Vol.8 Issue 3 July 2025

Multimedia Research

Received 20 May, Revised 26 May, Accepted 18 Juiy

Ensemble Deep Learning Architecture for Speech Emotion Recognition

Shivakumar Kagi
Department of Computer Science and Engineering
Sharnbasva University, Kalaburagi
shivakumarkagi@gmail.com

Abstract: Speech Emotion Recognition (SER) has also become a crucial building block in improving human-computer interaction by allowing systems to recognize and respond to human emotions via speech. This paper presents an innovative ensemble-based SER system that enhances recognition accuracy and robustness in varying acoustic conditions. The methodology starts with preprocessing the input audio via the Wiener filtering technique to significantly minimize background noise without sacrificing important speech features. This denoising step is particularly beneficial in real-world environments such as call centers or smart home devices, where background interference is common. After noise reduction, a discriminative feature set consisting of spectral features as well as high-level statistical descriptors is obtained to extract the emotional content present in speech. These features are selected for their proven relevance in capturing both the prosodic and spectral characteristics of emotional speech. For emotion classification, an ensemble model with three effective deep learning architectures (LinkNet, ShuffleNet, and GhostNet) is proposed. Each model predicts emotional states independently based on the extracted features, and their decisions are averaged to achieve the final decision. This fusion approach exploits the strengths of complementary models, adding to generalizability and performance stability. Experimentation results verify that the given method performs excellently in terms of emotion classification accuracy and can effectively support practical SER applications that demand reliability as well as responsiveness.

Keywords: Wiener Filtering, Ensemble Classifier, Linknet, Shufflenet, Ghostnet

Nomenclature

Abbreviation	Description
VQ	Vector Quantized
MAE	Masked Autoencoder
AV	Audiovisual
SER	Speech Emotion Recognition
CNN	Convolutional Neural Networks
DCNN	Deep Convolutional Neural Network
DL	Deep Learning
EQCNN	Equivariant Quantum Convolutional Neural Network
DMCL	Dynamic Multilevel Contrastive Loss
LEA	Local Enhancement Attention
SNN	Spiking Neural Network
PNEL	Perceptual Neuron Encoding Layer
BF	Bellman Filtering
SCO	Single Candidate Optimizer
RNN	Recurrent Neural Network
LN-SN-GN	Linknet- Shufflenet- Ghostnet

1. Introduction

Speech Emotion Recognition (SER) is a fast-emerging area of study in speech processing that aims at recognizing human emotions from voice signals. It is an important aspect of improving Human-Computer Interaction (HCI) and has extensive usage in application areas such as affective computing, virtual personal assistants, and healthcare systems. Through the ability to recognize emotional cues in speech through machines, SER enhances system flexibility and user interaction. In addition, it can potentially aid mental health tracking by detecting the early warning signs of emotional distress in real time [1]. This capability is particularly useful in telemedicine and remote patient monitoring systems, where continuous emotional assessment can lead to early intervention.

© Resbee Info Technologies Pvt Ltd

https://doi.org/10.46253/j.mr.v8i3.a5

Emotions are essentially multimodal and governed by many different factors, varying between the person's personality and contextual situation. They are communicated through verbal behavior in pitch, rhythm, intensity, and spectral profiles [2]. Such acoustic markers provide critical insights into affective states, even in the absence of visual or linguistic cues. Conventional SER systems have utilized handcrafted feature extraction like Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC), and prosodic cues based on pitch. These characteristics are usually categorized by machine learning models like Gaussian Mixture Models (GMM), Support Vector Machines (SVM), and Random Forests (RF). Although these methods have produced decent accuracy, they tend to fail in generalizability across datasets, speaker variability, and noisy environments [3]. Additionally, handcrafted approaches often require domain expertise and extensive preprocessing, limiting scalability in dynamic application settings.

The advent of deep learning has significantly enhanced SER by allowing automatic feature extraction and end-to-end learning from raw audio signals. Convolutional Neural Networks (CNNs) perform extremely well in extracting spatial features in spectrograms, while Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), can capture temporal dependencies in emotional speech [4]. Yet, while they perform well, these models face problems of vanishing gradients, not being able to handle long-range dependencies, and degradation in performance under varying conditions [5]. Moreover, their computational complexity poses a challenge for deployment in edge and resource-constrained environments, such as mobile devices or embedded systems.

To overcome these constraints, this paper suggests an efficient Speech Emotion Recognition system incorporating Wiener filtering-based noise cancellation, multi-level feature extraction, and lightweight deep learning model ensemble (LinkNet, ShuffleNet, and GhostNet) for better emotion classification. The use of lightweight architectures ensures a favorable balance between accuracy and real-time performance, making the proposed method suitable for latency-sensitive applications. The ensemble method improves accuracy, generalizability, and robustness in various acoustic environments. By combining complementary model strengths through decision-level fusion, the system effectively addresses the challenges of speaker variability and environmental noise.

The remainder of this paper follows this outline: Section 2.0 gives an overview of literature related to the work. Section 3.0 describes the proposed ensemble classification scheme of Speech Emotion Recognition. Section 4.0 discusses the experimental setup and results. Finally, Section 5 provides a summary and concluding statements of this work.

2. Literature Review

In 2025, Samir Sadok *et al.,* [6] proposed a self-supervised multimodal model, named VQ-MAE-AV that could utilize unlabeled AV speech data for emotion recognition. In this work, VQ variational autoencoders were utilized to map raw audio as well as visual inputs into discrete tokens, which were used to train MAE with attention mechanisms. This model could learn local as well as global speech representations. At pre-training, the model reconstructed the masked tokens that used contrastive learning to enhance representation learning. Similarly, at the fine-tuning phase, a lightweight classifier on top of the encoder learned for classifying emotion. This combination of tokenization and contrastive masking allowed the model to generalize well across multimodal domains. The proposed approach achieved traditional performance on various datasets under control as well as real-world settings.

In 2025, G. Balachandran *et al.,* [7] proposed a SER-EQCNN-ESC technique that used an optimized EQCNN for precise emotional state classification. This work began with extracting input speech samples and then preprocessing them through BF for noise elimination as well as normalization. A SCO was used to find out the most important speech emotion characteristics. Then, these attributes into affective categories by the EQCNN model. In contrast with conventional EQCNNs, the new method incorporated adaptive optimization for adjusting better parameters. This dynamic parameter tuning allowed the system to better adapt to complex acoustic conditions. The approach yielded much higher accuracy and precision than traditional SER approaches.

In 2025, Weiquan Fan *et al.,* [8] introduced a new multimodal SER system named LEDMCN, which combines a DMCL and a LEA mechanism. Here, DMCL proposed to handle the limitations in conventional contrastive learning methods by dynamically assigning distances to positive pairs according to similarity, thereby improving representation learning. Further, LEA was proposed to improve emotional feature extraction, which reinforced local token-level information in a global perspective to improve resilience against emotional oscillations. The integration of both mechanisms allowed the model to maintain robust performance even in emotionally ambiguous speech. In an end-to-end network integrating DMCL and LEA, the proposed model obtained traditional performance on the IEMOCAP and LSSED datasets.

In 2025, Chengyan Du *et al.,* [9] suggested a new dual-path SER model to tackle the poor accuracy issues in emotion recognition from speech. To enhance emotion recognition, they combined a traditional

CNN-based SER framework with an SNN capable of identifying weak and deeply embedded emotional signals triggered by dynamic pulses. To further facilitate signal processing, a PNEL was proposed, which improved the ability of the model to comprehend delicate emotional variations. This dual-path integration leveraged the strengths of both biologically-inspired neural processing and deep feature extraction. Through this coupled mechanism, more discriminative feature extraction was facilitated. Investigation of the IEMOCAP dataset displayed that the suggested scheme resulted in accuracy and exceeded other existing SER methods.

In 2025, Bineetha Vijayan and M. V. Judy [10] proposed EmoFusionNet, a new fusion-based framework to improve the efficiency and robustness of SER. They obtained MFCC features from whole speech signals and single frames and then fed them into two domain-specific sub-networks Att_BiLSTM and TempAwareNet. To support emotion classification, they employed a decision-level concatenation fusion strategy with subsequent fine-tuning. This architectural choice ensured comprehensive modeling of both global trends and transient emotional spikes. The model efficiently combined long-term dynamic and short-term static attributes, solving limitations in temporal dependency capturing and feature combination. EmoFusionNet obtained high accuracy on six benchmark datasets and surpassed state-of-the-art approaches, with significant gains in unweighted and weighted average recall.

3. Proposing an Ensemble-Based Speech Emotion Recognition Framework

This paper provides a speech emotion recognition system organized into three key stages:

- Initially, the audio input signal is denoised by the Wiener filtering process, which significantly minimizes background noise while keeping the integrity of the speech signal intact. This step is crucial for real-world applications, where background noise such as ambient chatter or mechanical sounds can severely impact recognition accuracy.
- From the preprocessed signal, appropriate spectral features and high-level statistical features are derived to represent frequency-domain as well as global statistical features related to emotion states. These features serve as the emotional signature of the speech and capture important variations in pitch, tone, and energy.
- Finally, a combination of LinkNet, ShuffleNet, and GhostNet is utilized as an ensemble model for emotion classification. The predictions from the three deep learning models are averaged to generate the final prediction, improving classification accuracy and stability. The decision-level averaging helps mitigate the weaknesses of individual models and ensures more reliable classification outcomes.

Fig. 1 shows the overall process of the proposed SER system, from input signal processing to final emotion recognition. The system architecture highlights a streamlined pipeline, balancing computational efficiency with high emotion recognition performance across diverse acoustic environments.

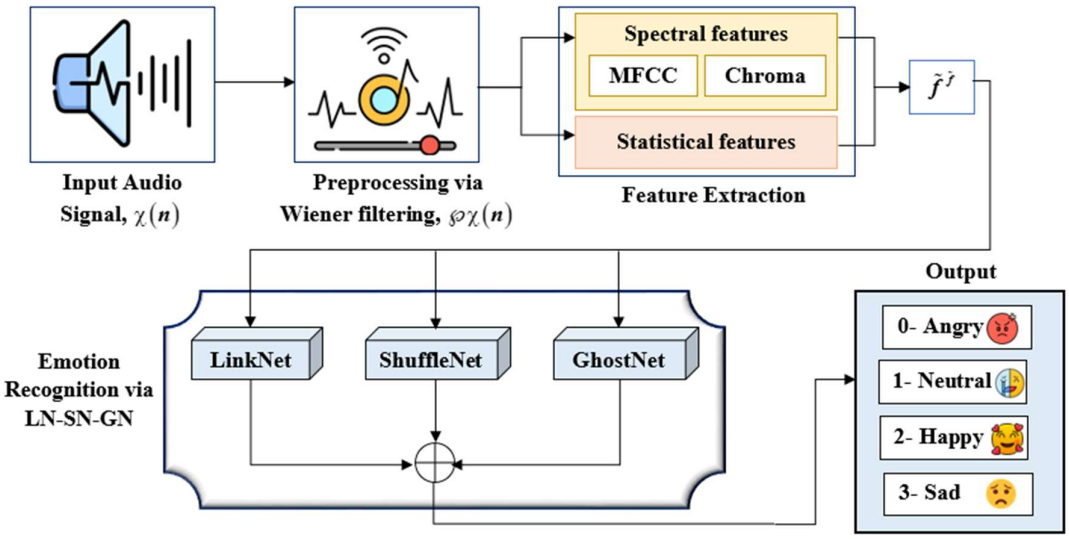


Fig. 1. *Process flow of the proposed speech emotion recognition approach*

3.1 Audio Preprocessing for Emotion Recognition

In the Speech Emotion Recognition system, preprocessing is a critical step in improving audio quality by suppressing environmental noise that may mask emotional information in speech. Wiener filtering [11] is used in processing the input audio signal, $\chi(n)$, which is usually interfered with by background noises. The Wiener filter is designed to give an estimation of the clean signal that will minimize the mean square error (MSE) between the target clean speech $\gamma(n)$ and the noisy signal $\xi(n)$. The estimated clean output signal $\hat{\chi}(n)$ is given by convolving the noisy signal with the impulse response of the filter $\upsilon(n)$, expressed mathematically in Eq. (1). This noise reduction process ensures that important emotional cues embedded in speech signals are preserved for subsequent feature extraction.

$$\hat{\chi}(n) = \upsilon(n) * \xi(n) = \sum_{i=-\infty}^{\infty} \upsilon(i)\xi(n-i) \tag{1}$$

The difference between the ideal signal and the filtered output is given by $\tilde{e}(n) = \gamma(n) - \hat{\chi}(n)$ and the goal is to minimize the MSE using Eq. (2). Optimization of this problem results in the optimal Wiener filter coefficients as per eq. (3).

$$J(\upsilon) = E\left[\tilde{e}^2(n)\right] \sim \min \tag{2}$$

$$\upsilon = A^{-1} A_{\chi\gamma} \tag{3}$$

where, A indicates the autocorrelation matrix of the noisy speech input signal $\chi(n)$ and \$ $A_{\chi\gamma}$ indicates the cross-correlation between $\chi(n)$ and the clean desired signal $\gamma(n)$. This filtering removes most background noise while retaining significant speech parameters. The denoised output signal, $\wp\chi(n)$, is subsequently sent to the feature extraction phase for further processing and classification. This step ensures the preservation of subtle prosodic and spectral cues that are essential for recognizing emotions such as anger, sadness, or joy.

3.2 Extraction of Spectral and High-Level Statistical Features

Following the process of denoising the raw speech signal with Wiener filtering, the clean signal $\wp\chi(n)$ is utilized to perform both spectral features and high-level statistical features extraction, which capture essential emotional information inherent in speech. During the process of feature extraction, the audio signal is translated into a meaningful representation that can be used in emotion classification. This dual-level feature extraction approach ensures that both short-term acoustic variations and global statistical trends are captured effectively.

3.2.1 Spectral Features, $\tilde{f}_{sp}^{\tilde{f}}$

Spectral features encode frequency-domain properties of the speech signal, which play a key role in capturing emotional patterns in human speech. Within this contribution, two major spectral feature types are extracted from the preprocessed audio signal, $\wp\chi(n)$ such as Mel-Frequency Cepstral Coefficients (MFCC) and Chroma features.

(a) MFCC Features

Mel-Frequency Cepstral Coefficients (MFCCs) [12] are commonly used in speech processing because they are capable of representing the response of the human auditory system. The initial step is framing, where the preprocessed speech signal $\wp\chi(n)$ is segmented into short, overlapping frames. A Hamming window, $\omega(n)$ is added to each frame to minimize signal discontinuity at the frame edges, which is computed in Eq. (4) and each frame's output is expressed in Eq. (5). which L signifies the number of samples per frame, $O(n)$ signifies the windowed output signal. This windowing process reduces spectral leakage by tapering the edges of each frame, which is essential for accurate frequency analysis. The overlapping frames ensure temporal continuity and capture dynamic changes in speech important for emotion recognition.

$$\omega(n) = 0.54 - 0.4\cos\left[\frac{2\pi n}{L-1}\right] \tag{4}$$

$$O(n) = \wp\chi(n) \times \omega(n) \tag{5}$$

After that, the speech signal is transformed from the time to the frequency domain by the Fast Fourier Transform (FFT), which allows frequency components of the speech signal to be analyzed. A Mel filter bank is computed from the magnitude spectrum, transforming the frequencies into the Mel scale to

simulate human hearing of pitch. The Mel frequency $m\tilde{f}$ corresponding to an actual frequency $\tilde{f}$ in Hz is determined by Eq. (6).

$$m\tilde{f} = 2595\log_{10}\left(1+\frac{\tilde{f}}{100}\right) \tag{6}$$

Subsequently, the energy in each Mel filter is sent to the Discrete Cosine Transform (DCT) for decorrelating the features and gaining a compact cepstral representation. The MFCC coefficients C_{MFCC} are calculated as per Eq. (7). where, N indicates the number of Mel filters, and m signifies the index of the coefficient.

$$C_{MFCC} = \sum_{i=1}^{N}\log\left(O(i)\right)*\cos\left[mx(i-0.5)x\pi \div N\right] \tag{7}$$

MFCCs are particularly effective in encoding the spectral envelope, which helps capture variations in tone, stress, and speech articulation that correspond to different emotions.

(b) Chroma Features

Chroma features [13] are derived from the preprocessed speech signal, $\wp\chi(n)$ to track the tonal and harmonic information that is perceptually significant in human emotion. Chroma features are the magnitude of the 12 semitone pitch classes independent of an octave, essentially compacting the harmonic content of a speech frame. This is especially helpful in emotion detection since variations in pitch and tone such as rising intonation for anger or falling tone for sadness, are good emotional indicators. By focusing on pitch classes rather than absolute frequencies, chroma features capture the relative harmonic content that remains consistent across different octaves, making them robust to pitch variations. To calculate chroma features, a log-frequency spectrogram $\Im_{LF}(m,p)$ is calculated from the signal, which p indicates pitch values. For every class, the combined spectral energy is computed by adding up the magnitude of all the frequencies mapping to said chroma value using Eq. (8). This grouping leads to a 12-dimensional chroma vector, or a chromagram, that highlights the pitch-related attributes invariant across octaves.

$$Ch(m,c) = \sum_{p\in[0:127|p \bmod 12=c]} \Im_{LF}(m,p) \tag{8}$$

Chroma features complement MFCCs by capturing tonal content, especially useful when the emotional state affects intonation patterns or harmonic structures.

3.2.2 High-Level Statistical Features, $\tilde{f}_{st}^{\tilde{f}}$

High-Level Statistical Features [14] are intended to reflect the overall statistical behavior of the speech signal across time. In this work, statistical features like mean, standard deviation, minimum, and maximum are extracted from the preprocessed signal, $\wp\chi(n)$. It helps the classification models to differentiate between emotions more accurately. These features provide temporal context by summarizing long-term patterns in acoustic behavior across the entire utterance.

(a) Mean

In the SER, the mean can be used to capture typical behavior of features such as energy, pitch, or MFCC coefficients over time. It is calculated by adding all the points d_k and dividing the sum by the number of samples Q, as indicated in Eq. (9).

$$\mu = \frac{1}{Q}\sum_{k=1}^{Q} d_k \tag{9}$$

(b) Standard Deviation

Standard deviation measures the degree of variation or spread of the data. Variability in properties such as pitch, intensity, or formants in emotional speech can indicate expressive emotions. The standard deviation is defined as per Eq. (10).

$$\sigma = \sqrt{\frac{1}{Q}\sum_{k=1}^{Q}\left(d_k-\mu\right)^2} \tag{10}$$

(c) Minimum and Maximum

The minimum and maximum measures the extent of variation in a specified feature set. These measures are useful in identifying emotional extremes. These values set limits within which the emotional expressions vary and assist in detailing the emotional profile of a segment of speech. Together, these descriptors help in identifying speech segments that exhibit unusual emotional expression, such as sudden spikes in intensity or drastic pitch changes.

Therefore, the entire feature vector employed for classification is a union of spectral and statistical descriptors $\tilde{f}^{\tilde{f}}$, $\tilde{f}^{\tilde{f}} = \left[\tilde{f}_{sp}^{\tilde{f}} \;\; \tilde{f}_{st}^{\tilde{f}}\right]$. This is used as input by the ensemble deep learning models in the classification stage to provide reliable and precise emotion recognition. This hybrid feature set ensures both fine-grained acoustic details and holistic emotional trends are utilized during classification.

3.3 Ensemble-Based Emotion Recognition via LN-SN-GN Models

In the proposed system's recognition phase, an ensemble classification approach is adopted to enhance the robustness and accuracy of SER. The final feature set, $\tilde{f}^{\tilde{f}}$ which is both spectral and statistical features combined, is fed to three compact yet efficient convolutional neural network (CNN) architectures, namely LinkNet, ShuffleNet, and GhostNet, at the same time. Each model processes the input features separately and generates a prediction vector. These prediction vectors are then combined using averaging to generate the final predicted emotion label, which is shown in Fig. 2. This decision-level fusion approach enables the ensemble to leverage the unique learning capacities of each architecture.

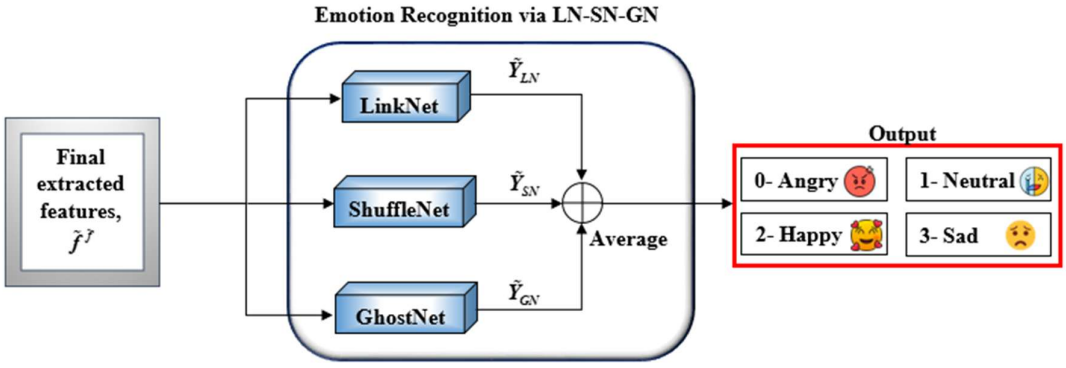


Fig. 2. *Process of Emotion recognition phase*

3.3.1 LinkNet

The LinkNet [15] is the first model in the ensemble, a deep encoder-decoder neural network architecture that is optimized for compact semantic segmentation. In this work, LinkNet is specifically used for classifying emotional states from the extracted feature set $\tilde{f}^{\tilde{f}}$. The output of the LinkNet classifier is represented as $\tilde{Y}_{LN}$.

The LinkNet architecture comprises two main components: Encoder and Decoder. The encoder down samples the spatial dimensions of the input with high-level feature extraction via convolutional layers, such as stridden convolution for downsampling, after which is applied batch normalization and non-linear activation. Also, the decoder mimics this pattern by employing upsampling operations and convolutions to regain the spatial dimensions and create the classification features. Skip connections are used between corresponding encoder and decoder blocks to preserve spatial context and enable better gradient flow during backpropagation. The structure of this architecture strikes an ideal balance between real-time responsiveness and precision, hence suitable for lightweight SER applications. The network starts with a convolutional layer with companion bottleneck blocks that preserve feature information across layers. This design ensures that while reducing the spatial complexity, key emotional features remain intact through each layer, improving the final classification accuracy.

3.3.2 ShuffleNet

Second, the feature vector $\tilde{f}^{\tilde{f}}$ is passed through ShuffleNet [16] for effective classification to obtain an output, $\tilde{Y}_{SN}$. ShuffleNet is an efficient convolutional neural network to ensures high accuracy with low computational overhead and hence is well-fitted for applications that need rapid processing. It introduces two key innovations: pointwise group convolutions and a channel shuffle. Pointwise group convolutions reduce the computation by partitioning channels into groups, yet traditional group convolutions limit information passing between groups. The channel shuffle operation sidesteps the limitation by rearranging the channels so that information passing between groups is facilitated without additional computational overhead. This structure achieves a significant reduction in floating-point operations (FLOPs), enabling the model to operate effectively on mobile or embedded devices. This shuffling mechanism allows more interaction across groups, overcoming the representational bottleneck found in earlier grouped convolution

methods. ShuffleNet takes a bottleneck approach, initially utilizing a 1×1 pointwise group convolution to reduce feature size, Depthwise 3×3 convolutions for low-cost spatial filtering within each channel, and subsequently another 1×1 group convolution to restore the feature map size. Skip connections are taken when input and output sizes are equal, which can enhance learning ability without extra computation. For downsampling, channel concatenation and 3×3 average pooling are applied to ensure efficiency. As a result, ShuffleNet maintains high inference speed while still retaining strong learning capability, making it an excellent candidate for real-time SER tasks. This renders ShuffleNet inference-fast with low resource consumption.

3.3.3 GhostNet

The third model in the ensemble is GhostNet [17], which also takes the feature vector $\tilde{f}^{\tilde{f}}$ as input and produces an output $\tilde{Y}_{GN}$. GhostNet tries to lower the computational expense of deep CNNs using a new Ghost module. This module effectively generates a huge number of feature maps by first producing a smaller set of intrinsic feature maps through standard convolution. These intrinsic maps represent the essential information of the input with minimal channel. Subsequently, by cheap linear transformations, the intrinsic maps are extended to generate more ghost feature maps, thereby augmenting the overall output without significant parameter or computation overhead. This separation between intrinsic and ghost features enables a high ratio of feature reuse while significantly reducing the number of multiply-accumulate operations (MACs). The GhostNet model starts with a 3×3 convolution for size reduction and low-level feature extraction, and then several Ghost blocks that make use of the Ghost module to produce efficient feature maps. The blocks consist of 1×1 convolution, Depth wise convolutions, batch normalization, and ReLU activation for stable training. Following the blocks, global average pooling reduces the feature maps to a compact form. Lastly, a Softmax activation fully connected layer is used to classify the input into emotion classes from the extracted features. GhostNet's low latency and reduced memory footprint make it well-suited for edge deployments and resource-constrained platforms.

3.3.4 Final Output Fusion

The results of the three classifiers such as $\tilde{Y}_{LN}$, $\tilde{Y}_{SN}$ and $\tilde{Y}_{GN}$, are fused by averaging their predicted scores or probabilities. This fusion method takes advantage of the capability of each individual model and generates a better overall classification result. Averaging ensures that predictions are smoothed and that no single model dominates the final output, leading to a more generalized and reliable decision. The final decision of the ensemble classifies the input into one of the four emotion categories "0 for Angry, 1 for Neutral, 2 for Happy, and 3 for Sad". This ensemble-based strategy not only boosts recognition accuracy but also reduces variance across samples, leading to higher consistency in emotionally diverse environments.

4. Results and Discussion

4.1 Simulation Procedure

The proposed ensemble Classification approach for Speech Emotion Recognition was executed with the PYTHON, specifically "Version 3.7". Simulations were performed on a machine set up with an "11th Gen Intel(R) Core (TM) i5-1135G7 @ 2.40GHz 2.42 GHz and the Installed RAM size was 16.0 GB". This setup ensures a balance between computational feasibility and performance for training lightweight deep learning models. The experiments used a benchmark dataset [18] referred to as LSSED (Large-Scale Dataset and Benchmark for Speech Emotion Recognition) which has rich annotations and varied emotional content appropriate for the testing of deep learning models' performance in mixed realistic situations. Python libraries such as NumPy, SciPy, TensorFlow, and Scikit-learn were used for preprocessing, model implementation, training, and evaluation.

4.2 Dataset Description

Experimental testing was conducted on the LSSED dataset, an extensive English speech corpus specifically intended for emotion recognition tasks. The dataset consists of 147,025 annotated sentences, representing more than 206 hours of audio, recorded from 820 distinct speakers. Each sentence is annotated with one or more of 11 emotional classes, such as angry, neutral, fearful, happy, sad, disappointed, bored, disgusted, excited, surprised, and other. Such diversity provides a robust foundation for training models with high generalization capability. For the present research, we considered four prevailing sentiments: Angry (0), Neutral (1), Happy (2), and Sad (3) to reduce and assess the classification

performance more heavily. This choice simplifies classification while preserving critical affective distinctions useful for practical applications such as healthcare, call centers, and virtual assistants.

4.3 Performance Analysis

The performance analysis of the proposed LN-SN-GN emotion recognition system exhibits its unequivocal superiority over traditional deep learners like LinkNet, ShuffleNet, GhostNet, CNN, and RNN under a range of criteria and proportions of training data. Each model was tested using the same preprocessed feature vectors and hyperparameter settings to ensure a fair comparison. The model outperforms baseline methods across all major positive metrics such as accuracy, precision, sensitivity, and specificity with improvements becoming even more significant with increased amounts of training data, reflecting strong generalization and increased capability to classify emotional classes accurately while minimizing false alarms, as depicted in Fig. 3. This trend confirms the ensemble's effectiveness in learning from increasing data volume, which is a critical trait for scalability in real-world deployments. The LN-SN-GN model also maintains the lowest false negative and false positive rates for all training sizes, thereby efficiently minimizing missed detections and classification mistakes, as reflected in Fig. 4. Low false negative rates are especially vital in applications such as mental health monitoring, where undetected distress may have serious implications. Supplemental objective metrics like F-measure, Matthews Correlation Coefficient, and Negative Predictive Value further reflect the balanced and accurate performance of the system on a wide variety of emotional categories, as presented in Fig. 5. These comprehensive results affirm the advantage of integrating multiple lightweight deep learners and a robust preprocessing pipeline for SER. These results collectively demonstrate that the ensemble method combining LN, SN, and GN along with robust feature extraction and fusion techniques significantly improves the classification accuracy of emotional speech data in large-scale environments.

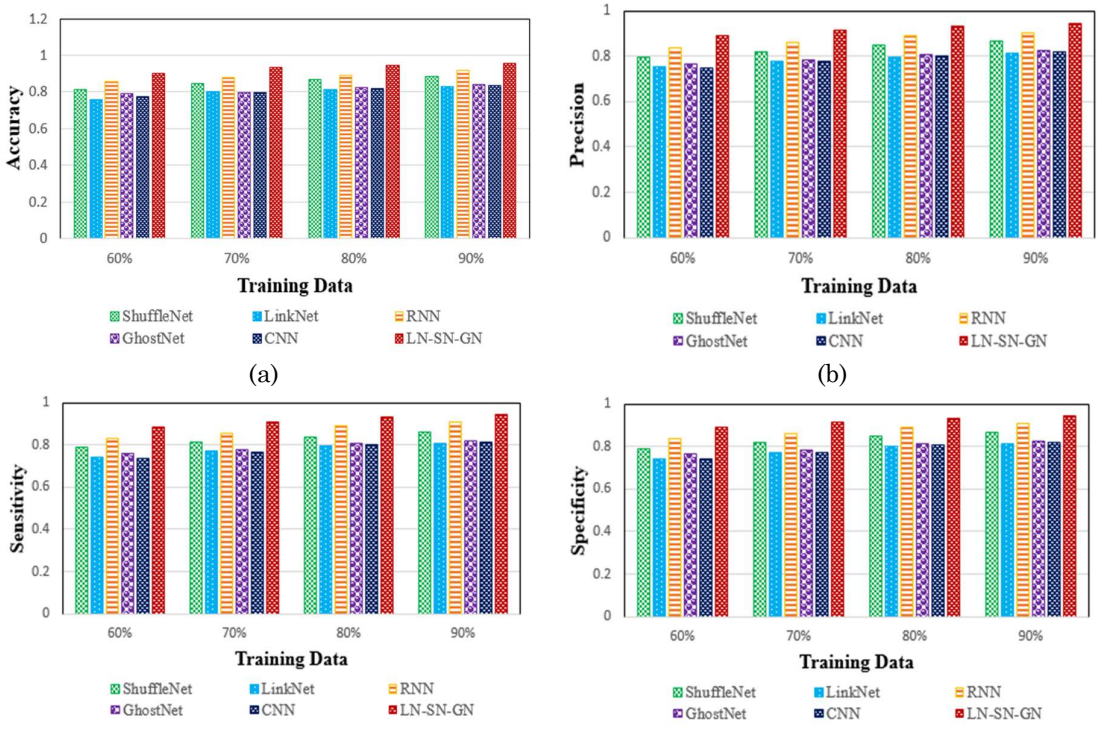
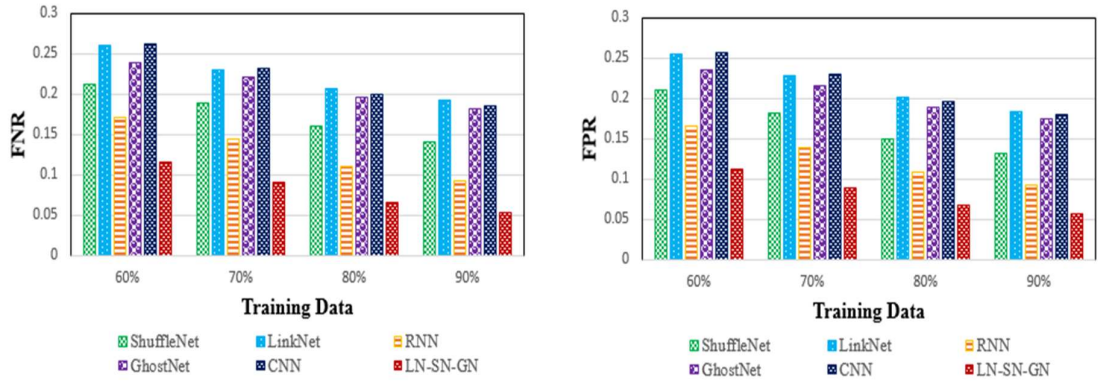
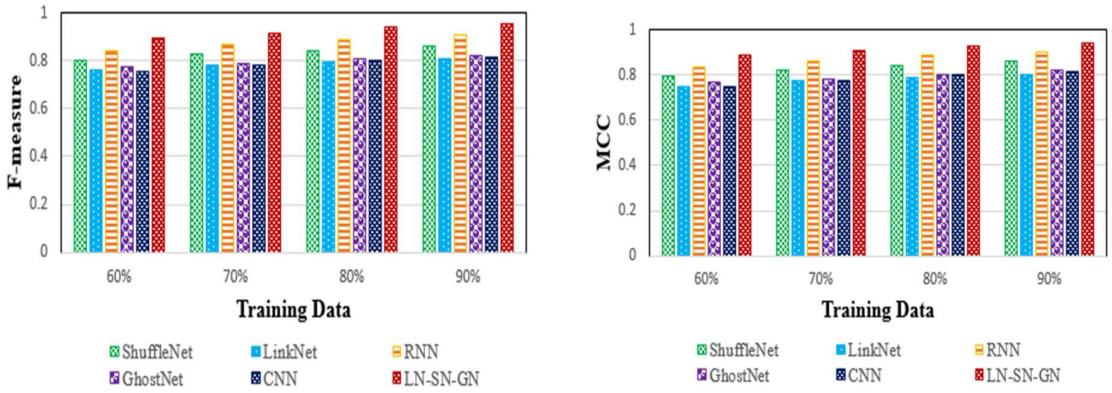


Fig. 3. *Positive Metric Assessment on LN-SN-GN and Conventional Methods a) Accuracy b) Precision c) Sensitivity and d) Specificity*

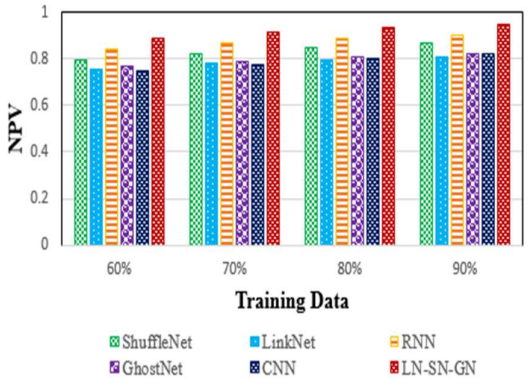


(a) (b)

Fig. 4. *Negative Metric Assessment on LN-SN-GN and Conventional Methods a) FNR and b) FPR*



(a) (b)



(c)

Fig. 5. *Neutral Metric Assessment on LN-SN-GN and Conventional Methods a) F-measure b) MCC and c) NPV*

4.4 Statistical Analysis on Accuracy

To further confirm the consistency and reliability of the performance of the suggested ensemble model (LN-SN-GN), A statistical analysis was conducted and contrasted with conventional classifiers such as ShuffleNet, LinkNet, RNN, CNN, and GhostNet. The statistical measures calculated are Minimum, Mean, Standard Deviation, Median, and Maximum values of classification accuracy for multiple runs with different random seeds and training-validation splits. As illustrated in Table I, the suggested LN-SN-GN ensemble outperformed all other models in each statistical aspect consistently. It recorded the highest mean accuracy of 0.947, followed by RNN (0.892) and ShuffleNet (0.861). Interestingly, the suggested model also recorded the lowest standard deviation of 0.025, reflecting better stability and generalization. This low standard deviation implies that the ensemble model yields reliable results even when subjected to variations in data partitions or initialization conditions. Its best accuracy level was 0.964, much higher than the rest, and its minimum value (0.910) as well as its median value (0.949) also indicate stable performance under all test conditions. These statistics collectively reveal a narrow range of accuracy

variation, demonstrating the ensemble model's robustness and consistent learning behavior. By contrast, baseline models such as CNN and LinkNet had larger variances, with larger standard deviations as well as lower minimum accuracy. Such inconsistency in performance highlights their vulnerability to fluctuations in training data composition and overfitting in certain scenarios. These findings combined confirm that the suggested combination of LinkNet, ShuffleNet, and GhostNet offers a statistically significant and high-performance approach to speech emotion recognition on big data. The statistical edge shown by the ensemble not only validates its superior architecture but also makes it a practical candidate for scalable, real-time emotion recognition systems.

Table 1. *Statistical Evaluation on Accuracy*

Statistical Metrics	ShuffleNet	LinkNet	RNN	CNN	GhostNet	LN-SN-GN
Mean	0.861	0.814	0.892	0.811	0.826	0.947
Standard Deviation	0.032	0.031	0.037	0.041	0.041	0.025
Maximum	0.896	0.855	0.926	0.856	0.871	0.964
Median	0.871	0.812	0.899	0.821	0.823	0.949
Minimum	0.834	0.779	0.863	0.776	0.788	0.921

4.5 Ablation Study

An ablation study is performed to reveal the individual and integrated contributions of spectral and statistical features in the proposed LN-SN-GN model. As illustrated from Table II, the proposed system combining both spectral and statistical features outperformed the single-feature configurations in all the evaluation measures consistently. The LN-SN-GN model produced the highest accuracy of 0.952, F-measure of 0.947, and MCC of 0.942, indicating high predictive accuracy and well-balanced classification. This clear performance gain confirms the complementary nature of spectral and statistical features, where spectral features capture fine-grained frequency information while statistical features provide global context over time. The integrated model also produced the lowest FPR of 0.075 and FNR of 0.069, showing its ability to minimize both error types in classification. These results unequivocally demonstrate that the integration of both feature sets enhances the model's ability to discriminate emotional classes towards more robust and legitimate emotion recognition outcomes. Moreover, the ablation study highlights the importance of multi-level feature representation in tackling the complex variability in emotional speech signals, further justifying the design choices in the feature extraction pipeline.

Table 2. *Ablation analysis*

Metrics	Proposed with Spectral Features	Proposed with Statistical Features	Proposed with both Spectral and Statistical features
Accuracy	0.922	0.889	0.952
FPR	0.097	0.132	0.075
F-measure	0.921	0.882	0.947
Sensitivity	0.919	0.864	0.949
FNR	0.099	0.158	0.069
Specificity	0.925	0.866	0.953
NPV	0.924	0.869	0.951
Precision	0.923	0.899	0.945
MCC	0.913	0.871	0.942

5. Conclusion

A novel ensemble-based approach (LN-SN-GN) to SER was introduced in this paper to improve the robustness and accuracy of speech emotion classification. The system was built as a three-stage pipeline, starting with noise removal by Wiener filtering, then spectral and high-level statistical feature extraction from preprocessed signals. An ensemble classification model of ShuffleNet, GhostNet, and LinkNet was then used, with the final prediction output achieved by averaging the predictions of each model. The ensemble method effectively leveraged the strengths of each architecture to achieve better generalization across different acoustic conditions. Experimental comparisons indicated that the proposed approach produced stable and effective emotion recognition performance, which rendered it a viable candidate for real-world SER usage. Furthermore, the lightweight nature of the chosen deep learning models makes the proposed system suitable for deployment in resource-constrained environments such as mobile and embedded devices. Future work could explore extending the ensemble to include transformer-based architectures or incorporating multimodal data inputs such as facial expressions or physiological signals to further enhance emotion recognition accuracy.

Compliance with Ethical Standards

Conflicts of interest: Authors declared that they have no conflict of interest.

Human participants: The conducted research follows the ethical standards and the authors ensured that they have not conducted any studies with human participants or animals.

References

[1] Huiyun Zhang, Puyang Zhao, Gaigai Tang, Zongjin Li, and Zhu Yuan, "Reproducible and generalizable speech emotion recognition via an Intelligent Fusion Network," *Biomedical Signal Processing and Control,* vol. 109, 2025.

[2] Enguerrand Boitel, Alaa Mohasseb and Ella Haig, "MIST: Multimodal emotion recognition using DeBERTa for text, Semi-CNN for speech, ResNet-50 for facial, and 3D-CNN for motion analysis," *Expert Systems with Applications,* vol. 270, 2025.

[3] Zengzhao Chen, Chuan Liu, Zhifeng Wang, Chuanxu Zhao, Mengting Lin, and Qiuyu Zheng, "MTLSER: Multi-task learning enhanced speech emotion recognition with pre-trained acoustic model," *Expert Systems with Applications,* vol. 273, 2025.

[4] Huiyun Zhang, Gaigai Tang, Heming Huang, Zhu Yuan, and Zongjin Li, "PulseEmoNet: Pulse emotion network for speech emotion recognition," *Biomedical Signal Processing and Control,* vol. 105, 2025.

[5] Xueliang Kang, "Speech emotion recognition algorithm of intelligent robot based on ACO-SVM," *International Journal of Cognitive Computing in Engineering,* vol. 6, pp. 131-142, 2025.

[6] Samir Sadok, Simon Leglaive, and Renaud Séguier, "A vector quantized masked autoencoder for audiovisual speech emotion recognition," *Computer Vision and Image Understanding,* vol. 257, 2025.

[7] G. Balachandran, S. Ranjith, G. C. Jagan, and T. R. Chenthil, "Advanced speech emotion recognition utilizing optimized equivariant quantum convolutional neural network for accurate emotional state classification," *Knowledge-Based Systems,* vol. 316, 2025.

[8] Weiquan Fan, Xiangmin Xu, Fang Liu, and Xiaofen Xing, "Multimodal speech emotion recognition via dynamic multilevel contrastive loss under local enhancement network," *Expert Systems with Applications,* vol. 281, 2025.

[9] Chengyan Du, Fu Liu, Bing Kang, and Tao Hou, "Speech emotion recognition based on spiking neural network and convolutional neural network," *Engineering Applications of Artificial Intelligence,* 2025.

[10] Bineetha Vijayan and M. V. Judy, "EmoFusionNet: A unified approach for robust speech emotion recognition," *Digital Signal Processing,* vol. 162, 2025.

[11] Mingduo Li, Jinhua Wang, Liwen Guo, Qinggang Meng, Mengqian Li, Jinliang Hou, Aoze Duan, and Haotian Sun, "A hybrid wavelet-wiener noise reduction algorithm for geomagnetic signals in dynamic positioning," *Alexandria Engineering Journal,* vol. 119, pp. 73-84, 2025.

[12] Siti Agrippina Alodia Yusuf and Risanuri Hidayat, "MFCC feature extraction and KNN classification in ECG signals," *In 2019 6th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE),* 2019.

[13] Meinard Müller, "Short-time Fourier transform and chroma features," *Lab Course, Friedrich-Alexander-Universität Erlangen-Nürnberg,* 2015.

[14] Buket Toptaş and Davut Hanbay, "Retinal blood vessel segmentation using pixel-based feature vector," *Biomedical Signal Processing and Control,* vol. 70, 2021.

[15] T. Ruba, R. Tamilselvi, M. Parisa Beham, and M. Gayathri, "Segmentation of a brain tumor using modified LinkNet architecture from MRI images," *Journal of Innovative Image Processing,* vol. 5, no. 2, pp. 161-180, 2023.

[16] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 6848-6856, 2018.

[17] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu, "Ghostnet: More features from cheap operations," *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* pp. 1580-1589, 2020.

[18] https://github.com/tobefans/LSSED.