GMR
The Original

VERG: Hybrid Machine Learning Based Speech Emotion Recognition System

[1]Anamika Shukla Sharma, [2]H.S. Hota, [3]Raveena Rajak

[1]Govt. E.R.Rao PG Science College, Bilaspur, India, Email: anamikashuklacs@gmail.com
[2]Atal Bihari Vajpayee Vishwavidyalaya, Bilaspur, India, Email: proffhota@gmail.com
[3]Atal Bihari Vajpayee Vishwavidyalaya, Bilaspur, India, Email:raveenarajak21@gmail.com

ABSTRACT
This research delineates a hybrid framework for Voice-based Emotion Recognition and Gender Classification (VERG) employing an ensemble approach that integrates Deep Learning (DL) and Machine Learning (ML) techniques. Two benchmark datasets were employed ; The Toronto Emotional Speech Set (TESS) comprises 2800 samples for emotion recognition, whereas the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains an initial 1440 samples for emotion and gender analysis. To mitigate the data scarcity problem in RAVDESS and improve generalization, data augmentation via SpecAugment-based data augmentation was implemented to expand the dataset to 5760 samples and to strengthen the model's resilience to voice variances. Features were derived utilizing Mel-Frequency Cepstral Coefficients (MFCCs). The framework was evaluated using three models: Random Forest (RF), Long Short-Term Memory (LSTM), and a stacking ensemble of both. The models exhibited exceptional performance on TESS, achieving 98.86% for LSTM, 99.43% for RF, and 99.71% for stacking. In the RAVDESS emotion-only classification, the Random Forest (RF) model attained an accuracy of 84.14%, while the Long Short-Term Memory (LSTM) model achieved 82.06%. The accuracy was enhanced to 87.27% with the application of stacking. The models exhibited strong performance in joint emotion and gender classification on the enlarged RAVDESS dataset (5760 samples), achieving accuracy rates of 86.10% with RF, 84.70% with LSTM, and 88.50% with stacking. The overall results affirm the competitiveness of RF and LSTM, while also demonstrating the stacking ensemble's superior accuracy, underscoring the efficacy of hybrid methodologies in capturing the temporal and statistical attributes of speech for practical applications.

KEYWORDS: health literacy, cardiovascular disease, concept paper, knowledge gap, adult learning, intervention design, teach-back, plain language, community health worker, health equity.

1. INTRODUCTION

Speech is one of the most natural and powerful human communications. It transmits not only linguistic information, but paralinguistic information such as gender and emotion. The detection of such attributes from speech is crucial for intelligent context aware systems in human computer interaction, sentiment analysis, virtual assistants, healthcare monitoring and many other fields. In Speech Emotion Recognition (SER) and gender classification, the conventional methods mainly depend on the hand-crafted features (e.g. pitch, formants and energy) and the classical machine learning (ML) algorithms. These methods do a reasonably good job, but do not generalize well in the real world, where variations in accent, age, emotional state and background noise have a significant effect on speech signals. Secondly, most of the existing systems treat gender and emotion separately, which limits the holistic nature of human communication.

A two-stage hybrid Voice-based Emotion Recognition and Gender Classification (VERG) System is proposed in this research work to tackle these challenges by recognizing both gender and emotion from speech simultaneously. VERG is different from standard SER frameworks as it casts the problem as a multi-class classification challenge by integrating gender and emotion into composite labels (e.g. angry_male, happy_female etc.). Most traditional SER systems focus on emotion recognition and neglect the importance of gender as a basic property of speech signals. Gender is a key factor impacting acoustic properties such as pitch, formant frequencies and spectral energy distribution. Therefore, a recognition system that concurrently incorporates Gender and Emotion can lead to more accurate, resilient and context-aware applications. This has led to the creation of VERG systems that extend the scope of SER by including gender classification as an additional analysis stage. The general idea behind VERG is to extract acoustic and spectral data (e.g., Mel-Frequency Cepstral Coefficients i.e. MFCCs, pitch, prosodic features) along with machine learning or Deep

Learning(DL) classifiers to detect emotional states and gender attributes (Lee & Nadeem, ,2025; Teja et al., 2024; Yasmin et al., 2022).

Such systems hold significant practical value. Organizations can provide more customized support by discerning the caller's gender and emotional condition. In healthcare, VERG systems can facilitate the early identification of mental health issues by monitoring emotional patterns and gender-specific speech signs. The recognition of gender and emotions in virtual assistants and interactive systems enhances naturalness, flexibility, and customization, resulting in an improved overall user experience.

Nonetheless, numerous challenges exist in the design of VERG systems. Emotions in speech are generally complex, intermingled, and heavily dependent on the speaker. Likewise, gender classification must address variations in pitch, age, and recording conditions. Such difficulties may lead to diminished accuracy when employing a singular model. Hybrid ML models are a potential methodology that can utilize the strengths of many classifiers to effectively capture temporal correlations and feature variability.

A hybrid modeling approach that integrates deep and conventional ML for enhanced performance is proposed herein. The Long Short-Term Memory (LSTM) effectively handles the temporal correlations of speech, whilst RF (Random Forest) ensures robust categorization through ensemble learning. To enhance the robustness of forecasts, the predictions of LSTM and RF are integrated through a stacking ensemble method, utilizing a meta-classifier to capitalize on the strengths of both models. This hybrid two-stage pipeline enhances the precision and dependability of gender-sensitive emotion identification. The proposed VERG system as shown in Figure 1 aims to reconcile established SER frameworks with practical applications where gender and emotion are significant factors. The suggested system utilizes ML and incorporates a two-stage stacking hybrid design, enhancing accuracy, robustness, and adaptability across many domains.

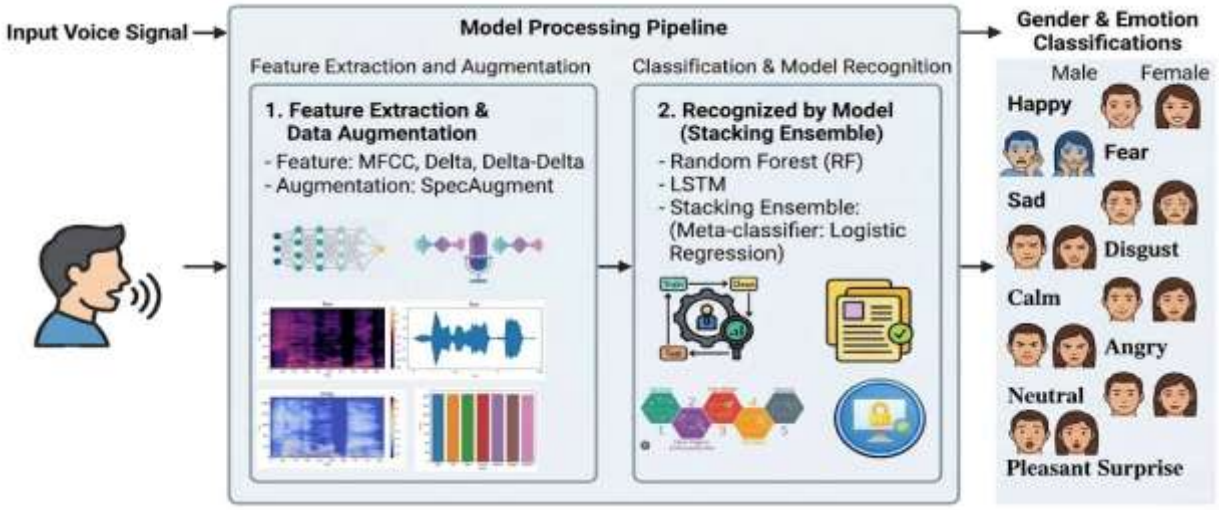


Fig. 1. Proposed VERG System flow diagram.

2.RELATED STUDIES

Deep Learning (DL) models for SER have garnered considerable attention recently because of their potential to enhance applications such as human-computer interaction, mental health monitoring, and affective computing. The field has evolved over the last ten years from traditional machine learning algorithms dependent on handcrafted features to advanced deep neural networks capable of autonomously extracting features and modeling temporal data. This paper critically evaluates novel DL models for voice emotion recognition, emphasizing technical aspects such as feature extraction, temporal modeling, and data augmentation techniques. This review seeks to consolidate findings from various architectures and datasets to delineate best practices, identify unresolved challenges, and propose future research directions to address existing gaps. The benefit is in delivering a thorough and technically detailed evaluation that informs the construction of more precise and robust SER systems.

Earlier models proposed by researchers (Lee & Narayanan, 2003; Chaves-Villota et al. ,2025; Balachandran et al., 2025; Garcia-Cuesta et al.,2023) do have reliance on carefully hand-crafted acoustic/prosodic features whereas, some rely heavily on features like MFCC (Poorna et al. ,2025; Shahin et al.,2019; Andayani et al., 2022, Shahin et al., 2023; Prayitno& Suyanto, 2019; Kaur et al., 2025; Mishra et al., 2025; Ezz-Eldin et al.,2021; Singh & Prasad, 2023; Wu et al., 2024; Madanian et al.,2023). Rather than just using raw MFCCs, recent studies are applying complex mathematical transformations before classification, such as the Fixed Frequency Range Empirical Wavelet Transform (Mishra et

al.,2025) for utilizing Temporal Bucketing specifically tailored for time-series data (Gurowiec& Nissim, 2025). DL approaches often utilize spectrograms or a fusion of hand-crafted features and automated feature extraction (e.g., BoAW, 3D Log Mel-Spectrograms). Models are transitioning from using all available data to using meta-heuristic optimization like the GREO algorithm by Dey et al., to select only the most relevant feature parameters (LPC/LPCC), which massively boosts accuracy while saving computational power.

There is a clear transition from classical statistical and fuzzy models FIS, GMM(Lee & Narayanan ,2003; Shahin et al.,2019), toward DL architectures like Convolutional Neural Networks(CNN), LSTM, Dilated Recurrent Neural Network (RNNs) and sophisticated ensemble methods (RF, Voting combinations).A significant advancement during this period was the transformation of audio into visuals (Mel-Spectrograms) and the application of image-recognition CNNs for emotion classification. Recent studies (Rasheed et al., 2024; Pham et al., 2023; Andayani et al., 2022) are transitioning from singular models to hybrid architectures such as CNN and LSTM or LSTM and Transformers to effectively capture spatial representations and long-term temporal contexts. This literature overview encompasses studies from foundational fuzzy logic and statistical models (GMM and FIS) to contemporary DL hybrids (CNN-BiLSTM with Time-Frequency Attention, Graph Convolutional Networks). As models increase in complexity, transitioning from standard ML to large language models and deep neural networks, there is an urgent demand for explainability. Di Luzio et al. (2025) exemplify the trend of ensuring that AI decisions in emotional computing are comprehensible and trustworthy for human operators. Esfahani & Adda (2024) illustrate the evolving domain of Natural Language Processing, wherein large models such as Mistral 7B are optimized to discern emotional context from text with greater efficacy than traditional Support Vector Machines (SVM). Frequency of methodology is demonstrated in Figure 2.

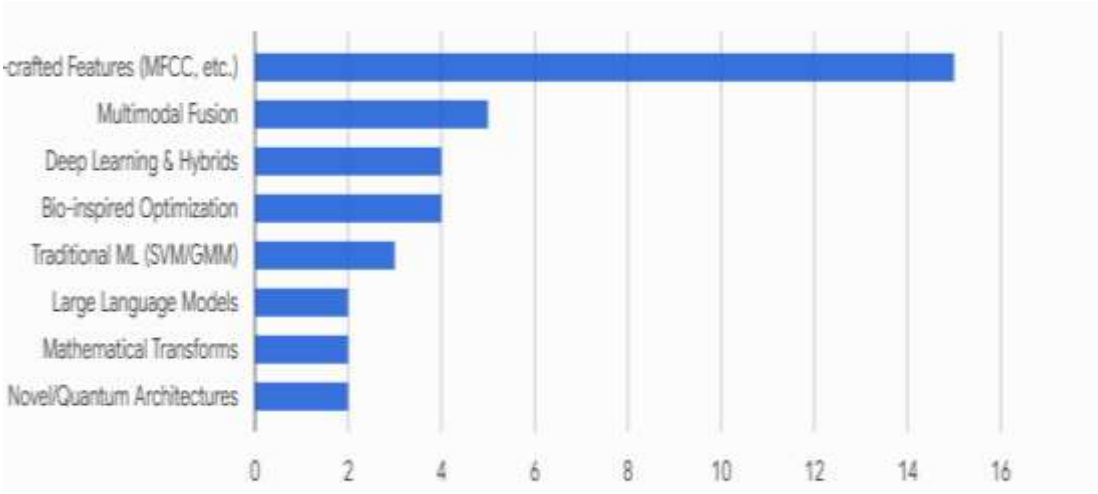


Fig. 2. Methodology frequency in SER.

Emotion recognition is rapidly moving beyond basic "happy/sad" classification into targeted clinical diagnostics. There is a pronounced shift toward applying SER in sensitive medical domains. Taşcı (2024) explores depression detection, Banos et al. (2024) review applications for autism spectrum disorder, and Dutta et al. (2025) focus specifically on early childhood speech processing. Jayasinghe et al. (2026) focus on the explainability required for mental health monitoring, Said et al., (2026) explore applications for Autism Spectrum Disorder (ASD), and Kaur et al. (2025), specifically target mental health counseling.

Single-modality recognition like just using speech is increasingly viewed as insufficient for real-world applications as presented in Figure 3. Researchers like Al-Saadawi (2024) and Pan (2023) highlight the necessity of fusing Text, Audio, and Visual data to accurately capture human sentiment. Rather than analyzing a single isolated voice clip, modern approaches like the one by Alhussein et al. (2025) are attempting to measure the "Emotional Climate" between two or more people, reflecting a more natural approach to human-computer interaction. Studies are increasingly looking beyond basic English audio by incorporating multimodal data (Kim et al.) and developing localized/custom language datasets (e.g., the new Spanish dataset by Garcia-Cuesta et al., (2023), Kaur et al. (2025), Shahin et al. (2023). Researchers are moving past training and testing on a single English dataset. Alroobaea (2024), Kaur et al. (2025), and Radhika et al. (2025) are actively trying to solve the problem of models failing when introduced to new languages, accents, or regional dialects (like Urdu, Arabic, and Indian regional languages).

Fusing audio with visual data is becoming the standard for higher accuracy (Geetha et al., 2024), and researchers are increasingly building localized datasets, such as the Arabic audiovisual database introduced by Al Roken & Barlas

(2023), to ensure models work globally. Several papers (Zhang, Jafari, Choo, Khare) emphasize that physiological signals Electroencephalogram(EEG) and Electrocardiogram(ECG), are much harder for humans to "fake" compared to facial expressions or tone of voice, making them a highly reliable (though harder to collect) source for emotion recognition. Researchers are moving beyond single-task SER. Chen et al. (2025) (MTLSER) simultaneously solve ASR, speaker ID, and SER. Chakhtouna et al. (2024) use multi-modal LLM embeddings (ImageBind), and Chaves-Villota et al. (2025) systematic review focuses heavily on combining acoustic and linguistic modalities. While models like the CNN-BiLSTM with Attention (Poorna et al., 2025) push the boundaries of accuracy, there is a counter-movement prioritizing efficiency and privacy. Bukhari et al. (2025) use evolutionary algorithms to lower computational costs, Taşcı (2024) reverts to highly optimized handcrafted features to save resources, and Dutta et al. (2025) emphasize data privacy.

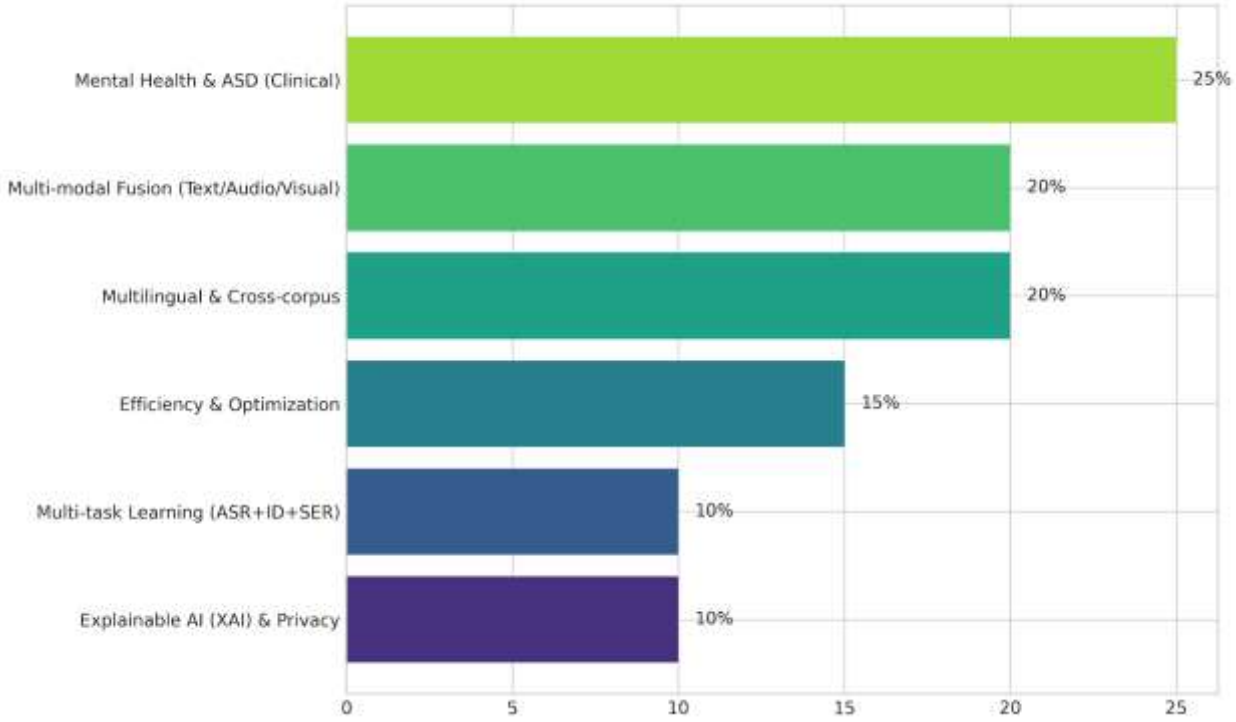


Fig. 3. Trend Analysis of application domains in SER.

To squeeze out better feature selection and computational efficiency, recent papers are deploying highly novel architectures, such as Quantum Convolutional Networks (Balachandran et al. (2025), Octonion algebra for the Metaverse (Daneshfar& Jamshidi, 2023) and unique bio-inspired algorithms like the Grey Wolf or Greater Cane Rat optimizers. A noticeable trend in optimizing feature selection and model hyperparameters using novel bio-inspired algorithms, such as Reptile Search Optimization (Zhang & Xiao ,2024) and Lyrebird Red Panda Optimization (Kanimozhi & Devi, 2025). This study also covered solving specific SER hurdles—such as data scarcity (via WaveGAN and SRHA augmentation) , environmental noise (via Binaural masks), hardware efficiency (via GCN for robots), and individual variability (via gender-dependent training).The systematic reviews (like Madanian et al., 2023, and Kumar & Jason, 2020) served as excellent anchoring points in summarizing the broader state of the field before diving into the specific methodologies of the experimental papers.

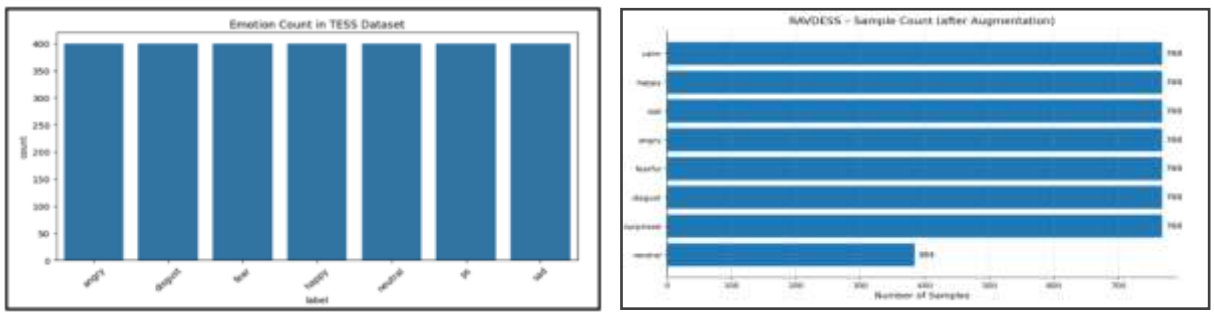
3. MATERIAL AND METHODS

3.1 Toronto Emotional Speech Set (TESS) SER Dataset

TESS is a high-quality emotional speech corpus containing 2,800 audio samples recorded by two female speakers aged 26 and 64 years. The dataset includes seven emotional categories—angry, disgust, fear, happy, sad, surprise, and neutral-spoken in a consistent phrase format, ensuring uniformity in articulation. All recordings are in 16-bit mono WAV format at a sampling rate of 24 kHz and captured in a noise-controlled environment, preserving fine acoustic details. Due to its balanced emotional representation, clear speech quality, and controlled recording conditions, TESS serves as a reliable resource for evaluating machine learning models in speech-based emotion and gender recognition research.

3.2 Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) SER Dataset

RAVDESS is a benchmark dataset consisting of 1,440 speech recordings from 24 professional actors (12 male and 12 female) expressing eight emotions—neutral, calm, happy, sad, angry, fearful, surprise, and disgust—at two intensity levels. All recordings are studio-quality 16-bit WAV files sampled at 48 kHz. In this study, the original RAVDESS dataset was augmented using SpecAugment, expanding it to 5,760 samples to increase data diversity and enhance model robustness.



(a)

(b)

Fig. 4. Distribution of unique emotional categories in (a) TESS and (b) RAVDESS datasets.

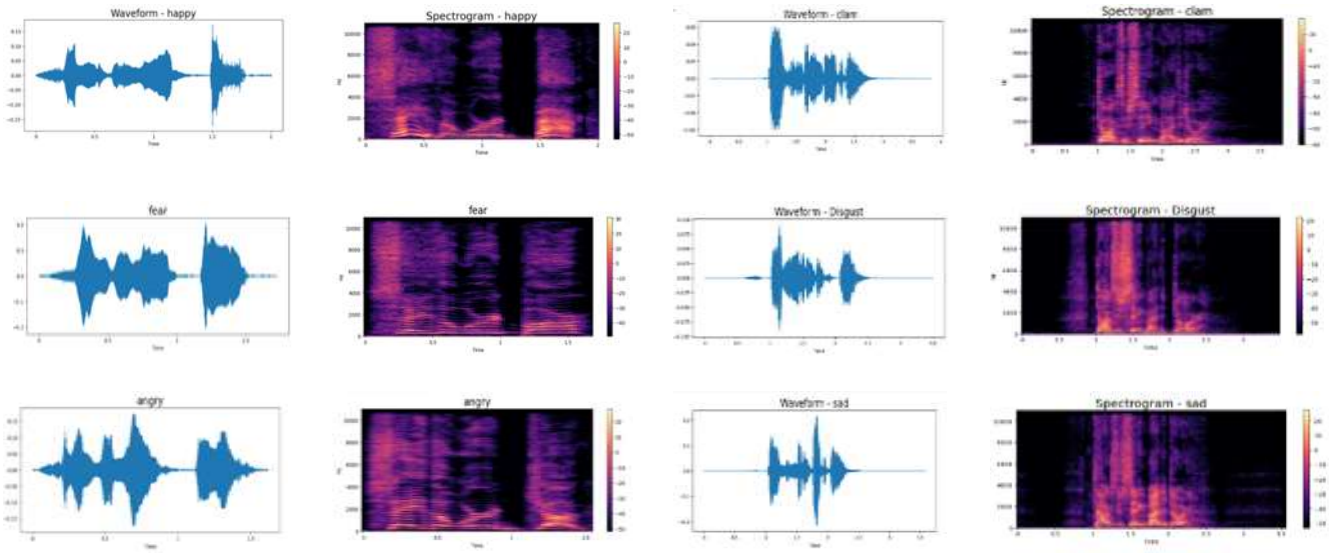


Fig. 5. Waveform and spectrogram representations of emotional speech samples.

Figure 4 illustrates the distribution of emotional categories in the TESS and RAVDESS datasets. The left chart represents the TESS dataset, where each of the seven emotions—fear, pleasant surprise, sad, angry, disgust, happy, and neutral—contains approximately equal numbers of samples, ensuring a balanced emotional representation. The right chart corresponds to the RAVDESS dataset, which includes eight emotions—neutral, calm, happy, sad, angry, fearful, surprise, and disgust. While most categories in RAVDESS have similar sample counts, the "calm" emotion exhibits a slightly lower frequency, introducing minor class imbalance. Such visual analysis of dataset distribution is crucial for understanding potential biases in training and evaluation of the ML models for gender-based voice and emotion recognition.

Figure 5 demonstrates waveform and spectrogram examples for six emotional speech utterances. The first three samples are taken from the TESS dataset, representing happy, fearful, and angry emotions, while the latter three are from the RAVDESS dataset, covering calm, sad, and disgust categories. In the TESS samples, happiness is marked by rhythmic amplitude fluctuations and a wide frequency range; fear displays irregular amplitude spikes with concentrated energy in

the mid-frequency band; and anger produces sharp, high-intensity peaks with dense high-frequency components. Conversely, in the RAVDESS samples, calm speech is reflected in smooth, low-energy patterns with evenly spread frequencies; sadness shows subdued amplitudes dominated by low-frequency regions; and disgust reveals sudden bursts of energy concentrated in the mid-to-high frequency range. These spectro-temporal characteristics emphasize how emotional states manifest differently across the two datasets.

Data augmentation was implemented to expand the size and variability of the RAVDESS speech emotion dataset and to enhance model robustness against overfitting. The original RAVDESS corpus consists of 1,440 audio samples, each representing different emotional states such as neutral, calm, happy, sad, angry, fearful, disgust, and surprised, spoken by both male and female actors. To enrich data diversity, five augmentation techniques—pitch shifting, time stretching, noise injection, time shifting, and volume scaling—were designed at the waveform level. However, instead of applying all five transformations to every sample, a random function was used to select and apply any three augmentations per original audio file. This ensured sufficient variability without introducing excessive redundancy. Consequently, each of the 1,440 original recordings produced three additional augmented versions, expanding the dataset to 5,760 audio samples in total and achieving a fourfold increase in data volume. Each augmented file was resampled to 22,050 Hz, trimmed, and normalized before feature extraction. This randomized augmentation approach introduced natural variability, improved generalization, and reduced overfitting in subsequent model training. Table 1 shows the distribution of augmented datasets.

Table 1: Distribution of Augmented RAVDESS Dataset across Emotion Classes and Gender.

	Gender		
Emotion classes	**Male**	**Female**	**Total**
Angry	384	384	768
Calm	384	384	768
Disgust	384	384	768
Fearful	384	384	768
Happy	384	384	768
Neutral	192	192	384
Sad	384	384	768
Surprised	384	384	768
Total			**5760**

3.3 Methodology

The proposed work combines conventional ML strategies with DL techniques to perform VERG. The dataset consisted of emotional speech samples, which were first pre-processed and converted into feature representations. For the TESS dataset, Mel-Frequency Cepstral Coefficients (MFCC) were extracted as the primary features, while for the RAVDESS dataset, the feature set was extended to include MFCCs along with their delta and delta-delta coefficients. Once the features were obtained, the data was split into training and testing sets, with the first used to build the model and the second to assess its accuracy.

Two models were developed individually: RF classifier, which served as a robust baseline ML model, and an LSTM, which is particularly well-suited for sequential data such as speech due to its ability to capture temporal dependencies. After training these models separately, an ensemble learning approach was applied using a stacking method, where the predictions of the RF and LSTM models were combined to create a meta-classifier for improved accuracy. This ensemble approach leveraged the strengths of both models: RF's capability to handle complex feature sets and LSTM's ability to model time-dependent patterns—resulting in a more discriminative and generalized system. Unlike conventional SER systems, the proposed VERG model simultaneously performs both emotion recognition and gender classification from a single audio input, thereby enhancing its applicability in real-world scenarios such as human-computer interaction, personalized assistants, and affective computing.

4. EXPERIMENTAL SETUP AND RESULTS

4.1 SER using TESS

The proposed framework employs the TESS dataset, comprising 2800 audio samples, which are split into 75% for training (2100 samples) and 25% for testing (700 samples). The audio samples undergo preprocessing followed by MFCC feature extraction. These representations are utilized to train both LSTM and RF models. To leverage their complementary strengths, a stacking ensemble is applied, in which a meta-model combines the predictions of LSTM and RF to enhance classification accuracy as shown in Figure 6. The performance of the system is comprehensively evaluated using multiple metrics.

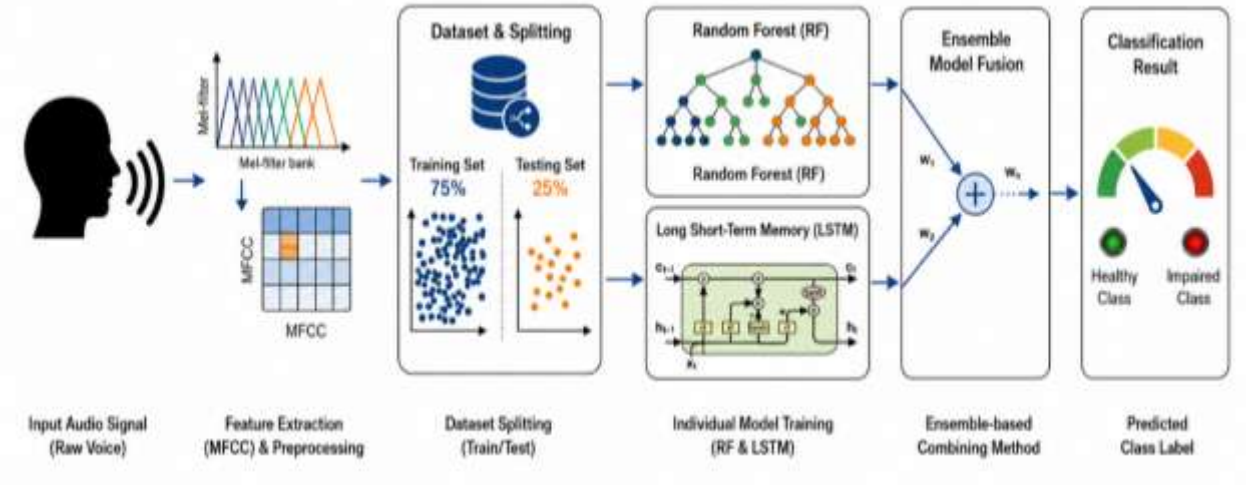


Fig. 6. Architecture of the SER System using the TESS Dataset

4.2 SER Using RAVDESS

The proposed framework utilizes the RAVDESS dataset, consisting of 1,440 original audio recordings across eight emotional categories. For each original audio file, 3 augmented versions were generated, resulting in a larger dataset. The audio samples are preprocessed through resampling, silence trimming, and segmentation into fixed-length clips, followed by data augmentation and SpecAugment for robustness. Feature extraction is performed in two ways: sequential MFCC representations for training an LSTM model and statistical spectral–temporal descriptors for a Random Forest classifier. The dataset is split as follows: 70% is used for training, while the remaining 30% is temporarily reserved for further splitting. This temporary 30% is equally divided into 15% for validation and 15% for testing. Finally, a stacking ensemble with Logistic Regression as the meta-model integrates the predictions of LSTM and RF, leading to improved classification performance as illustrated in Figure 7.

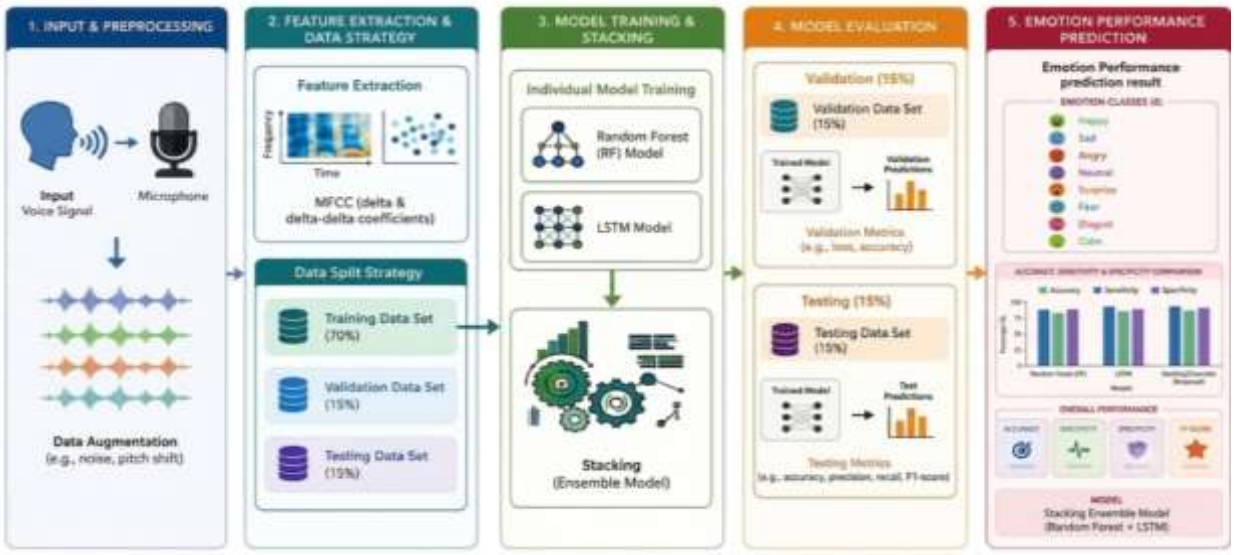


Fig. 7. Architecture of the Speech Emotion Recognition (SER) System using the RAVDESS Dataset.

4.3 VERG using RAVDESS

In the third case, the framework was extended to perform joint emotion and gender classification on the RAVDESS dataset. Both attributes were merged into a single composite label, such as (angry_male or happy_female), converting

the task into a multi-class classification problem. The original RAVDESS dataset contains 1,440 audio samples, and for each original file, 3 augmented versions were generated. The dataset was split as follows: 70% for training, and the remaining 30% was temporarily reserved and equally divided into 15% for validation and 15% for testing. Preprocessing steps, including resampling, silence removal, segmentation, and SpecAugment-based augmentation, were applied consistently to ensure robustness and balance across emotion–gender pairs. Feature extraction followed a dual-path strategy: sequential MFCC features were used to train the LSTM model to capture temporal dependencies, while statistical spectral–temporal descriptors were used with the Random Forest classifier to model distributional patterns. A stacking ensemble with Logistic Regression as the meta-model was then employed to combine the complementary strengths of RF and LSTM, as exhibited in Figure 8. This joint learning approach allowed the models to simultaneously capture emotional variations and gender-specific traits, resulting in strong classification performance. While RF and LSTM achieved competitive accuracies individually, the stacking ensemble consistently provided superior results, validating the effectiveness of the hybrid approach for multi-attribute speech analysis.

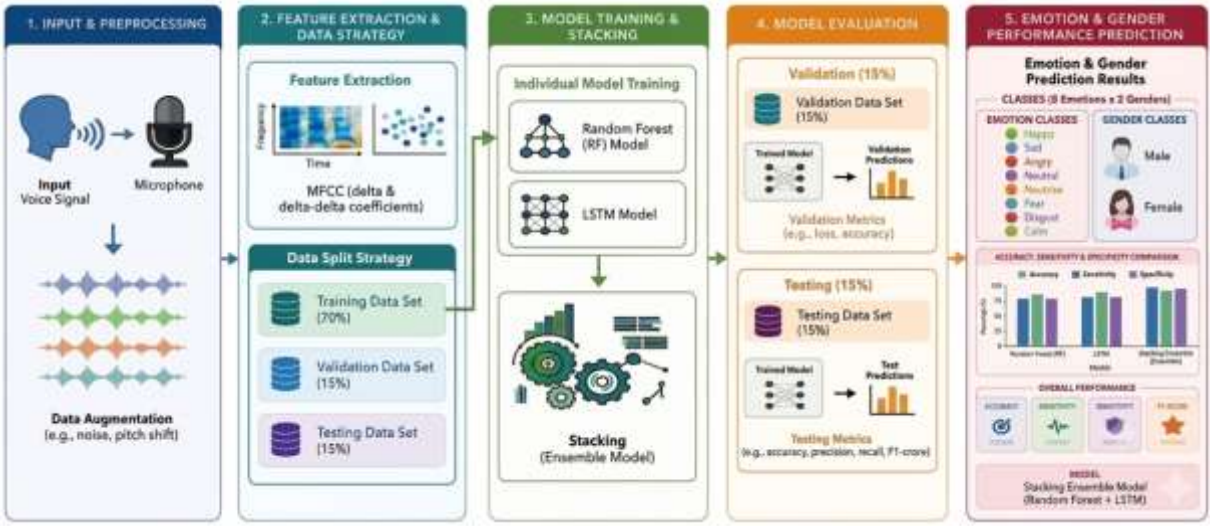


Fig. 8. Architecture of the VERG System on the RAVDESS Dataset.

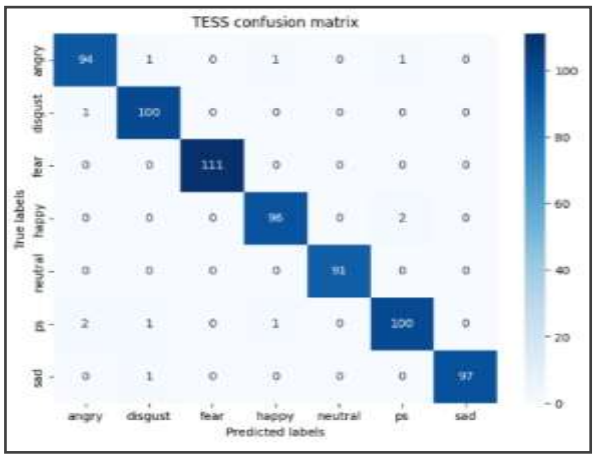
4.4 Result Analysis

The experimental evaluation was conducted using two benchmark datasets: TESS for emotion classification and RAVDESS for both emotion and joint emotion–gender recognition. MFCC features were extracted from all speech samples, with SpecAugment applied to the RAVDESS dataset to enhance robustness against variability in speaker expression. Three models were trained—RF, LSTM, and a stacking ensemble combining both.

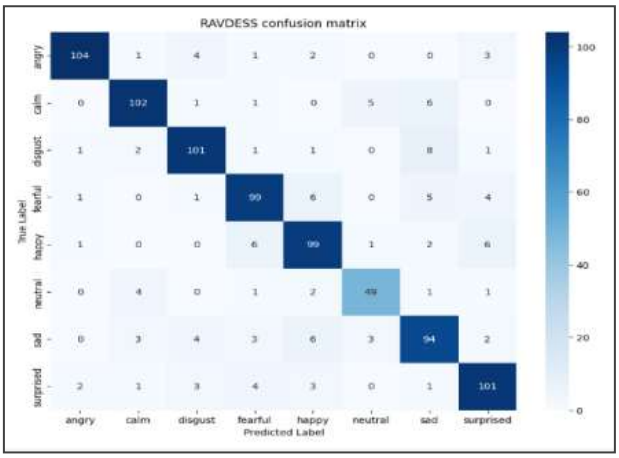
4.5 Confusion Matrices

To gain deeper insights beyond the overall accuracy values presented in the comparative performance table, confusion matrices were generated for all three experimental cases: SER using TESS, SER using RAVDESS and VERG using RAVDESS, each comprising RF, LSTM, and Stacking Ensemble models. These visualizations provide a detailed view of classification strengths and weaknesses by highlighting specific categories where misclassifications occur.

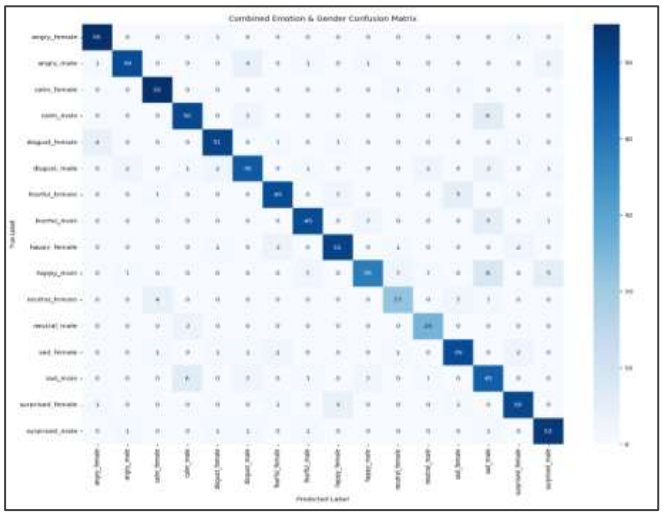
For the TESS dataset, the confusion matrices is shown in Figure 9 which reveal nearly flawless emotion recognition across all models, with the stacking ensemble achieving slightly superior performance. For the RAVDESS dataset, higher inter-speaker variability results in some overlap among acoustically similar emotions as shown in Figure 10. Most importantly, in case of VERG, which involves the joint emotion_gender classification task on the RAVDESS dataset the core objective of this study the stacking ensemble model demonstrates its full strength by effectively handling composite labels.



(a)

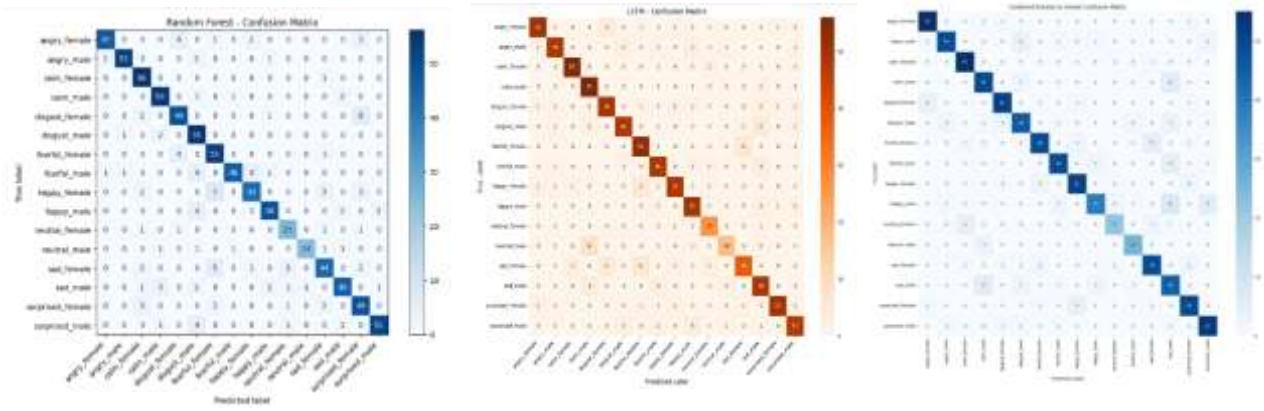


(b)



(c)

Fig. 9. Confusion matrices of the stacking ensemble model (a) SER on TESS (b) SER on DEVDASS (c) VERG RAVDESS



(a) (b) (c)

Fig. 10. Confusion matrices for (a) RF(b) LSTM, and (c) Stacking ensemble models on the RAVDESS dataset for joint emotion–gender recognition.

The accuracy–sensitivity–specificity comparison graph as shown in Figure 11 highlights the ensemble model's stable and reliable performance. Accuracy remains high across all categories, supported by specificity values that emphasize correct classification. Sensitivity metrics further illustrate the model's capability to recognize diverse emotional tones, reinforcing its generalization ability. Collectively, these measures confirm the ensemble framework's effectiveness in delivering consistently strong results for emotion–gender classification.

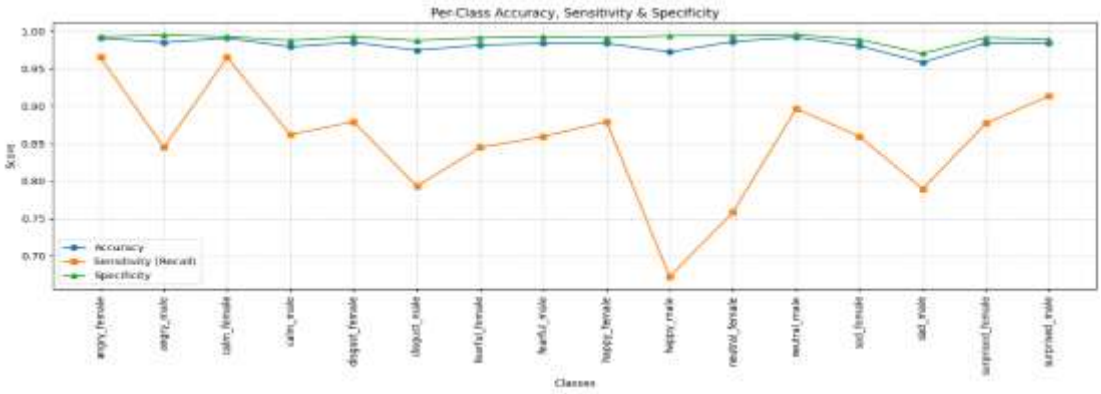


Fig. 11. Accuracy, sensitivity, and specificity comparison of the ensemble model across emotion–gender classes, demonstrating balanced and reliable performance in classification.

The results indicate that while RF and LSTM individually achieve competitive accuracy and sensitivity, the stacking ensemble provides a more balanced and reliable performance across all metrics. The VERG model with stacking demonstrates its effectiveness in jointly classifying emotions and gender with improved robustness. The improvements are particularly significant for the RAVDESS dataset, where variability in speech recordings poses greater challenges. Results shown in Table 2, Table 3 and Table 4 confirmed that the ensemble not only reduces misclassification rates but also maintains stability across multiple emotional states and gender categories.

Table 2: Class level performance evaluation of RF, LSTM, and Stacking models for SER on the TESS dataset

Emotion	RF			LSTM			STACKING		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Angry	99.00	98.00	100	97.00	96.00	99.00	99.00	99.00	99.00
Disgust	99.00	97.00	100	97.00	97.00	99.00	98.00	99.00	99.00
Fear	100	100	100	99.00	99.00	99.00	100	100	100
Happy	99.00	99.00	100	97.00	99.00	99.00	98.00	100	99.00
Neutral	100	100	100	99.00	99.00	100	100	100	100
PS (pleasant surprise)	97.00	98.00	100	98.00	99.00	99.00	99.00	97.00	98.00
Sad	100	100	100	99.00	98.00	100	100	100	100

Table 3: Class level performance evaluation of RF, LSTM, and Stacking models for SER on the RAVDESS dataset.

Emotion	RF			LSTM			STACKING		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Angry	84.00	83.00	99.00	82.00	82.00	97.00	87.00	87.00	99.00
Disgust	84.00	83.00	94.00	82.00	82.00	98.00	87.00	87.00	98.00
Fear	84.00	83.00	96.00	82.00	82.00	97.00	87.00	87.00	98.00
Happy	84.00	83.00	98.00	82.00	82.00	97.00	87.00	87.00	97.00
Neutral	84.00	83.00	98.00	82.00	82.00	97.00	87.00	87.00	98.00
Sad	84.00	83.00	99.30	82.00	82.00	98.40	87.00	87.00	99.00
Surprise	84.00	83.00	97.90	82.00	82.00	96.50	87.00	87.00	97.20
Calm	84.00	83.00	96.90	82.00	82.00	96.70	87.00	87.00	97.20

Table 4: Overall performance evaluation of Emotion and Gender classification on the RAVDESS dataset.

Emotion + Gender	Accuracy	Sensitivity	Specificity
angry_female	99.07	96.49	99.26
angry_male	98.50	84.48	99.50
calm_female	99.07	96.49	99.26
calm_male	97.92	86.21	98.76
disgust_female	98.50	87.93	99.26
disgust_male	97.45	79.31	98.76
fearful_female	98.15	84.48	99.13
fearful_male	98.38	85.96	99.26
happy_female	98.38	87.93	99.13
happy_male	97.22	67.24	99.38
neutral_female	98.61	75.86	99.40
neutral_male	99.19	89.66	99.52
sad_female	98.03	85.96	98.88

sad_male	95.83	78.95	97.03
surprised_female	98.38	87.72	9913
surprised_male	98.38	91.38	98.88

Table 5 presents a comparative evaluation of the models across three experimental cases: SER with the TESS dataset, SER with the RAVDESS dataset, and the proposed VERG model on the combined RAVDESS gender–emotion dataset. In Case of TESS – SER, all models achieved very high performance, with the stacking ensemble outperforming LSTM and RF individually, reaching 99.71% accuracy. In Case RAVDESS – SER, the performance decreased compared to TESS due to the greater complexity of the dataset, but stacking still improved the results, yielding 87.27% accuracy, higher than both RF 84.14% and LSTM 82.06%. In Case RAVDESS – VERG, when jointly classifying emotion and gender, stacking again achieved the best results (88.50% accuracy), showing a more balanced trade-off between sensitivity and specificity. Overall, the results indicate that the stacking ensemble consistently outperforms individual classifiers: LSTM and RF across all datasets, particularly in the more challenging joint classification task of VERG.

Table 5. Comparative performance of LSTM, RF, and Stacking ensemble models across three experimental cases: SER with the TESS dataset, SER with the RAVDESS dataset, and VERG with the combined RAVDESS gender–emotion dataset.

Case	Dataset	No. of Sample	No. of Classes	Model	Accuracy (%)	Sensitivity (%)	Specificity (%)
SER on TESS	TESS	2800	7	LSTM	98.86	98.85	99.81
				RF	99.43	99.46	99.90
				Ensemble method (Stacking)	99.71	99.71	99.71
SER on RAVDESS	RAVDESS	5760	8	LSTM	82.06	81.78	97.43
				RF	84.14	83.85	97.72
				Ensemble method (Stacking)	87.27	86.77	98.17
VERG on REVDESS	RAVDESS with Gender and Emotion	5760	8	LSTM	84.70	84.71	99.00
				RF	86.10	86.00	99.10
				Ensemble method (Stacking)	88.50	88.70	99.20

5. Conclusion

The proposed Two-Stage Hybrid VERG System successfully demonstrates the effectiveness of jointly recognizing both gender and emotion from speech signals using a hybrid machine learning framework. Unlike conventional SER systems that treat emotion and gender as independent tasks, VERG integrates them into a multi-class classification problem with composite labels (e.g., angry_male, happy_female), thereby capturing the holistic nature of human communication. Through the combination of RF for robust statistical classification and LSTM for modeling temporal dependencies, the system leverages an ensemble stacking approach to achieve enhanced accuracy and generalization. Experimental evaluations on benchmark datasets: TESS and RAVDESS confirmed that VERG not only establishes strong baselines for SER but also extends these capabilities to joint gender–emotion recognition, outperforming traditional standalone models. The strength of the VERG framework lies in its ability to handle the variability of real-world speech, including differences in accent, age, emotional expressiveness, and background noise. By incorporating both gender and emotion

in a unified pipeline, the system provides valuable applications across domains such as human–computer interaction, healthcare monitoring, adaptive learning environments, customer service, and emotion-aware virtual assistants. In conclusion, the VERG system bridges the gap between existing SER frameworks and practical real-world needs by offering a scalable, robust, and context-aware solution. Future research may extend this work by exploring multilingual speech datasets, deep attention-based architectures, and cross-corpus generalization to further improve adaptability and performance in diverse environments.

6. References

1. Al Roken, N., & Barlas, G. (2023). Multimodal Arabic emotion recognition using deep learning. Speech Communication, 155, 103005. https://doi.org/10.1016/j.specom.2023.103005
2. Alhussein, G., Alkhodari, M., Ziogas, I., Lamprou, C., Khandoker, A. H., &Hadjileontiadis, L. J. (2025). Exploring emotional climate recognition in peer conversations through bispectral features and affect dynamics. Computer Methods and Programs in Biomedicine, 265, 108695. https://doi.org/10.1016/j.cmpb.2025.108695
3. Alroobaea, R. (2024). Cross-corpus speech emotion recognition with transformers: Leveraging handcrafted features and data augmentation. Computers in Biology and Medicine, 179, 108841. https://doi.org/10.1016/j.compbiomed.2024.108841
4. Al-Saadawi, H. F. T., Das, B., & Das, R. (2024). A systematic review of trimodal affective computing approaches: Text, audio, and visual integration in emotion recognition and sentiment analysis. Expert Systems with Applications, 255, 124852. https://doi.org/10.1016/j.eswa.2024.124852
5. Andayani, F., Theng, L. B., Tsun, M. T., & Chua, C. (2022). Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files. IEEE Access, 10, 36018–36027. https://doi.org/10.1109/ACCESS.2022.3163856
6. A.V., G., T., M., D., P., & E., U. (2024). Multimodal Emotion Recognition with Deep Learning: Advancements, challenges, and future directions. Information Fusion, 105, 102218. https://doi.org/10.1016/j.inffus.2023.102218
7. Banos, O., Comas-González, Z., Medina, J., Polo-Rodríguez, A., Gil, D., Peral, J., Amador, S., & Villalonga, C. (2024). Sensing technologies and machine learning methods for emotion recognition in autism: Systematic review. International Journal of Medical Informatics, 187, 105469. https://doi.org/10.1016/j.ijmedinf.2024.105469
8. Bukhari, S. M. S., Zafar, M. H., Moosavi, S. K. R., & Sanfilippo, F. (2025). Emotion recognition with a Randomized CNN-multihead-attention hybrid model optimized by evolutionary intelligence algorithm. Array, 26, 100401. https://doi.org/10.1016/j.array.2025.100401
9. Chakhtouna, A., Sekkate, S., & Adib, A. (2024). Modeling Speech Emotion Recognition via ImageBind representations. Procedia Computer Science, 236, 428–435. https://doi.org/10.1016/j.procs.2024.05.050.
10. Chaves-Villota, A., Jimenez-Martín, A., Jojoa-Acosta, M., Bahillo, A., & García-Domínguez, J. J. (2026). Deep feature representations and fusion strategies for speech emotion recognition from acoustic and linguistic modalities: A systematic review. Computer Speech & Language, 96, 101873. https://doi.org/10.1016/j.csl.2025.101873
11. Chen, Z., Liu, C., Wang, Z., Zhao, C., Lin, M., & Zheng, Q. (2025). MTLSER: Multi-task learning enhanced speech emotion recognition with pre-trained acoustic model. Expert Systems with Applications, 273, 126855. https://doi.org/10.1016/j.eswa.2025.126855
12. Choo, S., Park, H., Kim, S., Park, D., Jung, J.-Y., Lee, S., & Nam, C. S. (2023). Effectiveness of multi-task deep learning framework for EEG-based emotion and context recognition. Expert Systems with Applications, 227, 120348. https://doi.org/10.1016/j.eswa.2023.120348
13. Daneshfar, F., & Jamshidi, M. (Behdad). (2023). An octonion-based nonlinear echo state network for speech emotion recognition in Metaverse. Neural Networks, 163, 108–121. https://doi.org/10.1016/j.neunet.2023.03.026
14. Dey, A., Chattopadhyay, S., Singh, P. K., Ahmadian, A., Ferrara, M., & Sarkar, R. (2020). A Hybrid Meta-Heuristic Feature Selection Method Using Golden Ratio and Equilibrium Optimization Algorithms for Speech Emotion Recognition. IEEE Access, 8, 200953–200970. https://doi.org/10.1109/ACCESS.2020.3035531
15. Di Luzio, F., Rosato, A., & Panella, M. (2025). An explainable fast deep neural network for emotion recognition. Biomedical Signal Processing and Control, 100, 107177. https://doi.org/10.1016/j.bspc.2024.107177
16. Dutta, S., Irvin, D., & Hansen, J. H. L. (2025). Exploring discrete speech units for privacy-preserving and efficient speech recognition for school-aged and preschool children. International Journal of Human-Computer Studies, 199, 103460. https://doi.org/10.1016/j.ijhcs.2025.103460
17. Esfahani, S. H. N., & Adda, M. (2024). Classical Machine Learning and Large Models for Text-Based Emotion Recognition. Procedia Computer Science, 241, 77–84. https://doi.org/10.1016/j.procs.2024.08.013
18. Ezz-Eldin, M., Khalaf, A. A. M., Hamed, H. F. A., & Hussein, A. I. (2021). Efficient Feature-Aware Hybrid Model of Deep Learning Architectures for Speech Emotion Recognition. IEEE Access, 9, 19999–20011. https://doi.org/10.1109/ACCESS.2021.3054345
19. G, B., S, R., G C, J., & T R, C. (2025). Advanced speech emotion recognition utilizing optimized equivariant quantum convolutional neural network for accurate emotional state classification. Knowledge-Based Systems, 316, 113414. https://doi.org/10.1016/j.knosys.2025.113414

20. Garcia-Cuesta, E., Salvador, A. B., &Pãez, D. G. (2023). EmoMatchSpanishDB: Study of speech emotion recognition machine learning models in a new Spanish elicited database. Multimedia Tools and Applications, 83(5), 13093–13112. https://doi.org/10.1007/s11042-023-15959-w
21. Gurowiec, I., & Nissim, N. (2025). Improving speech emotion recognition capabilities in the short and long term using temporal bucketing and active learning. Computers in Biology and Medicine, 196, 110912. https://doi.org/10.1016/j.compbiomed.2025.110912
22. Jafari, M., Shoeibi, A., Khodatars, M., Bagherzadeh, S., Shalbaf, A., García, D. L., Gorriz, J. M., & Acharya, U. R. (2023). Emotion recognition in EEG signals using deep learning methods: A review. Computers in Biology and Medicine, 165, 107450. https://doi.org/10.1016/j.compbiomed.2023.107450
23. Jayasinghe, H. M., Wong, K. W., &Nugaliyadde, A. (2026). A systematic review of interpretability and explainability for speech emotion features in automatic speech emotion recognition. Pattern Recognition, 171, 112122. https://doi.org/10.1016/j.patcog.2025.112122
24. Kaur, K., N, A., & Chellamani, G. K. (2025). Cross-lingual Speech Emotion Recognition for Mental Health Counselling and Aid. Procedia Computer Science, 258, 1425–1434. https://doi.org/10.1016/j.procs.2025.04.375
25. Khare, S. K., Blanes-Vidal, V., Nadimi, E. S., & Acharya, U. R. (2024). Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. Information Fusion, 102, 102019. https://doi.org/10.1016/j.inffus.2023.102019
26. Kim, D.-H., Son, W.-H., Kwak, S.-S., Yun, T.-H., Park, J.-H., & Lee, J.-D. (2023). A Hybrid Deep Learning Emotion Classification System Using Multimodal Data. Sensors, 23(23), 9333. https://doi.org/10.3390/s23239333
27. Kumar*, S., Jason, C. A., &M.Tech Student, Sreyas Institute of Engineering and Technology, Hyderabad, India. (2020). An Appraisal on Speech and Emotion Recognition Technologies based on Machine Learning. International Journal of Recent Technology and Engineering (IJRTE), 8(5), 2266–2276.
28. Lee, C. M., & Narayanan, S. (2003). Emotion recognition using a data-driven fuzzy inference system. 8th European Conference on Speech Communication and Technology (Eurospeech 2003), 157–160. https://doi.org/10.21437/Eurospeech.2003-88
29. Madanian, S., Chen, T., Adeleye, O., Templeton, J. M., Poellabauer, C., Parry, D., & Schneider, S. L. (2023). Speech emotion recognition using machine learning—A systematic review. Intelligent Systems with Applications, 20, 200266. https://doi.org/10.1016/j.iswa.2023.200266
30. Mishra, S. P., Warule, P., & Deb, S. (2025). Fixed frequency range empirical wavelet transform based acoustic and entropy features for speech emotion recognition. Speech Communication, 166, 103148. https://doi.org/10.1016/j.specom.2024.103148
31. N., K., & R., D. P. (2025). Hybrid Lyrebird Red Panda Optimization_Shepard Convolutional Neural Network for recognition of speech emotion in audio signals.Neurocomputing, 625, 129506. https://doi.org/10.1016/j.neucom.2025.129506
32. Pan, B., Hirota, K., Jia, Z., & Dai, Y. (2023). A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods. Neurocomputing, 561, 126866. https://doi.org/10.1016/j.neucom.2023.126866
33. Pham, N. T., Dang, D. N. M., Nguyen, N. D., Nguyen, T. T., Nguyen, H., Manavalan, B., Lim, C. P., & Nguyen, S. D. (2023). Hybrid data augmentation and deep attention-based dilated convolutional-recurrent neural networks for speech emotion recognition. Expert Systems with Applications, 230, 120608. https://doi.org/10.1016/j.eswa.2023.120608
34. Prayitno, B. A., & Suyanto, S. (2019). Segment Repetition Based on High Amplitude to Enhance a Speech Emotion Recognition. Procedia Computer Science, 157, 420–426. https://doi.org/10.1016/j.procs.2019.08.234
35. Radhika, S., Prasanth, A., & Sowndarya, K. K. D. (2025). A Reliable speech emotion recognition framework for multi-regional languages using optimized light gradient boosting machine classifier. Biomedical Signal Processing and Control, 105, 107636. https://doi.org/10.1016/j.bspc.2025.107636
36. Rasheed, B. H., Yuvaraj, D., Alnuaimi, S. S., & Priya, S. S. (2024). Automatic Speech Emotion Recognition Using Hybrid Deep Learning Techniques. International Journal of Intelligent Systems and Applications in Engineering, 12(15s), 87–96.
37. Said, Y., Saidani, T., Atri, M., Alsheikhy, A. A., &Shawly, T. (2026). Computational intelligence for emotion recognition in autism spectrum disorder: A systematic review of signal-based modeling, simulation, and clinical potential. Biomedical Signal Processing and Control, 111, 108367. https://doi.org/10.1016/j.bspc.2025.108367
38. Shahin, I., Alomari, O. A., Nassif, A. B., Afyouni, I., Hashem, I. A., &Elnagar, A. (2023). An efficient feature selection method for arabic and english speech emotion recognition using Grey Wolf Optimizer. Applied Acoustics, 205, 109279. https://doi.org/10.1016/j.apacoust.2023.109279
39. Shahin, I., Nassif, A. B., & Hamsa, S. (2019). Emotion Recognition Using Hybrid Gaussian Mixture Model and Deep Neural Network. IEEE Access, 7, 26777–26787. https://doi.org/10.1109/ACCESS.2019.2901352
40. Singh, V., & Prasad, S. (2023). Speech emotion recognition system using gender dependent convolution neural network. Procedia Computer Science, 218, 2533–2540. https://doi.org/10.1016/j.procs.2023.01.227

41. S.S., P., Menon, V., & Gopalan, S. (2025). Hybrid CNN-BiLSTM architecture with multiple attention mechanisms to enhance speech emotion recognition. Biomedical Signal Processing and Control, 100, 106967. https://doi.org/10.1016/j.bspc.2024.106967