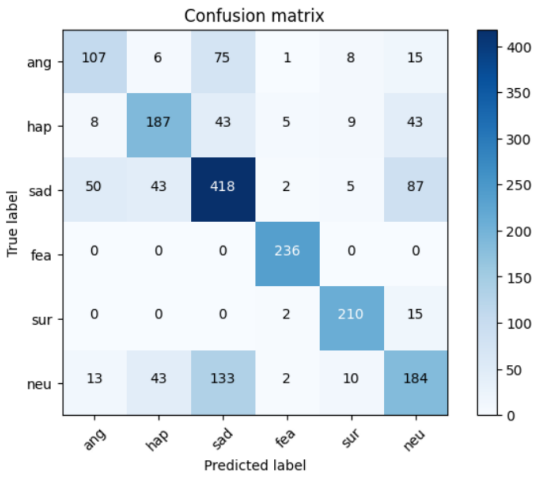

Fig.10. Confusion matrix of the Ensemble text-based setting

Table 6 shows that using BoW with the ensemble model (E2), which combines RF, XGB, SVM, MLP, and LR, outperforms all individual models with the highest accuracy (0.711), F1-score (0.720), precision (0.744), and recall (0.708). The confusion matrix (Figure 10) shows that our text-based models are able to discriminate between the six emotions fairly well. We see that the most difficult textual feature to distinguish very clearly is "sad".

For combined speech-text based settings table 7 presents the performance of different machine learning models when combining speech and text features. It is clear that the ensemble model (E3), which combines RF, XGB, MLP, and LR, Outperforms Individual Models with the highest accuracy (0.748), F1-score (0.761), precision (0.761), and recall (0.761). It proves that the combination of multiple models provides a more robust and generalizable approach.

In Figure 11. It is possible to show that the model was better able to identify the "sad" class according to the audio features, while the textual features assisted in correctly classifying the "angry" and "happy" classes.

Table 6. Text-based results using BoW

Model	Accuracy	F1	Precision	Recall
RF	0.658	0.655	0.712	0.656
XGB	0.610	0.595	0.693	0.587
SVM	0.646	0.663	0.660	0.675
MLP	0.698	0.698	0.691	0.706
LR	0.652	0.667	0.677	0.667
E (2)	0.711	0.720	0.744	0.708

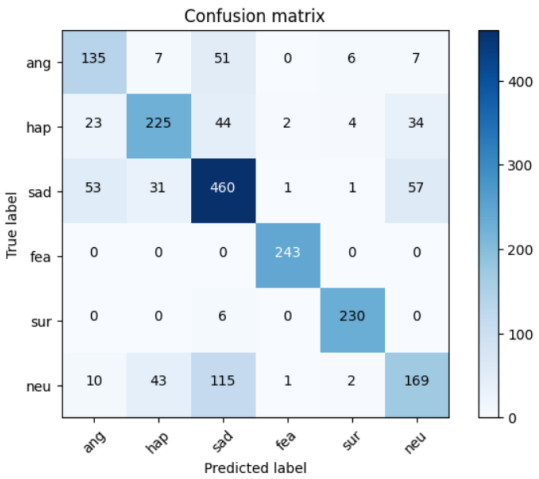

Fig.11. Confusion matrix of ensemble, fusion setting

Table 7. Combined speech and text results

Model	Accuracy	F1	Precision	Recall
RF	0.693	0.695	0.759	0.683
XGB	0.690	0.701	0.732	0.696
SVM	0.687	0.697	0.689	0.710
MLP	0.709	0.732	0.730	0.737
LR	0.699	0.705	0.709	0.709
E(3)	0.748	0.761	0.761	0.761

4.4. Results Summary and Comparison to Existing Approaches

The experimental results, as shown in Table 8, showed that the best classification performance is achieved when a) the dataset is preprocessed to increase the data quality through balancing the dataset and dimensionality reduction through the selection of the most contributing features. b) Combining acoustic (MFCC, pitch, spectral features) and textual BoW features. c) Using an ensemble learning approach to combine individual classifiers (i.e., RF, XGB, MLP, and LR) with a soft voting mechanism.

Table 8. Comparison between the existing and the proposed approach

Reference	Best Classifier Used	Highest Accuracy
(Gaurav Sahu, 2019)	Single and ensemble (RF, XGB, SVM, MNB, LR, MLP, LSTM) for both audio and text	70.1%
(Leonardo Pepino et al.,2021)	CNN, LSTM.	67.2%
(Junghun Kim et al.,2022)	CNN, LSTM, DNN.	72.83%
(Tao Meng, et al., 2023)	Bi-LSTM, CNN, DNN.	67.78 %
(Leyuan Qu et al. ,2023)	HuBERT, CNN, LSTM, DNN.	73.75 %
(Feng Li et al., 2024)	CNN, RNN.	70.53 %
Proposed Framework	Using BOW and Ensemble (RF, XGB, MNB, LR, MLP) for both audio and text.	74.8%

5. Conclusions

This paper presents a speech emotion recognition model that utilizes both speech and text features. A set of acoustic features - including Energy, Pitch, Harmonics, Pause, Central Momentum, Chroma, Zero Crossing Rate, and MFCCs - was extracted from the component of the speech utterances, while textual features were also employed to improve the model's performance. By implementing an ensemble method, the approach significantly improved emotion classification accuracy, particularly for related emotions, such as happiness vs. excitement and anger vs. frustration, where ambiguity was greatly reduced compared to individual classifiers. The proposed approach outperformed state-of-the-art findings on the same dataset and comparable modalities, with an accuracy of 74.8%.

However, there are a few limitations to consider. First, the model relies on a limited set of features that may restrict its ability to capture the full complexity of emotional expressions. Furthermore, while combining text and speech data can be advantageous, it might not completely solve the problems associated with multimodal emotion recognition in noisy or real-world settings. These constraints might be overcome in future research by enlarging the dataset to contain a wider range of samples, exploring new feature sets, and optimizing the ensemble approach. Moreover, the model could be tested on more datasets to evaluate its robustness across different contexts and domains.

References

[1] Taiba Majid Wani, Teddy Surya Gunawan, Syed Asif Ahmad Qadri, Mira Kartiwi, Eliathamby Ambikairajah, "A Comprehensive Review of Speech Emotion Recognition Systems", *IEEE Access*, Vol.9, pp.47795–47814, 2021. DOI:10.1109/ACCESS.2021.3068045.

[2] Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, Qianli Ma, "A Survey on Ensemble Learning", *Frontiers of Computer Science*, Vol.14, No.2, pp.241–258, 2020. DOI:10.1007/s11704-019-8208-z.

[3] Elena Rudkowsky, Martin Haselmayer, Matthias Wastian, Marcelo Jenny, Štefan Emrich, Michael Sedlmair, "More Than Bags of Words: Sentiment Analysis With Word Embeddings", *Communication Methods and Measures*, Vol.12, No.2–3, pp.140–157, 2018. DOI:10.1080/19312458.2018.1455817.

[4] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, Shrikanth Narayanan, "IEMOCAP: Interactive Emotional Dyadic Motion Capture Database", *Language Resources and Evaluation*, Vol.42, No.4, pp.335–359, 2008. DOI:10.1007/s10579-008-9076-6.

[5] Dias Issa, M.Fatih Demirci, Adnan Yazici, "Speech Emotion Recognition With Deep Convolutional Neural Networks", Biomedical Signal Processing and Control, Vol.59, pp.101894, 2020. DOI: 10.1016/j.bspc.2020.101894.

[6] Gaurav Sahu, "Multimodal Speech Emotion Recognition and Ambiguity Resolution", *arXiv preprint* abs/1904.06022, 2019. DOI:10.48550/ARXIV.1904.06022.

[7] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, Michael Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations", *Advances in Neural Information Processing Systems*, Vol.33, pp.12449–12460, 2020. DOI:10.5555/3454287.3455167.

[8] Luca Pepino, Pere Riera, Luis Ferrer, "Emotion Recognition From Speech Using wav2vec 2.0 Embeddings", *arXiv preprint* abs/2104.03502, 2021. DOI:10.48550/ARXIV.2104.03502.

[9] Eric Jang, Shixiang Gu, Ben Poole, "Categorical Reparameterization With Gumbel-Softmax", *arXiv preprint* abs/1611.01144, 2016. DOI:10.48550/ARXIV.1611.01144.

[10] *Feng Li, Jiusong Luo, Wanjun Xia, "WavFusion: Towards wav2vec 2.0 Multimodal Speech Emotion Recognition", arXiv preprint abs/2412.05558, 2024. DOI:10.48550/ARXIV.2412.05558.*

[11] Tao Meng, Yuntao Shou, Wei Ai, Nan Yin, Keqin Li, "Deep Imbalanced Learning for Multimodal Emotion Recognition in Conversations", *IEEE Transactions on Artificial Intelligence*, Vol.5, No.12, pp.6472–6487, 2024. DOI:10.1109/TAI.2024.3445325.

[12] Leyuan Qu, Wei Wang, Cornelius Weber, Pengcheng Yue, Taihao Li, Stefan Wermter, "Improving Speech Emotion Recognition With Unsupervised Speaking Style Transfer", *Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vol.1, No.1, pp.10101–10105, 2024. DOI:10.1109/ICASSP48485.2024.10446186.

[13] Junghun Kim, Yoojin An, Jihie Kim, "Improving Speech Emotion Recognition Through Focus and Calibration Attention Mechanisms", *arXiv preprint*, 2022. DOI/URL: https://arxiv.org/abs/2208.10491.

[14] Yanlin Liu, Aibin Chen, Guoxiong Zhou, Jizheng Yi, Jin Xiang, Yaru Wang, "Combined CNN LSTM With Attention for Speech Emotion Recognition Based on Feature-Level Fusion", *Multimedia Tools and Applications*, Vol.83, No.21, pp.59839–59859,

2024. DOI:10.1007/s11042-023-17829-x.
[15] Robert Toft, “Digital Audio: Sampling, Dither, Aliasing, and Bit Depth”, *Technical Matters*, Popular Music Forum, Western University, 4–5 May 2024. URL: https://ir.lib.uwo.ca/popmusicforum_techmatters/11/
[16] Weikuan Jia, Meili Sun, Jian Lian, Sujuan Hou, “Feature Dimensionality Reduction: A Review”, Complex & Intelligent Systems, Vol.8, No.3, pp.2663–2693, 2022. DOI:10.1007/s40747-021-00637-x.
[17] Daniel Hirst, Céline De Looze, "Measuring Speech: Fundamental Frequency and Pitch", *Cambridge Handbook of Phonetics*, Cambridge University Press, pp.336–361, 2021. DOI:10.1017/9781108644198.014.
[18] David Hason Rudd, Huan Huo, Guandong Xu, “Leveraged Mel Spectrograms Using Harmonic and Percussive Components in Speech Emotion Recognition”, *Advances in Knowledge Discovery and Data Mining: 26th Pacific-Asia Conference, PAKDD 2022, Proceedings*, Springer, Vol.13281, pp.392–404, 2022. DOI:10.1007/978-3-031-05936-0_31.
[19] Orhan Özhan, “Short-Time Fourier Transform”, in *Basic Transforms for Electrical Engineering*, Springer International Publishing, Cham, pp.441–464, 2022. DOI:10.1007/978-3-030-98846-3.
[20] Premjeet Singh, Goutam Saha, Md Sahidullah, “Non-Linear Frequency Warping Using Constant-Q Transformation for Speech Emotion Recognition”, *Proceedings of the 2021 International Conference on Computer Communication and Informatics (ICCCI)*, IEEE, pp.1–6, 2021. DOI:10.1109/ICCCI50826.2021.9402569.
[21] Zrar Kh. Abdul, Abdulbasit K. Al-Talabani, "Mel Frequency Cepstral Coefficient and Its Applications: A Review", IEEE Access, Vol.10, pp.122136–122158, 2022. DOI:10.1109/ACCESS.2022.3223444.
[22] Selva Birunda, S., Kanniga Devi, R., "A Review on Word Embedding Techniques for Text Classification", Innovative Data Communication Technologies and Application (Proceedings of ICIDCA 2020), Springer, pp.267–281, 2021. DOI:10.1007/978-981-15-9651-3_23.

Authors’ Profiles

Rania A. Anwer received her BSc in Electrical Engineering from the Communication and Electronics department at Zagazig University, Shubra Faculty of Engineering, in 2004. Her research interests include software engineering, Machine Learning, Big Data, and Security.

Mahmoud M. Hussein: received his BSc. and MSc. in Computer Science from Menoufia University, Faculty of Computers and Information in 2006 and 2009, respectively, and received his PhD in Software Engineering from Swinburne University of Technology, Faculty of Information and Communications Technology in 2013. His research interest includes Software Engineering, Data Mining, Machine Learning, Data Privacy, and Security

Arabi Keshk: received the B.Sc. degree in electronic engineering and the M.Sc. degree in computer science and engineering from the Faculty of Electronic Engineering, Menoufia University, in 1987 and 1995, respectively, and the Ph.D. degree in electronic engineering from Osaka University, Japan, in 2001. His research interests include software testing, software engineering, distributed systems, cloud computing, machine learning, the IoT, big data analytics, and bioinformatics.

How to cite this paper: Rania Ahmed, Mahmoud Hussein, Arabi keshk, "Ensemble Fusion Model for Enhanced Speech Emotion Recognition and Confusion Resolution", International Journal of Information Technology and Computer Science(IJITCS), Vol.18, No.1, pp.84-99, 2026. DOI:10.5815/ijitcs.2026.01.05