Article

Not peer-reviewed version

Combining Transformer, CNN, and LSTM Architectures: A Novel Ensemble Learning Technique That Leverages Multi-acoustic Features for Speech Emotion Recognition in Distance Education Classrooms

Eman Abdulrahman Alkhamali * , Arwa Allinjawi , Rehab Bahaaddin Ashari

Posted Date: 22 April 2024

doi: 10.20944/preprints202404.1456.v1

Keywords: Transformer, convolutional neural network, long short-term memory , speech emotion recognition, distance education, real-time, emotional stability, instructors

Article

Combining Transformer, CNN, and LSTM Architectures: A Novel Ensemble Learning Technique That Leverages Multi-Acoustic Features for Speech Emotion Recognition in Distance Education Classrooms

Eman Abdulrahman Alkhamali [1]*, Arwa Allinjawi [1] and Rehab Bahaaddin Ashari [1]

[1] Information Systems Department King Abdul-Aziz University, Jeddah, Saudi Arabia; eeedalenizi@stu.kau.edu.sa *, aallinjawi@kau.edu.sa, rashary@kau.edu.sa

Abstract: Speech emotion recognition (SER) is a technology that can be applied in distance education to analyze speech patterns and evaluate speakers' emotional states in real-time. It provides valuable insights and can be used to enhance the learning experience by enabling the assessment of instructors' emotional stability, a factor that significantly impacts information delivery effectiveness. Students demonstrate different engagement levels during learning activities, and assessing this engagement is an important aspect of controlling the learning process and improving e-learning systems. An important aspect that may influence student engagement is the emotional states of their instructors. Accordingly, this research uses deep learning techniques to create an automated system for recognizing instructors' emotions in their speech when delivering distance learning. This methodology entails integrating Transformer, convolutional neural network, and long short-term memory architectures into an ensemble to enhance SER. Feature extraction from audio data used Mel-frequency cepstral coefficients, chroma, Mel spectrogram, zero crossing rate, spectral contrast, centroid, bandwidth, roll-off, and root-mean square, with subsequent optimization processes adding noise to, conducting time stretching, and shifting the audio data. Notably, several Transformer blocks were incorporated, and a multi-head self-attention mechanism was employed to identify the relationships between the input sequence segments. The pre-processing and data augmentation methodologies significantly enhanced the precision of the results in that the model achieved accuracy rates of 96.3%, 99.86%, 96.5%, and 85.3% on the Ryerson Audio-Visual Database of Emotional Speech and Song, Berlin Database of Emotional Speech, Surrey Audio-Visual Expressed Emotion, and Interactive Emotional Dyadic Motion Capture datasets. Furthermore, it achieved 83% accuracy on another dataset created for this research—the Saudi Higher Education Instructor Emotions dataset. The results demonstrate this model's considerable accuracy in detecting emotions in speech data across different languages and datasets.

Keywords: Transformer; convolutional neural network; long short-term memory; speech emotion recognition; distance education; real-time; emotional stability; instructors

1. Introduction

Speech emotion recognition (SER) has the objective of detecting and comprehending emotions that are expressed as part of verbal communication. The field is currently undergoing significant development. Although humans express their feelings using, for example, body language, facial expressions, and speech, the latter is generally understood to be the most efficient interaction mode [1]. SER can be used to accurately pinpoint speakers' emotional states by processing the paralinguistic features within their speech signals. This presents transformative prospects, especially in human-machine interfaces [2], and including emotional understanding in these systems can promote more natural and seamless interactions.

Education is among the many fields to which SER can be applied, with emotions profoundly affecting student engagement and motivation, thereby impacting learning outcomes [3]. Several studies have shown that SER can be used to obtain important information about how students are feeling, especially in the distance learning context, and these results can inform adaptive teaching strategies. For example, if a student's speech signals indicate confusion or frustration, SER can prompt the educator to provide supplementary explanations or personalized aid [4]. Meanwhile, other studies have investigated how the emotional stability of instructors impacts how effectively they deliver information and the effect that this has on student engagement, information that can be used to improve teaching quality and student outcomes [5].

The field of deep learning (DL) has made significant progress in revolutionizing the ability of machines to comprehend and interpret human emotions. DL models can currently analyze intricate characteristics of unprocessed data—including voice recordings—to detect emotional cues. Emotional recognition technology has the potential to enhance human-machine interaction in a seamless and instinctive manner across a wide range of applications, and swift advancements in this domain underscore DL's profound capacity to facilitate machine recognition of and reactions to human emotions. Accordingly, the model proposed herein can significantly advance the field of SER by equipping machines with the human-like ability to discern emotions such as happiness, sadness, and anger in human speech [6]. By enabling machines to analyze subtle variations in tone, cadence, and timbre, DL allows for the deduction of emotional states from voice signals. The proposed model will allow for this in a manner that exceeds the capabilities of conventional methodologies. The technique is paramount for evaluating lecturers' emotional states because it enables the identification of potential stressors, anxiety, or emotional distress that might inhibit teaching efficacy. Upon detecting these emotional indicators, the SER technology can instigate suitable interventions, ranging from offering counseling to providing lecturers with appropriate support services, thereby ensuring that they reach their optimal teaching capacities [7]. Furthermore, the integration of SER in intelligent tutoring systems paves the way for more nuanced personalization of instructional content, tone, and feedback by fostering responsive and effective teaching-learning environments that are underpinned by a deeper understanding of emotional cues.

This study extracts Mel-frequency cepstral coefficients (MFCC), chroma, Mel spectrogram, zero crossing rate (ZCR), spectral contrast, centroid, bandwidth, roll-off, and root-mean square (RMS) from raw audio data, after which noise is added and the audio data is shifted and time-stretched. Then, an ensemble model comprising a combination of Transformer, convolutional neural network (CNN), and long short-term memory (LSTM) architectures is used to classify emotions from the static features in the data.

This paper makes several significant contributions to the field.

1. The creation and use of the Saudi Higher Education Instructor Emotions (SHEIE) dataset: This dataset is a distinct resource for SER research because it includes meticulous annotations of emotions and specifically focuses on Saudi Arabian instructors, making it unique and useful in terms of SER in education.
2. The undertaking of a comprehensive series of experiments: The study explores the effectiveness of various data augmentation (DA) and feature extraction techniques for speech emotion classification. Evaluation steps are conducted using five benchmark datasets, namely, the Ryerson Audio-Visual Database of Emotional Speech and Song, (RAVDESS), the Berlin Database of Emotional Speech (EMO-DB), the Surrey Audio-Visual Expressed Emotion (SAVEE), the Interactive Emotional Dyadic Motion Capture (IEMOCAP), and the SHEIE.
3. The proposal of a new model for classifying emotions from speech: This model uses a Transformer architecture that incorporates multi-head self-attention. Furthermore, the favorable attributes of CNN and LSTM networks are combined to extract spectral features and determine temporal dynamics.

The rest of the paper is organized as follows: Section 2 discusses the pertinent literature, Section 3 outlines the methodology adopted, Section 4 details the experimental outcomes, Section 5 discusses the research findings, and Section 6 concludes the paper and delineates future research avenues.

2. Literature Review

Various studies have used machine learning (ML) and DL methodologies to classify emotions in speech. Recognizing emotion from speech signals represents a significant yet complex aspect of human-computer interaction [8,9]. Furthermore, numerous approaches have been used in the field of SER to identify emotions from speech signals, including employing well-known methods for speech analysis and classification. For example, feature extraction involves transforming speech waveforms into parametric representations, which may include the use of statistical features or a spectrogram. Recognizing emotions can also encompass retrieving diverse characteristics from speech signals, for example, paralinguistic qualities such as pitch, intensity, and MFCC. For a more complete SER analysis, several scholars have proposed integrating prosodic and spectral data. Although substantial research has been conducted on this topic, it remains difficult for machines to discern between emotions.

2.1. SER Applications in Education

Recent studies have explored the application of SER in remote education settings. For instance, Zhu and Luo [10] detailed a new approach to recognizing emotions in speech in an effort to address emotion deficiency in e-learning systems. Their study focused on developing an SER system that used neural networks to extract prosodic features from emotional utterances. Similarly, Tanko et al. [11] proposed an automated speech emotion polarization model capable of evaluating lecturers on distance education platforms, with orbital local binary patterns and multi-level wavelet transforms employed for feature extraction and classification. Meanwhile, Chen et al. [12] analyzed the emotional deficiency in current e-learning systems and proposed a model that incorporates SER to enhance teacher-student connections by tracking changes in learners' emotional states through their speech cues, allowing for teaching strategies to be adjusted accordingly. Taking a similar approach, Huang et al. [13] discussed various techniques for transferring speech emotion classifiers trained on acted emotional data to naturalistic elicited data using online learning. This involved adopting a boosting algorithm that incrementally retrains models on unlabeled real-world data. Elsewhere, Bahreini et al. [4] developed and evaluated a real-time SER system that uses microphones to provide feedback in e-learning environments. This system's classification of basic emotional states significantly resembled human expert raters' classifications. Furthermore, Li et al. [14] constructed an e-learning model that incorporates SER and neural networks to track learners' emotional states, with the resulting system providing encouragement and advice that is tailored to the emotions it recognizes. Additionally, the method developed by Tanko et al. [5] automatically detects presenters' emotions from course materials, resulting in the researchers creating a lecture transcript database to study speaker personalities. The data were halved every 5 seconds, yielding 9,541 observations. This study also produced the Shoelace Pattern, a shoelace-inspired local feature generator for feature extraction. The Shoelace Pattern's sub-bands were created using a tunable q-wavelet transform to ensure enhanced performance. To create the final feature vector, the feature extraction model combined the top four performing feature vectors. After selecting the 512 most useful attributes using neighborhood component analysis, a support vector machine (SVM) with a 10-fold cross-validation scheme was used to categorize the data. The classifiers were determined to have accuracy rates of 94.97% and 96.41%. Finally, the SER method for educational settings introduced by Zhang and Srivastava [15] featured kernel canonical correlation analysis and SVMs and accuracy of over 90%. Therefore, the literature indicates that SER techniques could significantly enhance remote education by detecting emotional cues in speech.

2.2. Feature Extraction for Classification in SER

Ancilin and Milton [16] introduced a new feature extraction technique called Mel frequency magnitude coefficient (MFMC) for SER, which they comprehensively evaluated against conventional features such as MFCC, Log Frequency Power Coefficient (LFPC), and linear prediction cepstral coefficients (LPCC) using six diverse emotion databases: EMO-DB, RAVDESS, SAVEE, EMOVO, eNTERFACE, and Urdu. They used MFMC with an SVM classifier, which resulted in competitive accuracy of 95.25% for the Urdu Language Speech Dataset, 81.50% for the EMO-DB, 75.63% for the SAVEE, 73.30% for the EMOVO, 64.31% for the RAVDESS, and 56.41% for the eNTERFACE database. This extensive benchmarking highlights the potential of MFMC techniques in terms of SER across languages and for various uses. Guan et al. [17] focused on local dynamic features for SER. This saw them extract novel time-based segmentation and pitch probability distribution features from the EMO-DB database, including examples of seven emotions. Experiments were also conducted on global features including MFCC, ZCR, energy, and pitch. The results indicated that the local dynamic pitch features outperformed the global features by achieving 70.8% accuracy compared to the 66.4% achieved by the global features with an SVM classifier. This suggests that local dynamic pitch carries discriminative information that can be used to recognize emotions in speech. Similarly, Alsabhan [18] proposed an end-to-end CNN with LSTM attention mechanisms for multilingual SER, evaluating it using the multiple-language-spanning SAVEE, Arabic Natural Audio Dataset (ANAD), Basic Arabic Vocal Emotion(BAVED), and EMO-DB datasets. The model achieved accuracies of 97.13%, 96.72%, 88.39%, and 96.72% for these databases, outperforming even custom two-dimensional CNNs. Not only do these results highlight the efficacy of one-dimensional (1D) CNNs in combination with LSTMs, but they also attend to learning discriminatory representations in SER across languages. Furthermore, responding to the limited real-world emotional speech data, Atmaja and Sasou [19] analyzed DA techniques for SER by experimenting with, for example, pitch shifting and time stretching on the Japanese Twitter and IEMOCAP datasets. With wav2vec 2.0 speech embeddings and an SVM classifier, their DA approach improved accuracy by 77.25% for these datasets. Therefore, using DA proved valuable in the context of a lack of naturalistic emotional speech data.

Elsewhere, Zehra et al. [20] used ensemble learning for multilingual cross-corpus SER by combining spectral and prosodic features using an ensemble of multiple classifiers. From this, they achieved an accuracy above 96.75% for the Urdu dataset, outperforming even the most state-of-the-art approaches. This demonstrates the ability of ensemble techniques to learn complementary information from different features and classifiers, thereby improving SER performance across domains. Meanwhile, Parthasarathy and Busso employed an unsupervised auxiliary task in ladder networks that were used for emotional recognition [21], with emotion prediction through regression being the primary assignment. As a side task, denoising autoencoders were employed to recreate intermediate feature representations. Thus, the framework was trained in a semi-supervised manner using a large amount of unlabeled data from the domain of interest. This approach is more effective than fully supervised single-task learning and multi-task learning (STL and MTL) baselines because within-corpus evaluations have determined an increase from 3.0% to 3.5% in the concordance correlation coefficient (CCC) as a result of this architecture, with cross-corpus evaluations indicating that the CCC increased from 16.1% to 74.1%.

Soonil and Mustaqeem [22] introduced a clustering-based framework for SER that leverages learned features and a deep BiLSTM network to enhance recognition accuracy while reducing computational complexity. The framework employs a novel sequence selection strategy using Radial Basis Function Network (RBFN) for similarity measurement within clusters, selecting key segments that represent the emotional content of the speech. Spectrograms generated via the STFT algorithm are used to extract discriminative features with a pre-trained Resnet101 CNN model, which are then normalized and processed by a deep BiLSTM to capture temporal information. This approach has demonstrated robustness and effectiveness, outperforming state-of-the-art methods with accuracies of up to 72.25%, 85.57%, and 77.02% on the IEMOCAP, EMO-DB, and RAVDESS datasets, respectively. Yu and Xizhong [23] proposed an Attention-Augmented Convolutional Block Gated Recurrent Unit network to improve SER, which they tested using the IEMOCAP dataset. The model was employed to capture the spectrogram and first- and second-order derivative aspects of audio signals, and the residual blocks allowed the CNN to extract spatial characteristics from the inputs. The block-gated recurrent unit featured an attention layer and was used to mine the long-term data. Compared to existing methods, this network improved Weighted Accuracy (WA) and Unweighted Accuracy (UA accuracy) by 82%, demonstrating the effectiveness of using attention-augmented convolutional-recurrent networks for SER.

In another study by Ahmed et al [24] five English and German benchmark datasets were used to present an ensemble DL framework for SER. The ensemble used a 1D CNN-FCN (fully convolutional network) model to extract local features, a 1D CNN-LSTM-FCN model to capture long-term dependencies, and a 1D CNN-gated recurrent unit (GRU)-FCN model. Noise injection, pitch shifting, and temporal stretching were employed to boost sample sizes. The MFCC, chromogram, least mean square(LMS), RMS, and ZCR features were input into these models, and the ensemble model was found to outperform the previous techniques with a significant weighted average

doi:10.20944/preprints202404.1456.v1

accuracy of 99.46% for the Toronto emotional speech set dataset, 95.42% for the EMO-DB, 95.62% for the RAVDESS, 93.22% for the SAVEE, and 90.47% for the Crowd Sourced Emotional Multimodal Actors dataset. This ensemble framework uses multiple DL architectures and DA, which enables robust SER across various languages.

Thus, the research indicates that SER techniques have the significant potential to improve instructor-student connections, student well-being, and personalized learning in distance education environments. For example, detecting emotional cues such as prosody, pitch, and intensity in speech enables the adaptation of educational materials and experiences to improve student engagement. Importantly, SER is an unobtrusive and passive tool that can be used to analyze voice data during regular online learning activities. Moreover, advances in ML and multimodal affective computing will continue to improve the accuracy of SER for remote education. Potential future directions include the detection of more complex emotions, the analysis of conversational dynamics, and the combination of speech analysis with expressions. These findings also imply that combining feature extraction and modeling procedures can increase accuracy across datasets. However, the generalizability of SER models across languages, cultures, and application sectors needs further research; despite these considerable advances, models that can properly distinguish emotions in speech across various cultures and circumstances remain in demand.

3. Materials and Methods

This study proposes an automated system that can accurately recognize instructors' emotional states during remote instructional sessions, improving the results of the evaluations conducted by their institutions. This system is developed sequentially: data collected during normal lectures is subjected to pre-processing and feature extraction, enabling the model's development using advanced technologies such as DL, which combines Transformer, CNN, and LSTM architectures. The usefulness of the model is tested after training it on a standard dataset, as shown in Figure 1.

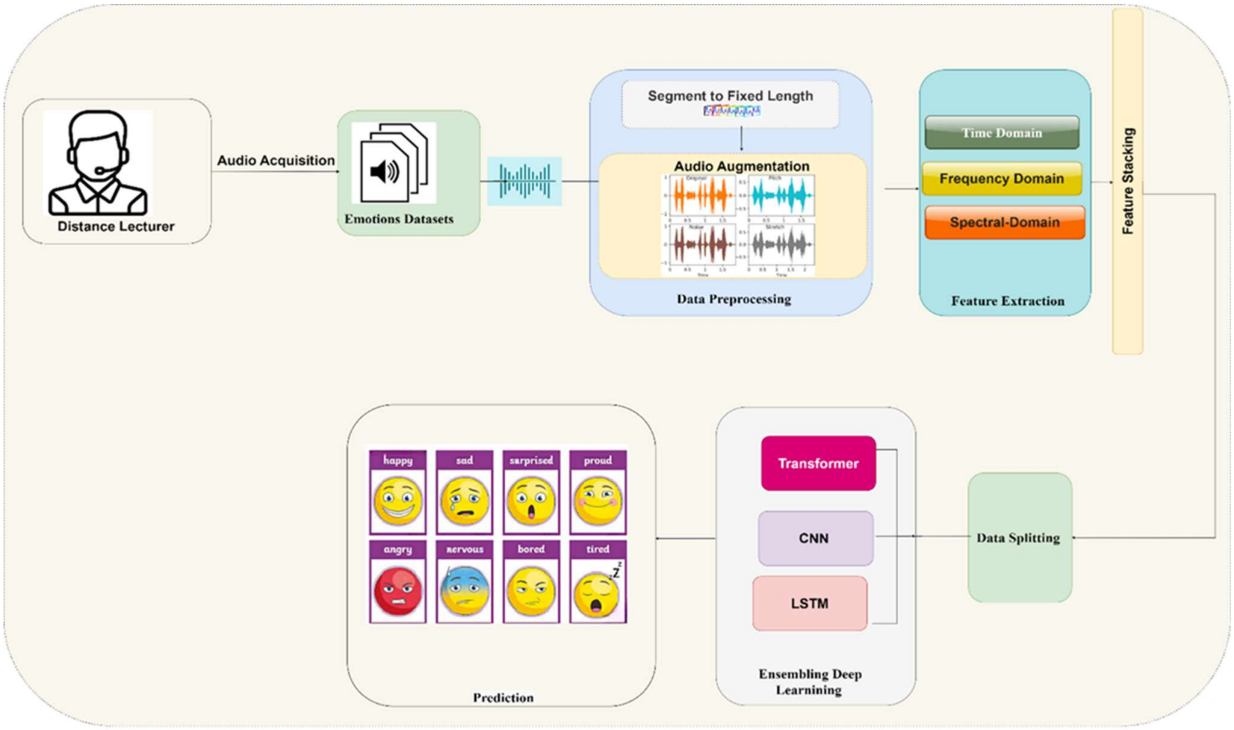


Figure 1. Proposed method.

3.1. Datasets

To ensure an exhaustive appraisal of the proposed model, it is tested on five datasets spanning three languages: English, Arabic, and German. To comprehensively train models using DL, a certain sample size is needed, which this study does not meet. Therefore, DA is applied to all the datasets. The following paragraphs summarize each of the datasets employed here.

The **RAVDESS** is a validated multimodal database containing emotional speech and song recordings. The database comprises 7,356 audio files recorded by 24 professional actors, divided equally between men and women. The actors were recorded speaking two linguistically neutral statements while adopting a neutral or standardized accent to minimize the influence of regional variations on the emotional content of the speech. By having professional actors record these controlled statements, high-quality emotional portrayals devoid of confounding regional accents were obtained. RAVDESS provides a substantial corpus of emotional speech and song for research and development in fields such as affective computing. The standardized validation protocols embedded in the dataset aid in comparative benchmarking and reproducibility across studies. Additionally, the spectrum of emotions conveyed by the speech in this dataset encompasses an array of affective states, including but not limited to tranquility, satisfaction, elation, melancholy, indignation, trepidation, astonishment, and revulsion [25]. Each expression is associated with a distinct level of emotional intensity, ranging from neutral to normal to strong.

The **EMO-DB** was created by the Technical University of Berlin Institute of Communication Science [26]. It is a free German emotional speech database [26] that includes 535 context-variable sentences used in everyday communication delivered by 10 expert actors (five men and five women) who, while speaking, simulated either happiness, anger, anxiety, fear, boredom, disgust, or neutrality. The 48-kHz data were down-sampled to 16-kHz.

The **SAVEE** dataset comprises 480 utterances in British English recorded by four male postgraduate students and researchers at the University of Surrey, all of whom were native English speakers aged between 27 and 31 [27]. Seven different emotions were expressed through these utterances (happiness, sadness, surprise, fear, disgust, neutral, and anger), with sentences from the phonetically balanced Texas Instruments/Massachusetts Institute of Technology corpus chosen for each emotion (See Table 1).

Table 1. Description of the datasets.

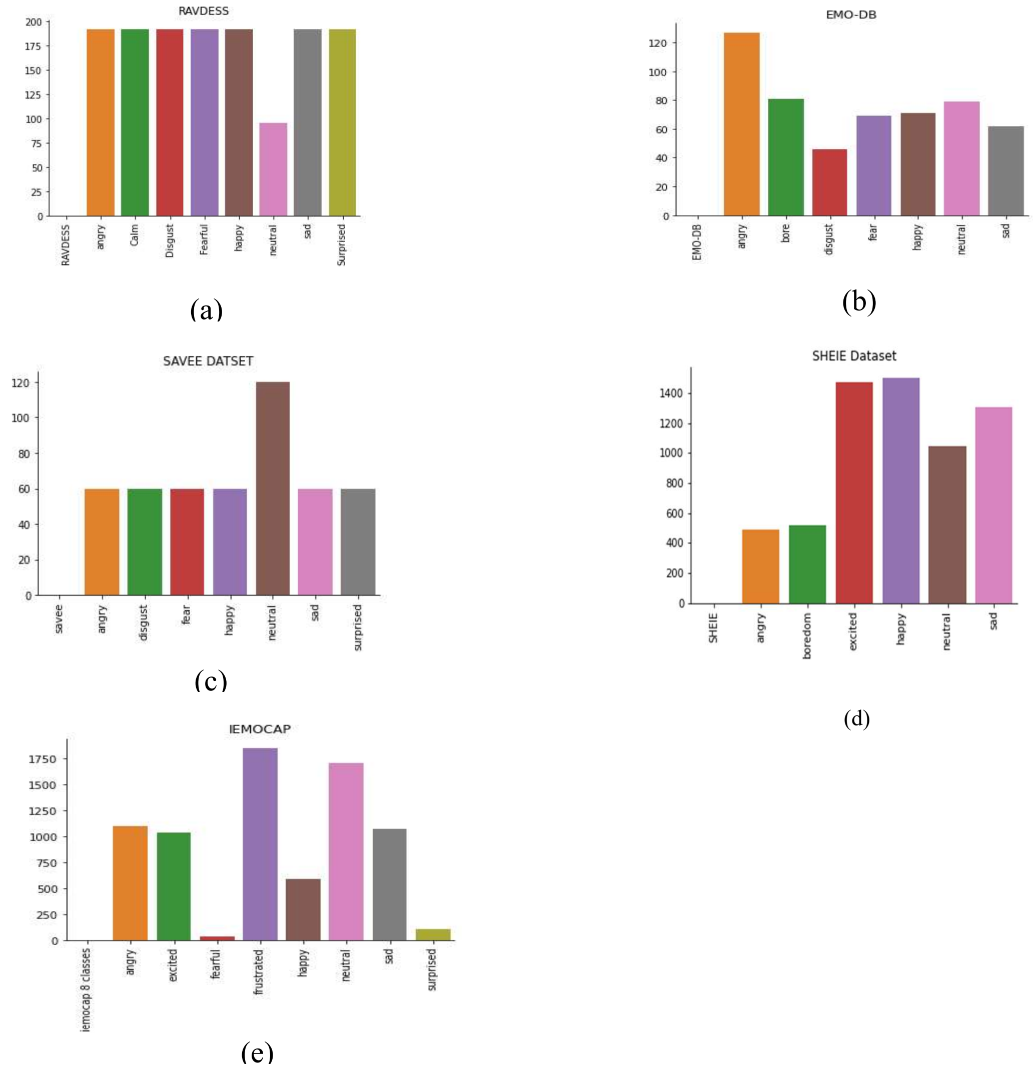
Dataset Name	Number and Type of Emotions	Sample Size	Type of Dataset	Status
Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)	Eight: disgust, fear, sadness, angriness, happiness, surprise, calmness, neutrality	7,356	Multimodal: speech and song	Acted
Berlin Database of Emotional Speech (EMO-DB)	Seven: anger, disgust, fear, happiness, neutrality, sadness, surprise, boredom	535	Speech	Acted
Surrey Audio-Visual Expressed Emotion (SAVEE)	Seven: anger, disgust, fear, happiness, neutrality, sadness, surprise	480	Speech	Acted
Interactive Emotional Dyadic Motion Capture (IEMOCAP)	Eight: happiness, anger, sadness, frustration, surprise, fear, excitement, neutrality	7,513	Multimodal (speech, video, motion capture)	Acted
Saudi Higher Education Instructor Emotions (SHEIE)	Six: anger, happiness, sadness, excitement, boredom, neutrality	7,515	Speech	Natural

doi:10.20944/preprints202404.1456.v1

The **IEMOCAP** was developed by the Signal Analysis and Interpretation Laboratory (SAIL) at the University of Southern California. This database represents an acted multimodal, multi-speaker collection of data. The dataset spans 12 hours and includes videos, audio, face tracking, and text transcriptions of paired performances in which actors try to elicit a particular feeling from the viewer using a combination of improvisation and prepared scenes. Multiple annotators contributed labels to the IEMOCAP database, classifying the data according to dimensional labels, including valence, activation, and dominance, as well as category labels, such as anger, happiness, sadness, and neutrality [28].

The **SHEIE** dataset is a unique emotional speech dataset developed to test the model proposed herein. It features real interactions from the realm of higher education and focuses on instructors. Recognizing the scarcity of datasets comprising genuine interactions, this dataset was designed to address the gap in existing speech emotion studies. It includes six universal emotions: anger, happiness, sadness, excitement, boredom, and neutrality. The data were collected from the various synchronous online lectures held for the Computer Science Department at King AbdulAziz University and the Islamic Studies Department at Al Jouf University, which were delivered in both Arabic and English. The dataset contains a total of 20 hours and 50 minutes of speech data from 19 lecture sessions. The audio data were collected by two instructors (one male, one female) and were carefully segmented and labeled. Developing the SHEIE dataset involved a four-step process. First, it was determined that real interactions rather than acted emotions would be recorded during live lectures via the Blackboard e-learning system. Second, the emotions represented in the dataset were selected based on their relevance to an instructor’s experience during a lecture. Third, volunteer instructors from the two universities were selected to participate in the study, and they provided the speech data via their lecture recordings. Finally, the emotions in the dataset were labeled based on self-reporting by instructors, which meant selecting an emotion from a prompt every ten minutes during their lectures. The time interval was chosen in response to research indicating that student attention begins to wane around the ten-minute mark. The SHEIE dataset underwent extensive pre-processing, which included splitting the data into ten-minute segments, manually labeling the emotions, cleaning the data to exclude noise and unwanted sounds, and normalizing the data into equal three-second intervals. This produced a total of 7,515 audio files, each labeled with a specific emotion and formatted for use in SER research: 490 files for anger, 1,888 files for happiness, 1,439 files for sadness, 516 files for boredom, 1,654 files for excitement, and 1,528 files for neutrality.

Figure 2 details the distribution of the different emotional classes in these five datasets. Some emotional classes occur more often in most datasets, while others demonstrate the inverse. Thus, there are imbalances in the emotional class distributions across the datasets, indicating the need for DA before model training. Table 1 presents a description of all five datasets as well as the distribution of the classes in these datasets.



(a)

(b)

(c)

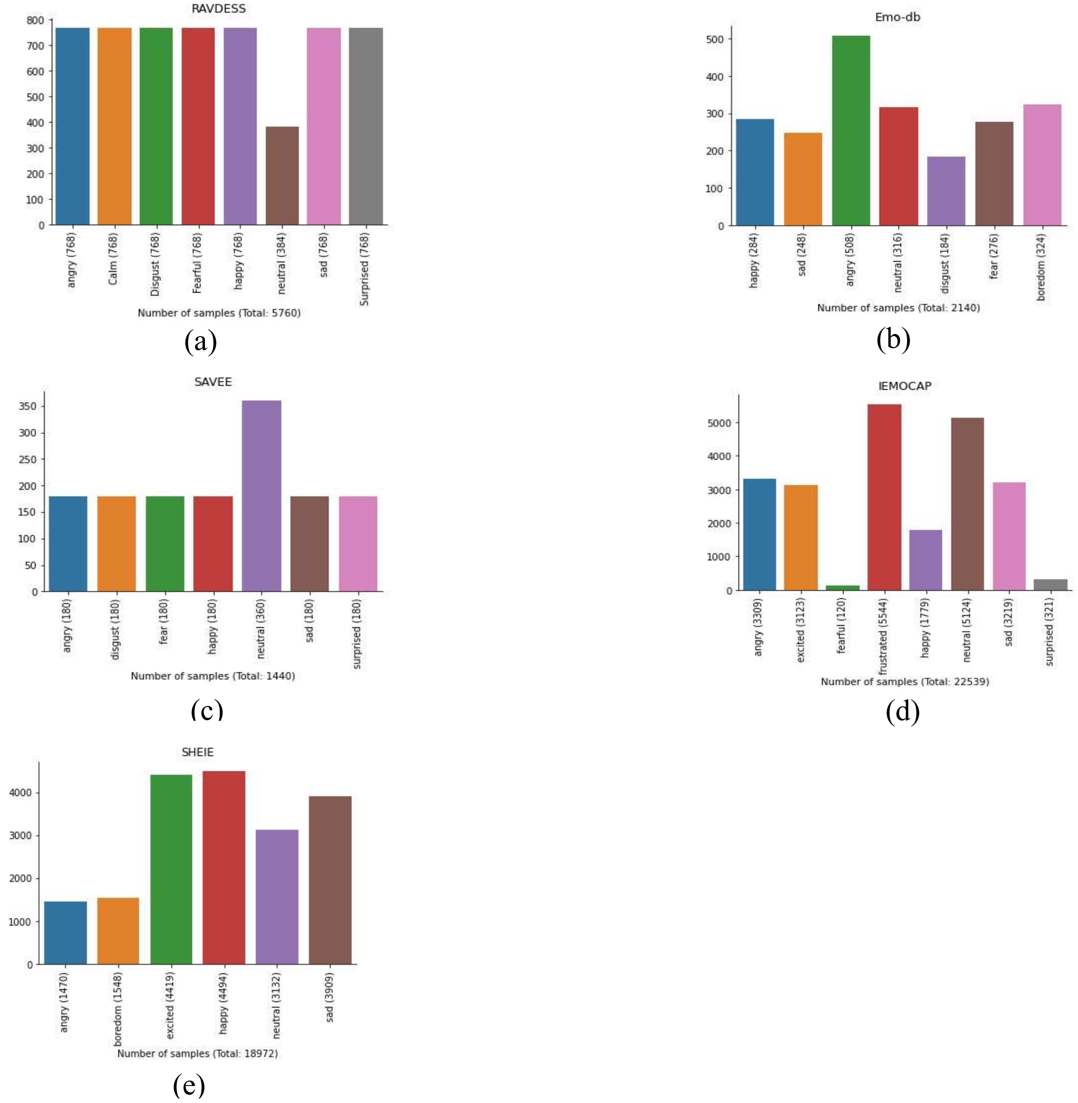
(d)

(e)

Figure 2. Distribution of the emotional classes in the five datasets: (a) RAVDESS; (b) EMO-DB; (c) SAVEE; (d) SHEIE; (e) IEMOCAP.

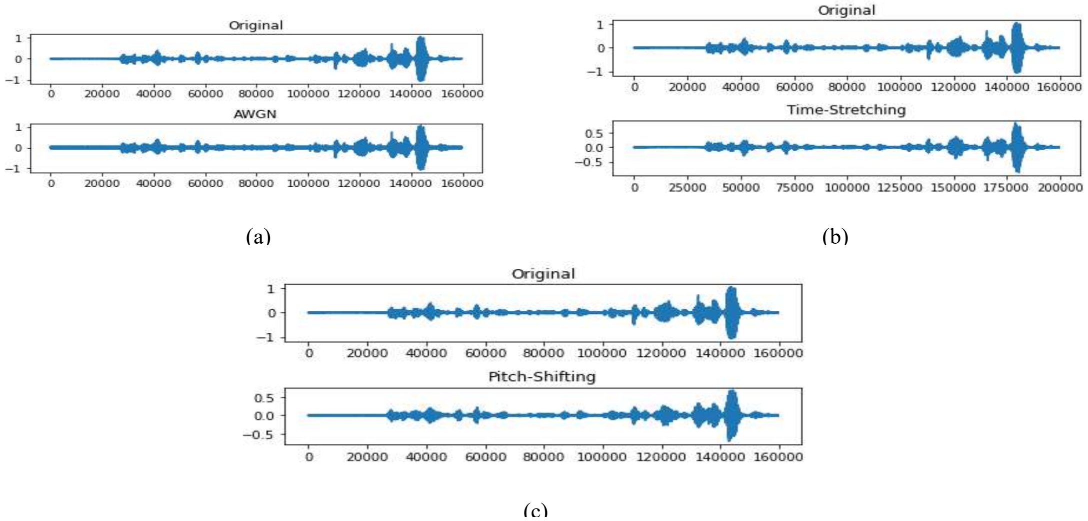
3.2. Data Augmentation

DA is essential for assessing SER model performance because SER systems often lack sufficient and diverse training data. This is observed in the imbalance between emotional classes in this study's datasets as shown in Figure 2. Therefore, speed, pitch, noise, and time stretching are used to generate variant data to give the model more context. This DA greatly improves the ability of SER systems to generalize and recognize emotions from speech under uncontrolled and varied conditions [29]. DA reduces overfitting, making the SER model more stable during the training process. Figure 3 illustrates the influence of DA on SER tasks for this study's five different datasets. DA techniques, such as adding additive white Gaussian noise (AWGN) to the samples, are employed to balance class distributions across these datasets, effectively addressing the imbalances in the class distributions. Furthermore, a custom noise function is used to add AWGN to the samples, as shown in Figure 4. This results in DA by adding random noise that is multiplied by 0.01 to the data. In addition, time stretching is used to stretch the data by the given rate, and the pitch shift function is applied with factors of 0.5 and 0.6. Time shifting is also employed using a custom shift function that incorporates the data, sampling rate, maximum shift value, and shift direction. A random shift value within the maximum shift value that is multiplied by the sampling rate is also generated. The shift value is negated for the right shift direction, and if the direction is labeled as "both," the direction is deemed to be random. The positive shift values set the first shift values of the DA to 0, and the last shift values are 0 if the shift values are negative, resulting in DA. Figure 4 indicates the changes in the waveforms after applying these DA techniques.



(a) (b) (c) (d) (e)

Figure 3. Distribution of the emotional classes in the five datasets after data augmentation: (a) RAVDESS; (b) EMO-DB; (c) SAVEE; (d) IEMOCAP; (e) SHEIE.



(a)

(b)

(c)

Figure 4. The effects of data augmentation on SER: (a) original audio with additive white gaussian noise; (b) original audio with time stretching; (c) Original sound with pitch shifting.

3.3. Feature Extraction

SER extracts features from speech waves to determine a person's emotional state because transforming speech into features or parameters allows for the determination of emotional characteristics. Features that can be extracted from the time domain include ZCR, energy, and amplitude, which can be used to identify anger and excitement by revealing speech rate and volume. Pitch, formant, and other features can be extracted from the frequency domain, with ZCR, chroma, and roll-off representing some of these features, as Figure 5 shown. Formants (the vocal tract's resonant frequencies) reveal the shape of the vocal tract, which determines the characteristics of articulated speech sounds. Pitch (the perceived fundamental frequency of a sound) can indicate emotions such as fear and surprise. MFCC, spectral contrast, and chroma also affect the spectral domain [30]. MFCC offer the best approximations of the human ear's nonlinear frequency perception and represent a sound's short-term power spectrum. Thus, they allow systems to more effectively understand emotions due to the systems acting like a human. This means that combining features from different domains guarantees a complete speech signal representation, allowing SER systems to identify emotions accurately and precisely.

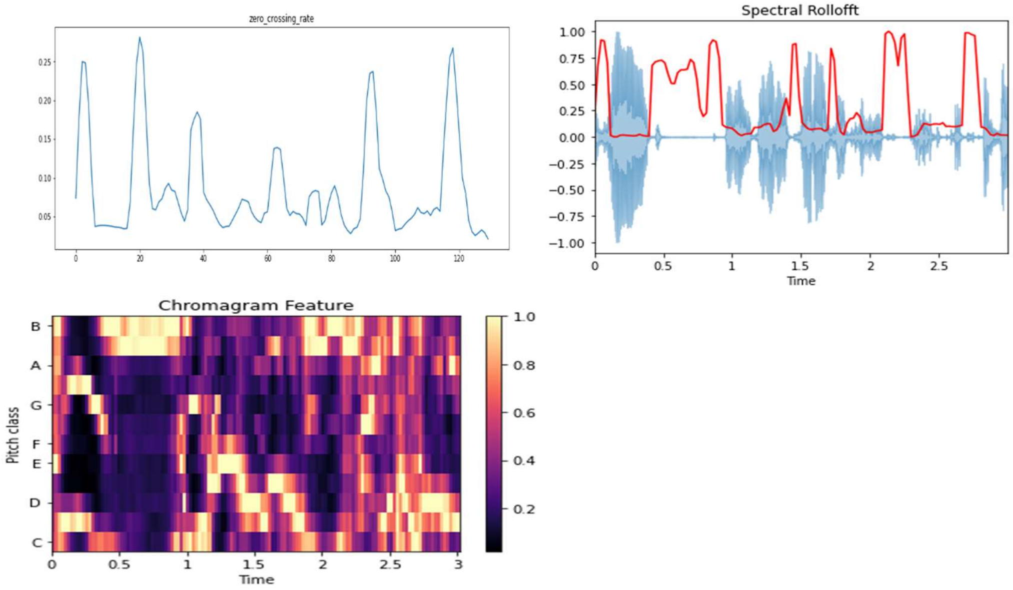


Figure 5. Zero cross rate (ZCR), chroma, and roll-off values for the anger sample.

doi:10.20944/preprints202404.1456.v1

ZCR measures how often a signal crosses the zero axis by determining signal sign changes per frame [31]. As such, it denotes the total number of times the wave flips from positive to negative, and it is contrariwise distributed by frame. Mathematically, ZCR can be determined using the following equation:

$$zcr = \frac{1}{T-1}\sum_{t=1}^{T-1} \mathbb{1}_{\mathbb{R}_{<0}}(s_t s_{t-1}), \qquad (1)$$

where s is a signal of length (T) and 1R<0 is a sign function.

A **chromatogram** visualizes audio by mapping frequencies onto 12 bins that match the 12 semitones of an octave using Chroma features [32]. This compresses pitch content into time windows, enabling music analysis applications to recognize chords and harmonic similarity despite timbre and instrumentation changes.

The **Mel spectrogram** visualizes audio signals by mapping frequencies onto the Mel scale, which aligns with human hearing. This technique captures important sound characteristics, facilitating its widespread use in speech and audio processing.

Spectral contrast, centroid, bandwidth, and roll-off are features extracted from sound signals. The differences in the levels between spectrum peaks and valleys can reveal a sound's timbre, while the spectrum's "center of mass" (also called the spectral centroid) can indicate sound brightness. Additionally, spectral bandwidth indicates a sound's spectral shape by measuring the spectrum's spread around its centroid, with spectral roll-off used to determine a sound's high-frequency content.

RMS is a feature that can be extracted from a speech wave and used in SER tasks [33]. It is used to measure the energy of a signal and provides information about the overall loudness of the signal.

MFCC is used in SER applications to parametrize speech signals generated by the Mel spectrogram. This scale matches the human auditory system more closely than linear frequency bands. To obtain the MFCC, the Mel spectrogram is transformed using discrete cosine transform (DCT) [26], [27] These coefficients capture speech signal characteristics and are resistant to timbre and instrumentation changes. This process entails determining the energy band of the speech wave, mapping the power spectrum onto the Mel scale using corresponding deltoid windows to obtain the Mel spectrogram logarithm, and applying the DCT to this logarithm to obtain the MFCC. The formula for mapping a frequency (f) in Hertz to a Mel frequency (m) is as follows:

$$m = 2595\log_{10}(1 + \frac{f}{700}). \qquad (2)$$

The inverse formula for mapping a Mel frequency (m) to a frequency (f) in Hertz is:

$$f = 700(10^{\frac{m}{2595}} - 1). \qquad (3)$$

Emotion classification studies typically use 40 MFCC coefficients for feature extraction, but a more nuanced representation of speech data using more coefficients can improve the detection of emotional states. MFCC, especially when combined with RMS and ZCR, has performed well in the complex task of SER [28,29].

3.4. Proposed Model

This study's methodology centers around the construction of an ensemble model that combines Transformer, CNN, and LSTM architectures. Transformer models use self-attention mechanisms to extract contextual features from input sequences, CNN filters extract local temporal features from input sequences, and LSTM models infer long-term dependencies from input sequences using recurrent connections. Therefore, in the study context, the Transformer self-focuses on the interconnectedness of the input elements regardless of sequence, the CNN layers identify the local audio feature patterns, and the LSTM layers capture and learn the long-term dependencies and temporal relationships within the audio sequences. LSTM layers use a series of gates (input gate, forget gate, and output gate) and a memory cell to selectively retain, update, or forget information from previous time steps, enabling them to effectively model and learn from long-term dependencies in the input sequences. The outputs of these three models are then merged and input into a dense final layer for classification, as Figure 6 shows. Thus, these architectures are combined to determine a feature of the input sequence that accounts for contextual, local temporal, and long-term dependencies. The SoftMax activation function in the dense layer classifies the combined feature representation to produce the final output. The transformer model features three Transformer block layers that contain multi-head self-attention layers and feedforward neural networks (FFNNs). The multi-head self-attention layer embeds 64 units and eight heads, and the FFNNs have 64-unit hidden layers with rectified linear unit (ReLU) activation functions. The CNN model comprises four Conv1D layers with 64 filters and three kernel sizes, with the ReLU activating each Conv1D layer. The flatten layer then flattens the last Conv1D layer outputs, which comprise three 64-unit layers. The first two LSTM layers then return the sequences, with the last layer returning the final hidden state. Then, the final output for emotion classification is generated by concatenating the outputs of the Transformer, CNN, and LSTM architectures and running them through a dense final layer comprising six to eight units and a SoftMax activation function.

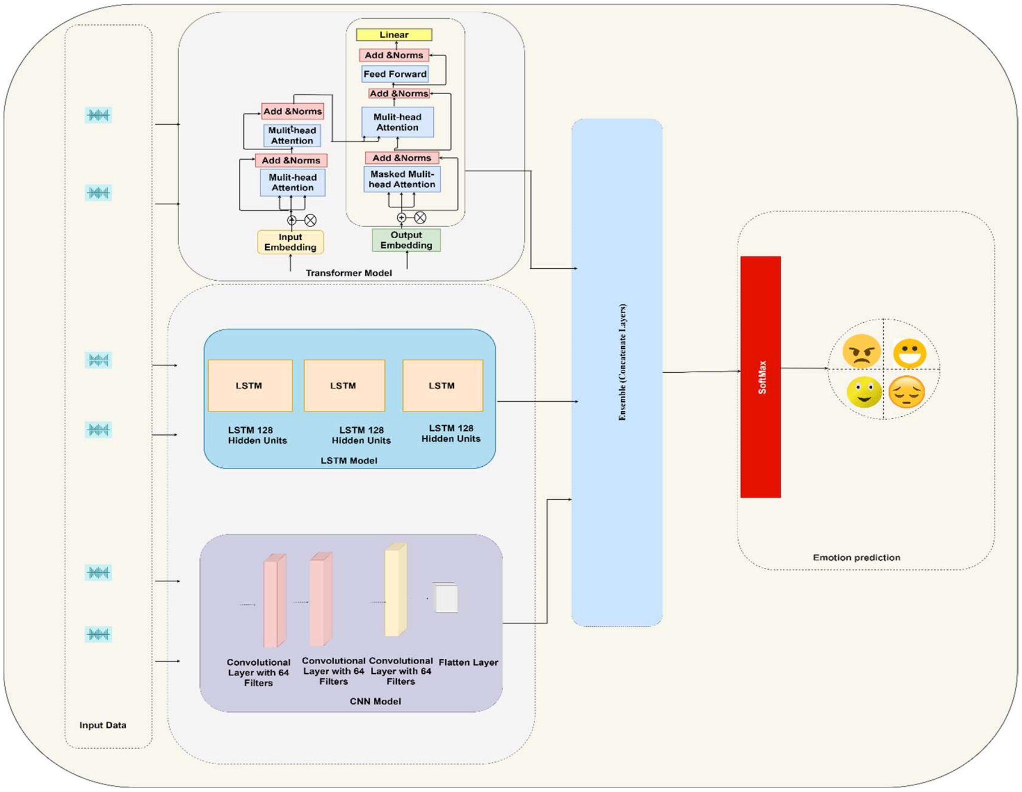


Figure 6. Ensemble model’s architecture.

1) *Transformer Block:* Transformer DL models use self-attention mechanisms to focus on different words in the input sequence when producing an output, with the Transformer block using multi-head self-attention to focus on different positions and understand data. FFNNs and dropout layers reduce overfitting, with layer normalization stabilizing learning [36].

The matrices W_Q, W_K, and W_V are subject to training and are capable of being modified by the learning process. Furthermore, attention scores (S) are calculated by multiplying the query and key matrices by the square root of their dimension using the following equation:

$$S = QK\text{^}T/sqrt(d\_k). \qquad (4)$$

W denotes the SoftMax (S), which yields the attention weights, and the attention layer output is a weighted sum of the value matrices: Z = WV.

The FFNN comprises a nonlinear activation function and two linear transformations: W_1, b_1, W_2, and b_2. These denote the adjustable weights and biases for the attention layer. The FFNN outputs are residually connected by adding the input to the output using the following equation:

$$\mathrm{Y} = \mathrm{X} + \mathrm{F(X)}. \qquad (5)$$

doi:10.20944/preprints202404.1456.v1

Layer normalization is the process by which the feature dimension input is normalized according to the attention layer and FFNN output using the following equation:

Y = (X – mean(X))/std(X). * gamma + beta (trainable scale and shift parameters).

The Transformer block in this model comprises an architectural component that accepts a feature vector of size (192, 1) as input (see Table 2). It consists of a multi-head self-attention mechanism and an FFNN, each of which employs residual connections and layer normalization. In addition, the multi-head self-attention's embed_dim and num_heads parameters represent the size of the input embeddings and the number of attention heads. Each attention head individually processes the input, allowing the model to concurrently learn different types of information from a singular input sequence. In this mechanism, the query_dense, key_dense, and value_dense are the dense layers that transform the inputs into their corresponding query, key, and value vectors. These vectors are further separated into different heads in the separate_heads method, ensuring parallel and independent computations for each head. The call method subsequently orchestrates the self-attention mechanism's computation flow by calling the previous components in order. After these computations, the combine heads dense layer merges the outputs from all the attention heads back into the original embedding dimension. The Transformer block also features an FFNN characterized by ffdim, which denotes the size of its hidden layer. To prevent overfitting, two dropout layers that are defined by rate are employed. The FFNN, denoted as ffn, essentially comprises two dense layers: The first applies a ReLU activation function; the second, which is the same size as the embed_dim, applies none. Post ffn, the output enters another dropout layer and residual connection before undergoing normalization using a second layer normalization layer. In summary, the Transformer block transforms the input embedding and outputs a transformed embedding of the same size that has tunable hyperparameters, such as the number of attention heads and the size of the hidden layer in the FFNN. To prevent overfitting, two dropout layers with a rate defined by the hyperparameter 'rate' are employed. The feed-forward network (FFN) in the Transformer block consists of two dense layers: the first one applies a "relu" activation function, and the second one, which has the same size as 'embed_dim', applies no activation function. The FFN is applied to the output of the multi-head attention mechanism, which computes the attention weights between all pairs of positions in the input sequence using multiple attention heads. The purpose of the FFN is to process the concatenated output from the different attention heads, allowing the model to capture more complex dependencies and transformations. After the FFN, the output traverses another dropout layer and undergoes a residual connection, followed by normalization using a second LayerNormalization layer. The Transformer block transforms the input embedding, outputting a transformed embedding of the same size, with tunable hyperparameters like the number of attention heads and the size of the hidden layer in the FFN.

Table 2. Ensemble model summary.

Layer	Output Shape	Param #	Connected to	Parameter	Value
Input_1 (InputLayer)	(None, 192, 64)	0			
conv1d (Conv1D)	(None, 192, 64)	256	input_1[0][0]	filters	64
transformer_block_1	(None, 192, 64)	0	input_1[0][0]	embed_dim	64
conv1d_1 (Conv1D)	(None, 192, 64)	12,352	conv1d[0][0]	kernel_size	3
				stride	
conv1d_2 (Conv1D)	(None, 192, 64)	12,352	conv1d_1[0][0]	padding	'same'
lstm (LSTM)	(None, 192, 64)	16,896	input_1[0][0]	activation	'relu'
transformer_block_2	(None, 192, 64)	25,216	transformer_block_1[0][0]	ff_dim	64
conv1d_3 (Conv1D)	(None, 192, 64)		conv1d_2[0][0]	rate	0.1
lstm_1 (LSTM)	(None, 192, 64)	33,024	lstm[0][0]	nodes	64
flatten (Flatten)	(None, 12288)	0	conv1d_3[0][0]	return_sequences	True
lstm_2 (LSTM)	(None, 64)	53,824	lstm_1[0][0]		
flatten_1 (Flatten)	(None, 12288)	0			

doi:10.20944/preprints202404.1456.v1

concatenate (Concatenate)	(None, 24640)	0	flatten_1[0][0], lstm_2[0][0]	concatenate_axis	1
dense_13 (Dense)	(None, 8)	197,128	concatenate[0][0]		

2) *CNN:* After the Transformer block, the CNN is used, which comprises two Conv1D layers. The (192, 1) feature vector feeds these layers. Each filter (f) is a vector of the weights of size (k), which denotes the kernel size. The output of the convolutional layer at position I is determined using the following equation:

$$(\mathrm{x} \ * \ \mathrm{f})(\mathrm{i}) \ = \ \mathrm{sum}(\mathrm{x}[\mathrm{i}+\mathrm{j}] \ * \ \mathrm{f}[\mathrm{j}]) \text{ for } \mathrm{j} \ = \ 0,\ldots,\mathrm{k}-1. \qquad (6)$$

If padding is utilized, the input sequence is expanded with zeroes before the filters are applied. Additionally, when a nonlinear activation function, such as ReLU, is employed, it is implemented on every individual element of the convolutional layer's output.

The Conv1D layers comprise 64 filters and three kernel sizes. The padding technique known as "same" is utilized to pad the sides of the input to achieve a width that matches that of the output. The ReLU activation function then introduces non-linearity to the model by producing the input value as an output if it is positive and zero if it is negative. The CNN layers then use filters to obtain the local features from the input vectors through convolution.

3) *LSTM:* The model's long-term LSTM layers process the sequence data, with the LSTM accepting the original (192, 1) input feature vector. The LSTM comprises two 64-unit layers. The first outputs its hidden state at every time step because the return sequences are labeled "True," and the text LSTM layer receives each output. This means that this setting is required when stacking LSTM layers. The second LSTM layer does not include this parameter, so it only returns the last output, which is then fed into dense layers to obtain the final predictions. Notably, LSTM networks can predict sequence data, remember information, and avoid the vanishing gradient problem of traditional recurrent neural networks.

For each time step (t), the LSTM simultaneously receives x_t from the input sequence as well as c_t-1 and h_t-1. The LSTM then calculates the input gate (i_t), an output gate (o_t), and a cell candidate (g_t) using various combinations of the present input, the preceding hidden state, and trainable weights according to the following equations:

$$\text{“}\mathrm{i\_t} \ = \ \mathrm{sigmoid}\left(\mathrm{W_i} * \left[\mathrm{h_{\{t-1\}}}, \mathrm{x_t}\right] + \mathrm{b_i}\right). \qquad (7)$$

$$\mathrm{f\_t} = \mathrm{sigmoid}\left(\mathrm{W_f} * \left[\mathrm{h_{\{t-1\}}}, \mathrm{x_t}\right] + \mathrm{b_f}\right). \qquad (8)$$

$$\mathrm{o\_t} \ = \ \mathrm{sigmoid}\left(\mathrm{W_o} * \left[\mathrm{h_{\{t-1\}}}, \mathrm{x_t}\right] + \mathrm{b_o}\right). \qquad (9)$$

$$\mathrm{g\_t} \ = \ \tanh\left(\mathrm{W_g} * \left[\mathrm{h_{\{t-1\}}}, \mathrm{x_t}\right] + \mathrm{b_g}\right). \qquad (10)$$

4. Experimental Results

The results from this study's model are compared to models employed by previous researchers, enabling evaluation of the different models based on established metrics. This study's SER system was scrutinized via speaker-independent experiments conducted on the five datasets. The data were partitioned using a percentage-based stratification approach, with 75% of the data used to train the SER models and the remaining 25% reserved for testing. Figure 17 depicts the performance of the Transformer model proposed. The model was trained using a feature set comprising 192 features from the five datasets. Table 8 visualizes the model's accuracy across the five datasets.

4.1. Experimental Design

Several packages and software programs were used, including TensorFlow and Librosa for audio pre-processing and WavePad for audio segmentation. The training and testing procedures were conducted using Google Collaboratory, a platform equipped with a 2.20 GHz Intel Xeon CPU, 25 GB of RAM, and Tesla graphics processing units (GPUs). The DL libraries of Keras and TensorFlow were used to develop and train the neural network models, and the GPU allowed for the efficient processing of the large matrix operations required for training. Keras provides a simple interface for building neural networks, while TensorFlow offers more flexibility for customizing and fine-tuning models. This research adopted Python as the implementation language for the proposed method.

4.2. Measuring Tools Used for the Evaluation

Several metrics were used to measure the performance of the SER model on the test set across the five datasets in the evaluation, including accuracy, loss, precision, F1-score, recall, confusion matrix, and receiver operating characteristic (ROC) curve [37].

Evaluation of each class [i] was conducted by quantifying true positives (TP_[i]), true negatives (TN_[i]), false positives (FP_[i]), and false negatives (FN_[i]). This process resulted in the determination of the performance metrics.

Accuracy is a metric used to determine the model's prediction accuracy by determining the ratio between accurate predictions and the sum of all predictions using the following equation:

$$\mathrm{Accuracy} = \frac{\mathrm{TP+TN}}{\mathrm{FP+FN+TP+TN}} \times 100. \qquad (11)$$

Precision is a metric that denotes the proportion of true positive predictions relative to all positive predictions. Determining this value involves calculating the proportion of true positives in relation to the combined total of true and false positives using the following equation:

$$\mathrm{Precision} = \frac{\mathrm{TP\_i}}{\mathrm{TP\_i+FP\_i}} \times 100. \qquad (12)$$

The **F1-score** provides a mean that balances precision and recall. It represents the weighted average of precision and recall that is determined using the following equation:

$$\mathrm{F1-score} = 2 * \frac{\mathrm{precisio}_{\mathrm{i}} \times \mathrm{Sensitivit}_{\mathrm{i}}}{\mathrm{precisio}_{\mathrm{i}} + \mathrm{Sensitivit}_{\mathrm{i}}} \times 100. \qquad (13)$$

The **recall metric** denotes the percentage of true positive predictions among all positives. It is defined as the relationship between accurate positives and the sum of true positives and false negatives. It is calculated as follows:

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}. \qquad (14)$$

The **confusion matrix** indicates a classification model's class prediction distribution. To calculate the precision, recall, sensitivity, and specificity, the confusion matrix evaluates TP, TN, FP, and FN, metrics used to identify model performance issues.

Finally, the **ROC curve** is a metric that plots the true positive rate (TPR) against the false positive rate (FPR) at different classification thresholds. The TPR and FPR are formally defined using mathematical expressions, with the TPR formula determining the ratio of true positives compared to the sum of true positives and false negatives by identifying the ratio of false positives compared to the sum of false positives and true negatives.

$$\mathrm{TPR} = \frac{\mathrm{TruePositives}}{\mathrm{TruePositives+FalseNegative}}. \qquad (15)$$

$$\mathrm{FPR} = \frac{\mathrm{FalsePositives}}{\mathrm{FalsePositives+TrueNegatives}}. \qquad (16)$$

4.3 Ensemble Model Evaluation

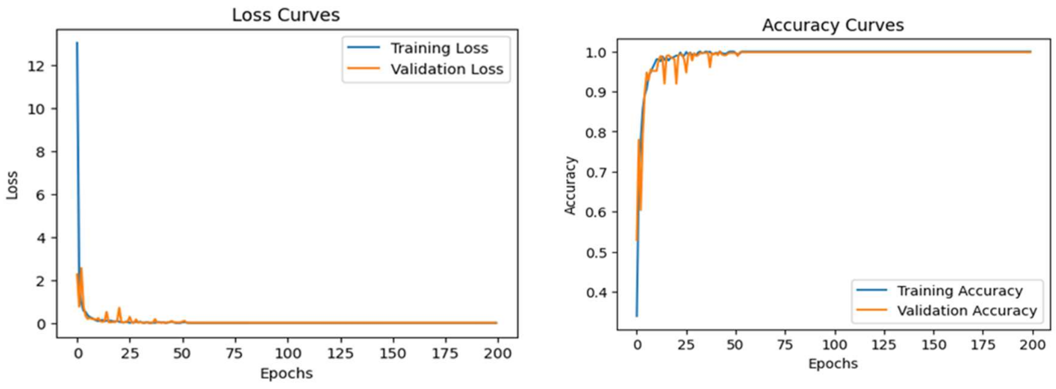
The evaluation process demonstrated that the ensemble model performed well across all datasets, thereby demonstrating its effectiveness in making accurate predictions. The following details the evaluation metrics for each dataset:

4.3.1. The EMO-DB Dataset

The ensemble model underwent 200 epochs of training with a batch size of 20, with the Adamax optimizer and sparse categorical cross-entropy loss being incorporated. The data were partitioned into the training and testing sets according to the prescribed ratio. The model was determined to have an average testing accuracy of 99.86%, a precision of 99.71%, a recall of 99.71%, and an F1-score of 99.71% for all emotions (see Table 3). Figure 7 details the loss and accuracy curves for this dataset. Figure 8 is the confusion matrix, which summarizes the model's performance in terms of classification by indicating how many predictions were correct and incorrect for the seven classes. In this metric, the diagonal lines of the matrix indicate the true positives for each class, whereas the off-diagonal elements represent the false positives and negatives.

Table 3. Ensemble model's performance on the EMO-DB dataset.

Emotion	Accuracy	Precision	Recall	F1-Score
Anger	100%	100%	100%	100%
Boredom	100%	100%	100%	100%
Disgust	100%	100%	100%	100%
Fear	99%	99%	99%	99%
Happiness	98%	100%	99%	99%
Neutral	100%	100%	100%	100%
Sadness	100%	99%	99%	99%
Average	99.86%	99.71%	99.71%	99.71%



(a) (b)

Figure 7. Ensemble model's loss and accuracy curves for the EMO-DB dataset.

doi:10.20944/preprints202404.1456.v1

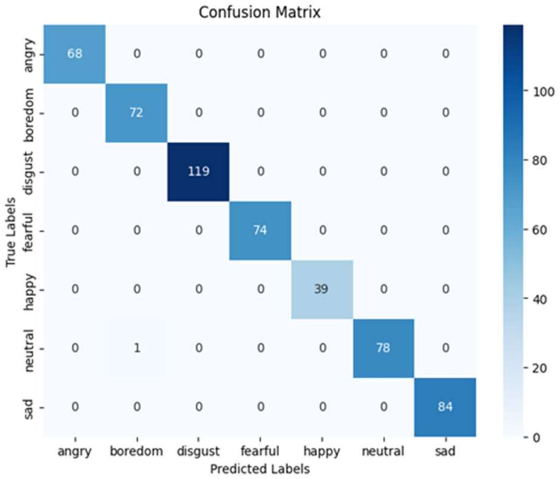


Figure 8. Ensemble model's confusion matrix for the EMO-DB dataset.

4.3.2. The RAVDESS Dataset

The ensemble model was trained over 200 epochs using a batch size of 20. The Adamax optimizer was also employed in conjunction with sparse categorical cross-entropy loss during the training process. The dataset was partitioned into distinct subsets for training and testing. The results indicate an average testing accuracy of 96.3%, a precision of 95.7%, a recall of 96.3%, and an F1-score of 95.9% for all emotions. These findings are presented in Table 4. Figure 9 details the loss and accuracy curves for this dataset, and Figure 10 demonstrates how many predictions were correct and false for the eight classes in the classification process.

Table 4. Ensemble model's performance on the RAVDESS dataset.

Emotion	Accuracy	Precision	Recall	F1-Score
Happiness	91%	97%	94%	94%
Sadness	89%	94%	92%	92%
Anger	93%	95%	94%	94%
Neutral	94%	90%	93%	92%
surprise	98%	99%	98%	98%
Calm	96%	96%	96%	96%
Fear	99%	92%	96%	94%
Disgust	95%	97%	96%	96%
Average	96.3%	95.7%	96.3%	95.9%

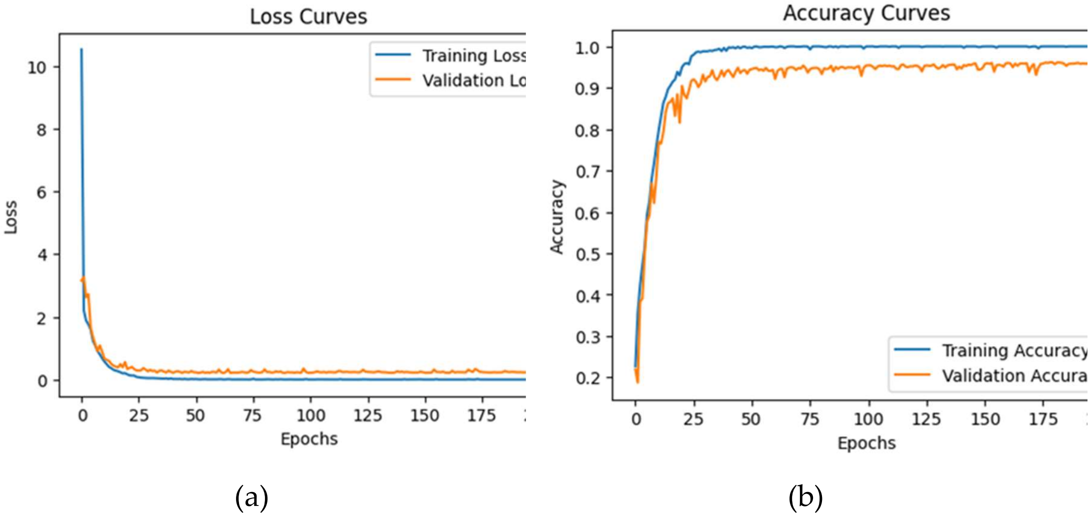


Figure 9. Loss and accuracy curves for the RAVDESS dataset.

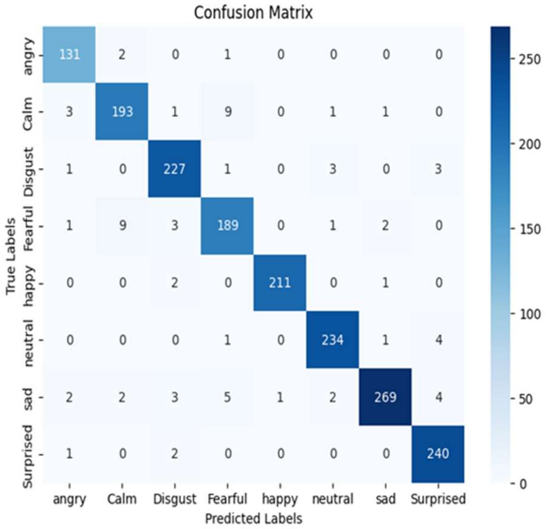


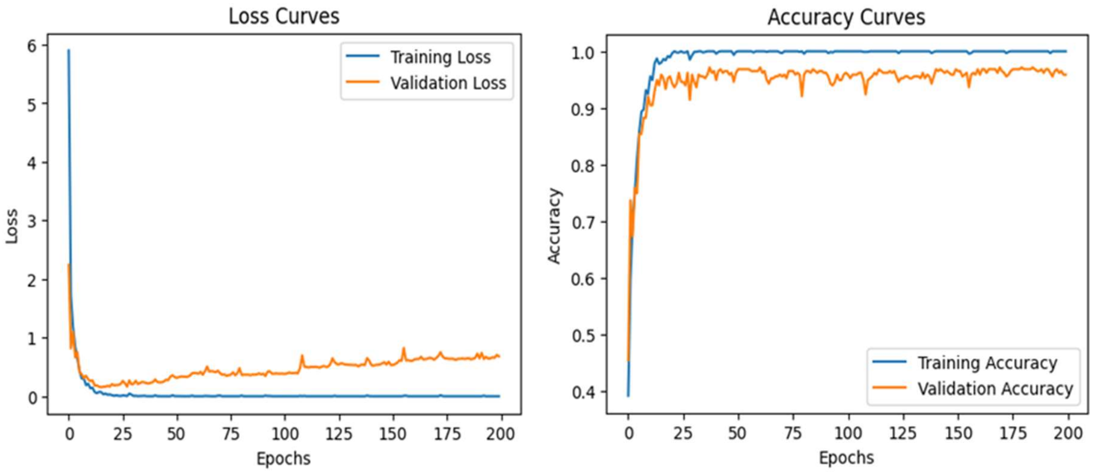
Figure 10. Ensemble model's confusion matrix for the RAVDESS dataset.

4.3.3. The SAVEE Dataset

The ensemble model was trained over 200 epochs using 20 batches in combination with the Adamax optimizer and sparse categorical cross-entropy loss. The dataset was split according to the 75/25 ratio discussed above. The model's average testing accuracy, precision, recall, and F1-score for all the emotions in this dataset were determined to be 96.5%, 95.3%, 95.6%, and 95.2% (see Table 5). Figure 11 visualizes the loss and accuracy curves for this dataset, and the confusion matrix is included as Figure 12.

Table 5. Ensemble model's performance on the SAVEE dataset.

Emotion	Accuracy	Precision	Recall	F1-Score
Anger	96.2%	95.2%	97.3%	96.3%
Disgust	100%	100%	99.4%	100%
Fear	82.8%	81.0%	99.9%	89.3%
Happiness	87.9%	86.4%	72.2%	77.6%
Neutral	106.2%	100%	100%	107.4%
Sadness	100%	100%	100%	100%
Surprise	99.2%	96.5%	93.0%	94.8%
Average	96.5%	95.3%	95.6%	95.2%



(a) (b)

Figure 11. Ensemble model's loss and accuracy curves for the SAVEE dataset.

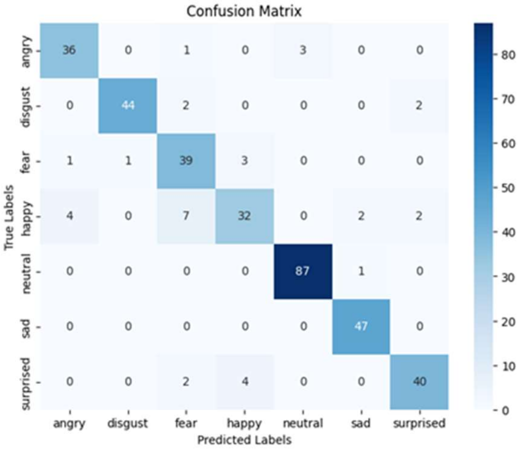


Figure 12. Ensemble model's confusion matrix for the SAVEE.

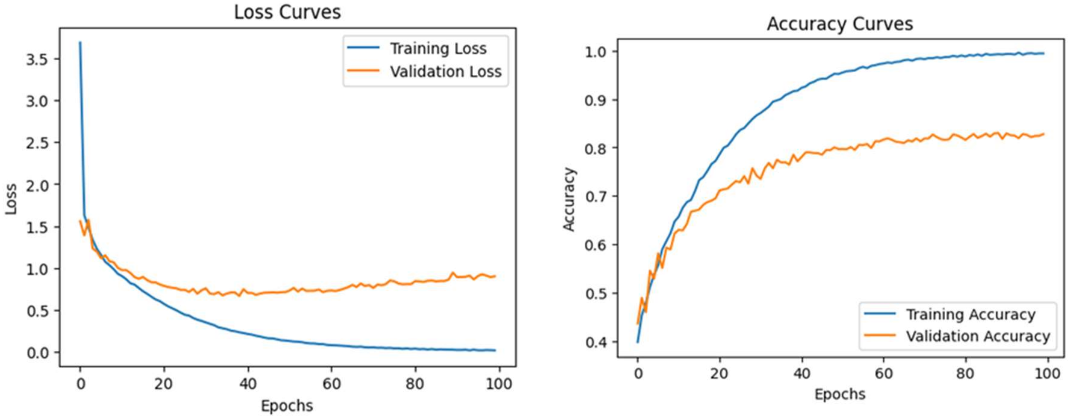
4.3.4. The IEMOCAP Dataset

After adjusting the classification thresholds to achieve the desired accuracy, the model demonstrated average accuracy of 85.3%, average precision of 94.1%, average recall of 61.4%, and an F1-score of 73.6% for all the emotions in the IEMOCAP dataset. These results are depicted in Table 6. While the precision and recall values were adjusted to achieve the desired accuracy, the overall F1-score reflects the trade-off between precision and recall. Figure 13 shows the loss and accuracy curves for the IEMOCAP dataset, with the confusion matrix for the emotion classification model appearing as Figure 14. The neutral and surprise classes recorded the highest accuracy 95.0%, while excitement recorded the highest precision at 99.0%. Neutral and surprise produced the highest recall rates 92.0%, and surprise produced the highest F1-score 94.0%.

Table 6. Ensemble model's performance on the IEMOCAP dataset.

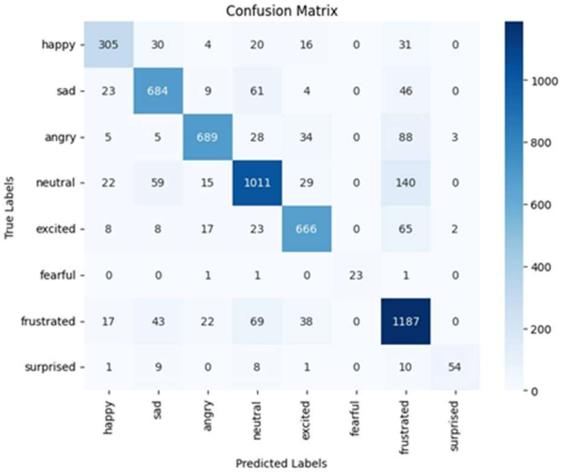
Emotion	Accuracy	Precision	Recall	F1-Score
Anger	80.0%	95%	50.0%	65.0%
Excited	90.0%	99%	70.0%	81.8%
Fearful	85.0%	92%	60.0%	72.7%
Frustrated	80.0%	87%	40.0%	54.8%
Happy	85.0%	97%	58.0%	72.8%

Neutral	95.0%	90%	92.0%	91.0%
Sadness	85.0%	94%	60.0%	72.7%
Surprised	95.0%	96%	92.0%	94.0%
Average	85.3%	94.1%	61.4%	73.6%



(a) (b)

Figure 13. Ensemble model’s loss and accuracy curves for the IEMOCAP dataset.



	happy	sad	angry	neutral	excited	fearful	frustrated	surprised
happy	305	30	4	20	16	0	31	0
sad	23	684	9	61	4	0	46	0
angry	5	5	689	28	34	0	88	3
neutral	22	59	15	1011	29	0	140	0
excited	8	8	17	23	666	0	65	2
fearful	0	0	1	1	0	23	1	0
frustrated	17	43	22	69	38	0	1187	0
surprised	1	9	0	8	1	0	10	54

Figure 14. Ensemble model’s confusion matrix for the IEMOCAP dataset.

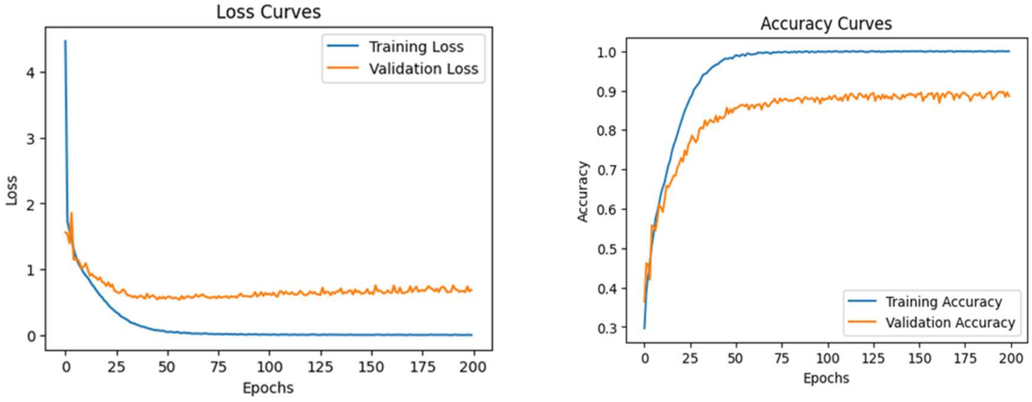
4.3.5. The SHEIE Dataset

The ensemble model was trained using 200 epochs, 20 batches, the Adamax optimizer, and sparse categorical cross-entropy loss. The model demonstrated an average accuracy of 83%, an average precision of 83%, an average recall of 81.17%, and an average F1-score of 81.71% across all emotions (see Table 7). Figure 15 displays the loss and accuracy curves for this dataset, and Figure 16 represents the confusion matrix, indicating the predicted and actual outcomes for each emotional class. These results demonstrate the effectiveness of the proposed ensemble model in terms of accurately classifying the emotions in the SHEIE dataset.

Table 7. Ensemble model’s performance on the SHEIE dataset.

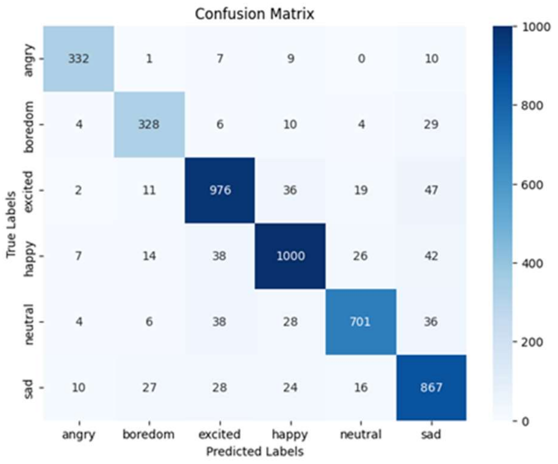
Emotion	Accuracy	Precision	Recall	F1-Score
Anger	84.13%	84.13%	86.07%	85.10%
Boredom	81.07%	81.07%	73.33%	76.92%
Excitement	84.13%	84.13%	85.71%	84.91%
Happiness	82.10%	82.10%	85.54%	83.75%

Neutral	80.2%	80%	77.8%	78.3%
Sadness	79.87%	79.87%	79.87%	79.87%
Average	83%	83%	81.17%	81.71%



(a) (b)

Figure 15. Ensemble model's loss and accuracy curves for the SHEIE dataset.



	angry	boredom	excited	happy	neutral	sad
angry	332	1	7	9	0	10
boredom	4	328	6	10	4	29
excited	2	11	976	36	19	47
happy	7	14	38	1000	26	42
neutral	4	6	38	28	701	36
sad	10	27	28	24	16	867

Figure 16. Ensemble model's confusion matrix for the SHEIE dataset.

Table 8. The Ensemble model's overall accuracy.

Dataset	Model	Testing Accuracy	F1-Score
EMO-DB	Ensemble model	99.86%	99.71%
RAVDESS	Ensemble model	96.3%	95.9%
SAVEE	Ensemble model	96.5%	95.2%
IEMOCAP	Ensemble model	85.3%	73.6%
SHEIE	Ensemble model	83%	81.71%

5. Analysis and Discussion

Evaluation of the efficacy of the proposed model involved conducting experiments using not only the diverse EMO-DB, RAVDESS, SAVEE, and IEMOCAP datasets but also the purpose-built SHEIE dataset. Table 8 delineates the model's performance across these datasets. Its performance also improved with the DA strategies of noise, time stretching, and audio data shifting, which diversified the training datasets, providing the model with more robust features and enabling the generalizability of new data. Table 9 compares the precision of the model developed to the models used by previous studies.

The model developed recognized the input sequence link using several Transformer blocks and a multi-head self-attention mechanism. DA, pre-processing, and complicated model design further increased the model's reliability. The ensemble model outperformed those developed in prior studies [11]–[23], demonstrably increasing the accuracy from 56.41% to 95.25% of the multiclass SVM model that included the MFMC, MFCC, LFPC, and LPCC features. For the first dataset EMO-DB, the ensemble model recorded an average accuracy of 99.8%, 96.3% in the RAVDESS dataset, 96.5% in the SAVEE dataset, 85.3% in the IEMOCAP dataset, and 83% in the SHEIE dataset. Thus, this model enhances SER. The model's average accuracy was 84.13% for anger, 81.07% for boredom, 84.13% for excitement, 82.10% for happiness, and 79.87% for sadness. Thus, the model can accurately classify numerous emotions. Voice-assisted emotion detection and online teaching tools could benefit from this feature. The ROC curve in Figure 18 visualizes the ensemble model's performance in terms of classifying emotions from the five datasets, with the x-axis representing the FPR and the y-axis representing the TPR. This indicates that the model has a high TPR (90%) and a low FPR (10%), signaling that it can effectively identify emotions.

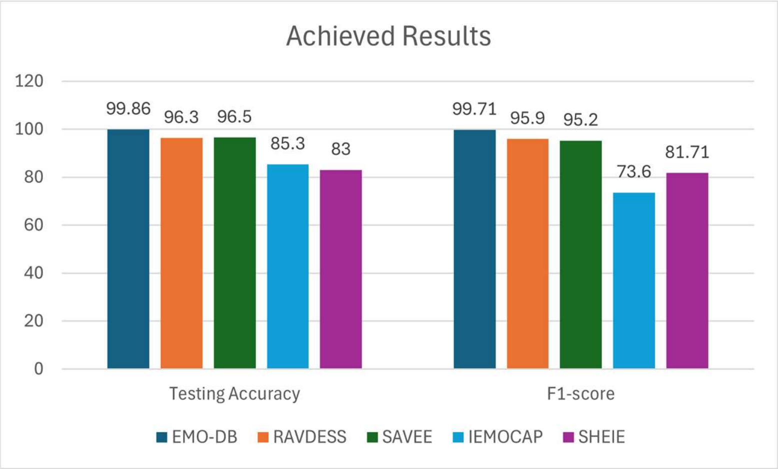
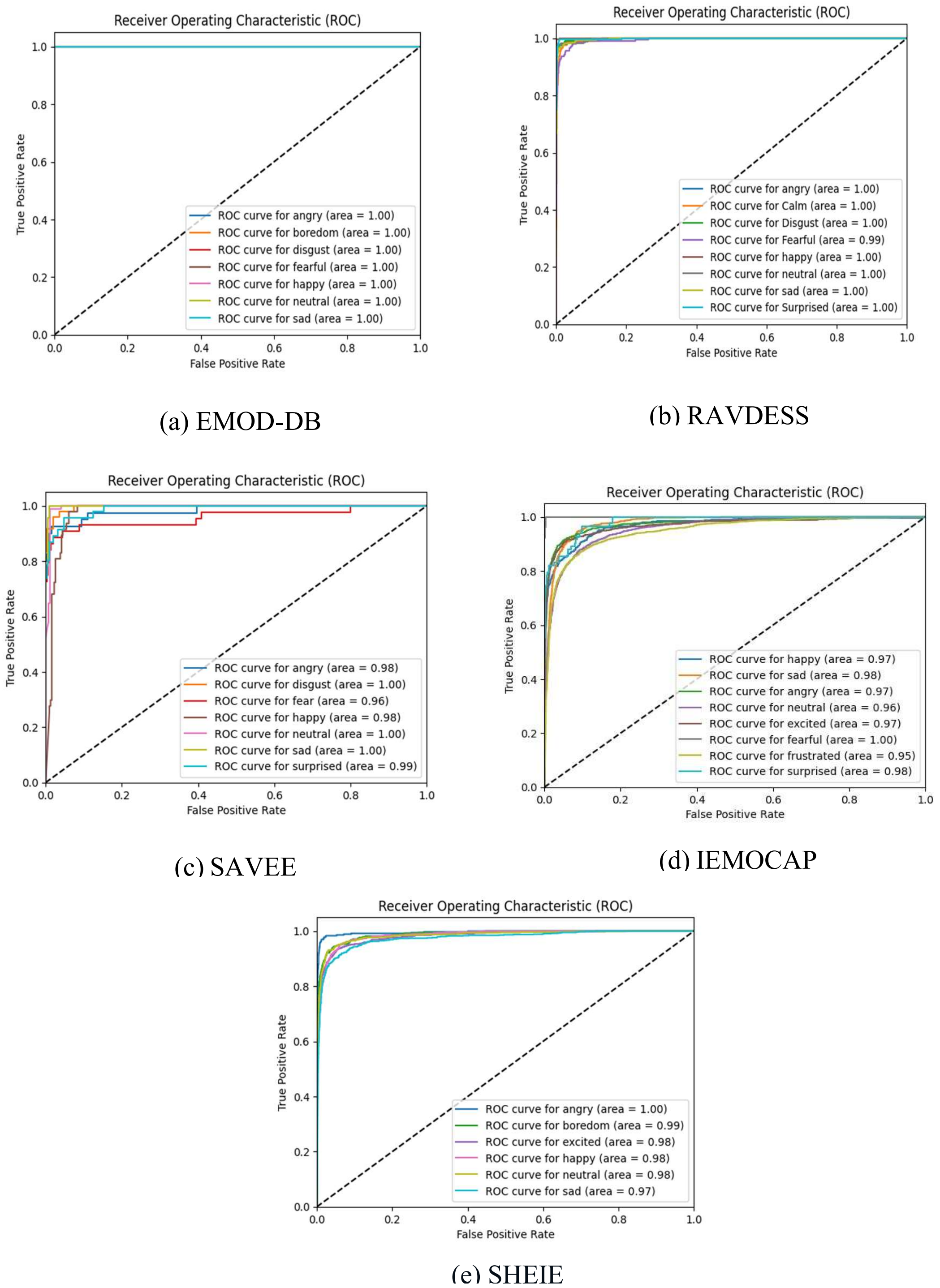


Figure 1. Summary of the model's accuracy and F1-scores across the datasets.



(a) EMOD-DB

(b) RAVDESS

(c) SAVEE

(d) IEMOCAP

(e) SHEIE

Figure 18. Receiver operating characteristics for the datasets: (a) EMO-DB; (b) RAVDESS; (c) SAVEE; (d) IEMOCAP; (e) SHEIE.

Table 9. The ensemble model's performance compared to models from previous studys.

Authors	Dataset	Feature Extraction and Techniques	Model	Results
Ancilin and Milton *t al.* [16]	EMO-DB, RAVDESS, SAVEE, EMOVO, eNTERFACE, Urdu	Mel frequency magnitude coefficient (MFMC), Mel-frequency cepstral coefficients (MFCC), LFPC, linear prediction cepstral coefficients (LPCC)	Multiclass support vector machine (SVM)	Urdu: 95.25%, EMO-DB: 81.5%, SAVEE: 75.63%, EMOVO: 73.3%, RAVDESS: 64.31%, eNTERFACE: 56.41%
Haotian *et al.* [17]	EMO-DB	Zero cross rate (ZCR), root-mean square energy (RMSE), pitch, harmonics-to-noise ratio, MFCC 1–12 (first twelve Mel-frequency cepstral coefficients)	SVM	66.4%, 70.8%
Alsabhan *et al.* [18]	ANAD, BAVED, SAVEE, EMO-DB	ZCR, RMSE, MFCC	One-dimensional (1D) convolutional neural network (CNN) with long short-term memory (LSTM) and self-attention mechanisms, custom two-dimensional (2D) CNN architecture	EMO-DB: 96.72%, SAVEE: 97%
Atmaja *et al.* [19]	IEMOCAP, Japanese Twitter-based emotional speech	Wav2vec 2.0 and pitch shifting, time stretching, silence removal used for data augmentation (DA)	SVM	77.25%
Tanko *et al.* [11]	Lectures dataset	Multi-level discrete wavelet transform and 1D orbital local binary pattern	SVM	Proposed method achieves 93.40% accuracy on the lectures dataset that contains three emotions. Also evaluated using the EMO-DB, Toronto Emotional Speech Set, and Sharif Emotional Speech Database datasets, achieving 86.14%, 99.82%,

				and 73.60% accuracy, respectively.
Soonil and Mustaqeem [22]	IEMOCAP, EMO-DB, RAVDESS	Short-time Fourier transform, CNN, Bidirectional LTSM	Object similarity measured using radial basis function networks	IEMOCAP: 72.25%, EMO-DB: 85.57%, RAVDESS: 77.02%
Yu and Xizhong [23]	IEMOCAP	Spectrogram	Bidirectional Gated Recurrent Unit Neural Network	82% increase in accuracy
Ahmed et al [24]	TESS, EMO-DB, RAVDESS, SAVEE, Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D)	MFCC for 13 coefficients , log Mel-scaled spectrogram, ZCR, chromogram, RMS, and DA	Ensemble model (1D CNN, LSTM, gated recurrent unit) with local feature-acquiring blocks and global feature-acquiring blocks	TESS: 99.46%, EMO-DB: 95.42%, RAVDESS: 95.62%, SAVEE: 93.22%, CREMA-D: 90.47%
Bahreini [4]	Collected dataset	Pitch, energy	Framework for Improving Learning Through Webcams And Microphones	Accuracy of 74.3 %
Tanko [5]	Lecturer speeches	Shoelace Pattern, tunable-Q wavelet transform sub-bands, Neighborhood Component Analysis	SVM	Accuracy of 94.97%, 96.41%
Zhang and Srivastava [23]	EMO-DB, Chinese Academy of Sciences' Chinese Affective Corpus	Speech essential frequency, quality intensity	Canonical Correlation Analysis , SVM	Average accuracy of 90%–95% for both
This study	RAVDESS, EMO-DB, SAVEE, IEMOCAP, SHEIE	MFCC 40 coefficient, Chroma, Mel spectrogram, ZCR, spectral contrast, centroid, bandwidth, roll-off, RMS	Ensemble model (Transformer, CNN, LSTM) with multi-head self-attention mechanism	RAVDESS: 96.3%, EMO-DB: 99.86%, SAVEE: 96.5%, IEMOCAP: 85.3%, SHEIE : 83%

6. Conclusions and Future Research Avenues

This study proposed a comprehensive system for measuring the emotional stability of remote educators, aiming to improve the quality of distance education. The ensemble model proposed combines Transformer, CNN, and LSTM architectures to enhance the identification of emotions in speech. MFCC, chroma, Mel spectrogram, ZCR, spectral contrast, centroid, bandwidth, roll-off, and RMS were extracted from audio files to enhance the model's effectiveness, with noise, time-stretching, and audio data shifting used as DA methods to improve the model's performance. This system was demonstrated to recognize emotions with an accuracy of 96.3% for the RAVDESS dataset, 99.86% for the EMO-DB dataset, 85.3% for the IEMOCAP dataset, 96.5% for the SAVEE dataset, and 83% for the SHEIE dataset. The SHEIE dataset developed for this study, which comprises recordings of instructor emotions during online teaching sessions, advances SER research. However, the various limitations of the proposed system should be addressed by improving pre-processing, adding features, and incorporating additional DA. Furthermore, we intend to broaden the number of languages included in the dataset to improve the model's ability to recognize emotion independent of language. Refining this method should increase its practical applicability, particularly in terms of analyzing the emotional states of distant educators, because it will allow the model to recognize emotional states across languages and datasets. Future research should test this model in real-world educational settings to determine its effects on teaching and learning.

Conflicts of interest: The authors assert that they have no conflict interests.

Data availability: Not applicable. We will share upon request.

References

1. L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. A. Mahjoub, and C. Cléder, "Automatic Speech Emotion Recognition Using Machine Learning," *Social Media and Machine Learning*, Mar. 2019, doi: 10.5772/INTECHOPEN.84856.
2. S. Ramakrishnan and I. M. M. El Emary, "Speech emotion recognition approaches in human computer interaction," *Telecommun Syst*, vol. 52, no. 3, pp. 1467–1478, Mar. 2013, doi: 10.1007/S11235-011-9624-Z/METRICS.
3. E. J. de Visser, R. Pak, and T. H. Shaw, "From 'automation' to 'autonomy': the importance of trust repair in human–machine interaction," *https://doi.org/10.1080/00140139.2018.1457725*, vol. 61, no. 10, pp. 1409–1427, Oct. 2018, doi: 10.1080/00140139.2018.1457725.
4. K. Bahreini, R. Nadolski, and W. Westera, "Towards real-time speech emotion recognition for affective e-learning," *Educ Inf Technol (Dordr)*, vol. 21, no. 5, pp. 1367–1386, Sep. 2016, doi: 10.1007/S10639-015-9388-2/TABLES/6.
5. D. Tanko, S. Dogan, F. Burak Demir, M. Baygin, S. Engin Sahin, and T. Tuncer, "Shoelace pattern-based speech emotion recognition of the lecturers in distance education: ShoePat23," *Applied Acoustics*, vol. 190, p. 108637, Mar. 2022, doi: 10.1016/J.APACOUST.2022.108637.
6. J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Information Fusion*, vol. 59, pp. 103–126, Jul. 2020, doi: 10.1016/J.INFFUS.2020.01.011.
7. Z. Taha, R. M. Musa, A. P. P. Abdul Majeed, M. R. Abdullah, M. M. Alim, and A. F. A. Nasir, "The application of k-Nearest Neighbour in the identification of high potential archers based on relative psychological coping skills variables," *IOP Conf Ser Mater Sci Eng*, vol. 342, no. 1, p. 012019, Apr. 2018, doi: 10.1088/1757-899X/342/1/012019.
8. B. Wang, M. Liakata, H. Ni, T. Lyons, A. J. Nevado-Holgado, and K. Saunders, "A Path Signature Approach for Speech Emotion Recognition," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-September, pp. 1661–1665, 2019, doi: 10.21437/INTERSPEECH.2019-2624.
9. X. Cheng and Q. Duan, "Speech Emotion Recognition Using Gaussian Mixture Model," *Proceedings of the 2012 International Conference on Computer Application and System Modeling, ICCASM 2012*, pp. 1222–1225, 2012, doi: 10.2991/ICCASM.2012.311.
10. A. Zhu and Q. Luo, "Study on speech emotion recognition system in E-learning," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 4552 LNCS, no. PART 3, pp. 544–552, 2007, doi: 10.1007/978-3-540-73110-8_59/COVER.
11. D. Tanko, F. B. Demir, S. Dogan, S. E. Sahin, and T. Tuncer, "Automated speech emotion polarization for a distance education system based on orbital local binary pattern and an appropriate sub-band selection technique," *Multimed Tools Appl*, pp. 1–18, Apr. 2023, doi: 10.1007/S11042-023-14648-Y/TABLES/5.
12. K. Chen, G. Yue, F. Yu, Y. Shen, and A. Zhu, "Research on speech emotion recognition system in E-learning," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 4489 LNCS, no. PART 3, pp. 555–558, 2007, doi: 10.1007/978-3-540-72588-6_91/COVER.

13. C. Huang, R. Liang, Q. Wang, J. Xi, C. Zha, and L. Zhao, “Practical speech emotion recognition based on online learning: From acted data to elicited data,” *Math Probl Eng*, vol. 2013, 2013, doi: 10.1155/2013/265819.
14. W. Li, Y. Zhang, and Y. Fu, “Speech emotion recognition in E-learning system based on affective computing,” *Proceedings - Third International Conference on Natural Computation, ICNC 2007*, vol. 5, pp. 809–813, 2007, doi: 10.1109/ICNC.2007.677.
15. Y. Zhang and G. Srivastava, “Speech emotion recognition method in educational scene based on machine learning,” *EAI Endorsed Transactions on Scalable Information Systems*, vol. 9, no. 5, pp. e9–e9, Feb. 2022, doi: 10.4108/EAI.10-2-2022.173380.
16. S. Parthasarathy and C. Busso, “Semi-Supervised Speech Emotion Recognition with Ladder Networks,” *IEEE/ACM Trans Audio Speech Lang Process*, vol. 28, pp. 2697–2709, 2020, doi: 10.1109/TASLP.2020.3023632.
17. Mustaqeem, M. Sajjad, and S. Kwon, “Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM,” *IEEE Access*, vol. 8, pp. 79861–79875, 2020, doi: 10.1109/ACCESS.2020.2990405.
18. M. Rayhan Ahmed, S. Islam, A. K. M. Muzahidul Islam, and S. Shatabda, “An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition,” *Expert Syst Appl*, vol. 218, p. 119633, May 2023, doi: 10.1016/J.ESWA.2023.119633.
19. S. R. Livingstone and F. A. Russo, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English,” *PLoS One*, vol. 13, no. 5, p. e0196391, May 2018, doi: 10.1371/JOURNAL.PONE.0196391.
20. F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, “A Database of German Emotional Speech”, Accessed: Jun. 12, 2023. [Online]. Available: http://www.expressive-speech.net/emodb/
21. “Surrey Audio-Visual Expressed Emotion (SAVEE) Database.” Accessed: Jun. 16, 2023. [Online]. Available: http://kahlan.eps.surrey.ac.uk/savee/
22. “IEMOCAP- Home.” Accessed: Jun. 12, 2023. [Online]. Available: https://sail.usc.edu/iemocap/
23. A. A. Torres-García, O. Mendoza-Montoya, M. Molinas, J. M. Antelis, L. A. Moctezuma, and T. Hernández-Del-Toro, “Pre-processing and feature extraction,” *Biosignal Processing and Classification Using Computational Learning and Intelligence: Principles, Algorithms, and Applications*, pp. 59–91, Jan. 2021, doi: 10.1016/B978-0-12-820125-1.00014-2.
24. M. Tawfik, S. Nimbhore, N. M. Al-Zidi, Z. A. T. Ahmed, and A. M. Almadani, “Multi-features Extraction for Automating Covid-19 Detection from Cough Sound using Deep Neural Networks,” *Proceedings - 4th International Conference on Smart Systems and Inventive Technology, ICSSIT 2022*, pp. 944–950, 2022, doi: 10.1109/ICSSIT53264.2022.9716529.
25. S. Jothimani and K. Premalatha, “MFF-SAug: Multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network,” *Chaos Solitons Fractals*, vol. 162, p. 112512, Sep. 2022, doi: 10.1016/J.CHAOS.2022.112512.
26. X. Guang, “Buddhist Impact on Chinese Culture,” *https://doi.org/10.1080/09552367.2013.831606*, vol. 23, no. 4, pp. 305–322, 2013, doi: 10.1080/09552367.2013.831606.
27. S. B. Davis and P. Mermelstein, “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences,” *IEEE Trans Acoust*, vol. 28, no. 4, pp. 357–366, 1980, doi: 10.1109/TASSP.1980.1163420.
28. J. Ancilin and A. Milton, “Improved speech emotion recognition with Mel frequency magnitude coefficient,” *Applied Acoustics*, vol. 179, p. 108046, Aug. 2021, doi: 10.1016/J.APACOUST.2021.108046.
29. H. Guan, Z. Liu, L. Wang, J. Dang, and R. Yu, “Speech Emotion Recognition Considering Local Dynamic Features,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10733 LNAI, pp. 14–23, 2017, doi: 10.1007/978-3-030-00126-1_2.
30. A. Vaswani *et al.*, “Attention Is All You Need,” *Adv Neural Inf Process Syst*, vol. 2017-December, pp. 5999–6009, Jun. 2017, Accessed: Jun. 14, 2023. [Online]. Available: https://arxiv.org/abs/1706.03762v5
31. R. Yacouby Amazon Alexa and D. Axman Amazon Alexa, “Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models,” pp. 79–91, Nov. 2020, doi: 10.18653/V1/2020.EVAL4NLP-1.9.
32. W. Alsabhan, “Human–Computer Interaction with a Real-Time Speech Emotion Recognition with Ensembling Techniques 1D Convolution Neural Network and Attention,” *Sensors 2023, Vol. 23, Page 1386*, vol. 23, no. 3, p. 1386, Jan. 2023, doi: 10.3390/S23031386.

33. B. T. Atmaja and A. Sasou, "Effects of Data Augmentations on Speech Emotion Recognition," *Sensors 2022, Vol. 22, Page 5941*, vol. 22, no. 16, p. 5941, Aug. 2022, doi: 10.3390/S22165941.
34. Y. Yan and X. Shen, "Research on Speech Emotion Recognition Based on AA-CBGRU Network," *Electronics 2022, Vol. 11, Page 1409*, vol. 11, no. 9, p. 1409, Apr. 2022, doi: 10.3390/ELECTRONICS11091409.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.