Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.080025

ARTICLE

A Hybrid CNN–BiLSTM Framework for Speech Emotion Recognition with TimeGAN-Augmented Data and Contrastive Learning

Rashid Jahangir[1,*], Muhammad Asif Nauman[2], Oumaima Saidani[3] and Faisal Ramzan[2]

[1]Department of Computer Science, COMSATS University Islamabad, Vehari Campus, Vehari, Pakistan
[2]Riphah School of Computing & Innovation, Riphah International University, Lahore, Pakistan
[3]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

*Corresponding Author: Rashid Jahangir. Email: rashidjahangir@cuivehari.edu.pk

Received: 02 February 2026; Accepted: 15 April 2026; Published: 23 July 2026

ABSTRACT: Speech Emotion Recognition (SER) is a critical component of affective computing with broad applications in human–computer interaction, mental health monitoring, and intelligent multimedia systems. However, SER remains challenging due to the emotional ambiguity, lack of labeled data, class imbalance, and speaker variability. This study presents an effective SER framework that integrates contrastive representation learning, optimized spectrogram-based data augmentation, and selective synthetic data generation by using TimeGAN to enhance emotion classification performance. Contrastive learning enables the model to better discriminate acoustically similar emotions while Optuna automatically tunes augmentation strategies such as noise injection, time shifting, and time-frequency masking. Unlike existing approaches that apply synthetic generation uniformly across all classes, the proposed method targets only confusing or under-represented emotion classes to preserve the inter-class separability. A CNN-BiLSTM architecture is used to extract spectral and temporal information of the speech. The framework is evaluated with benchmark SER datasets—EMO-DB and RAVDESS—under speaker independent protocols. Experimental results demonstrate improved accuracy, robustness, and generalization under limited and imbalanced data conditions, supported by confusion matrices, UMAP, and t-SNE visualizations.

KEYWORDS: Speech emotion recognition; data augmentation; optuna; TimeGAN; synthetic data; contrastive learning

1 Introduction

Speech signals serve as an effective mode of interaction that conveys valuable information, such as gender, emotional state, pitch, and other distinctive speaker characteristics. These traits enable discrimination of a speaker's emotional state during remote voice communication. By leveraging these characteristics, algorithms can be trained to identify emotional expression in speech similar to human perception. The problem of predicting the emotional state of a person based on information from the speech is called speech emotion recognition (SER)—a topic that has attracted growing research interest in recent years. SER has found utility across a range of real-world applications. In healthcare, for instance, a patient's emotional state can be assessed through speech analysis [1]. Another significant application lies in intelligent driving systems, where SER enables real-time detection of a driver's emotional state [2]. Furthermore, SER has been widely adopted in call centers, automatic translation systems, and broader human–machine interaction contexts [3].

Numerous machine learning (ML) classification methods have been proposed to predict emotions from speech, including sadness, happiness, anger, and fear. Based on the modality used for emotion inference, existing studies can be broadly categorized into speech-based, facial-expression-based, and video-streaming-based approaches. Video-based systems detect emotions by identifying emotionally salient frames and uniformly sampling them for analysis [4]. In real-world applications, computer systems are increasingly required to identify human emotions from facial expressions. This approach involves the extraction of discriminative features from image datasets, constructing a master feature, and training an ML classifier [5]. Nevertheless, visual disturbances in emotional frames significantly degrade the performance of such systems. Moreover, the formation of excellent image or video datasets that effectively reflect emotion expressions is laborious and time-consuming [6]. In contrast, audio-based techniques are better suited to real-time SER. These systems are cost-effective, computationally efficient, and largely language-independent.

SER involves several sequential stages to identify emotional states as illustrated in Fig. 1. Signal pre-processing is the initial stage, required to develop a robust and effective system. Pre-processing involves pre-emphasis, silence, noise reduction, and speech framing to samples [7], collectively assisting to normalize the signal and reduce its dynamic range. The second stage is feature extraction, wherein handcrafted acoustic features are derived from the processed speech signals to construct a numerical feature vector for model training. These handcrafted acoustic characteristics can be broadly grouped into spectral, prosodic, Teager Energy Operator (TEO)-based, and qualitative features. Following feature extraction, feature selection is applied to identify the most discriminative attributes and construct a compact master feature vector for ML classifier training. Feature selection further enhances system efficiency and reduces classification error by eliminating redundant or irrelevant features. In the final stage, an ML or deep learning (DL) model is trained to classify emotional states based on the selected feature vector. Established classification methods applied in SER include Gaussian Mixture Models (GMM) [8], Support Vector Machines (SVM) [9], and Hidden Markov Models (HMM) [10]. Feature extraction is the most critical stage of the SER pipeline, as it directly impacts recognition performance. This stage aims to extract discriminative features that effectively represent a speaker's emotional state. Although extensive research has concentrated on hand-crafted feature extraction to enhance SER performance [11], such approaches rely heavily on domain-specific expert knowledge. Consequently, their limited generalizability poses a major challenge to developing a more accurate and multilingual SER system.

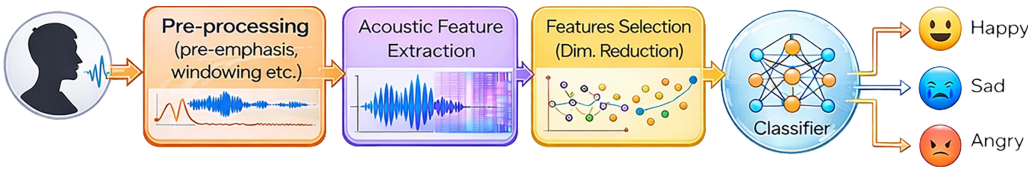


Figure 1: Classical speech emotion recognition process.

TimeGAN has demonstrated effectiveness in generating realistic synthetic speech features while preserving temporal dependencies, making it well-suited for addressing data imbalance in SER. However, the integration of TimeGAN with contrastive learning within SER frameworks remains largely unexplored. Contrastive learning enhances intra-class compactness and inter-class separability, thereby reducing confusion between acoustically similar emotions [12]. Motivated by this gap, this work proposes an integrated SER framework that combines optimized data augmentation, targeted TimeGAN-based synthetic data generation for minority and confusing emotion classes and supervised contrastive learning for robust feature representation. A CNN–BiLSTM architecture is employed to capture both spectral and temporal speech

characteristics. Experiments on benchmark SER datasets confirm improved robustness and generalization under imbalanced and limited data conditions.

To overcome the shortcomings of ML methods, recent research has investigated the use of speech-based feature extraction via the use of DL methods, which are much less reliant on human effort. DL can automatically learn salient and hierarchical representations of raw speech signals and consequently has grown to be more significant in SER studies [13,14]. The typical architectures of the DL are convolutional neural networks (CNNs), restricted Boltzmann machines (RBMs), deep autoencoders, and recurrent neural networks (RNNs). These models can be layered to form better classification accuracy, robustness and flexibility without depending on expert-crafted features. Recent research indicates that the performance of SER can be enhanced by optimization of data augmentation parameters to particular emotion classes, especially emotions that are often confused with others. TimeGAN has been shown to be effective in realistic speech features synthesis and maintaining the time relations, thus it is applicable to tackle data imbalance in SER. Nevertheless, a combination of TimeGAN and contrastive learning in SER is still limited. Contrastive learning improves the intra-class compactness, inter-class separability, and minimizes confusion between acoustically related emotions [15]. Driven by this limitation, this study proposes a unified SER architecture that integrates data augmentation optimization, controlled TimeGAN-like synthesizing minority and confusing emotion classes and contrastive learning in supervision of strong features representation. To obtain the characteristics of spectral and temporal speech, a CNN-BiLSTM architecture is used. Experiments on benchmark SER datasets show increased robustness and generalization in imbalanced and limited data settings. The key contributions of this study are:

1. This paper presents an Optuna-driven search-based approach that dynamically identifies the optimal augmentation hyperparameters for noise injection, time shifting, and frequency masking.
2. The research solves the common problem of class imbalance and confusable emotion regions through a simplified TimeGAN module. The method generates 250 high-fidelity synthetic samples per targeted class by selectively augmenting emotion classes with recall less than 0.93, which greatly enlarges the decision boundaries of the model with challenging emotions.
3. This study incorporates a contrastive pre-training phase to equip the encoder with robust, speaker-invariant emotional representations by maximizing agreement between differently augmented views of the same signal, prior to supervised fine-tuning.

The remainder of the paper is organized as follows. Section 2 presents literature on SER, data augmentation, and DL techniques. Section 3 introduces the proposed SER framework, which comprises the data augmentation, TimeGAN-based data synthesis, and contrastive learning. Section 4 presents the performance analysis and results. Lastly, Section 5 concludes the paper and provides future work.

2 Related Word

Recent advances in SER have largely focused on improving performance under limited and imbalanced data conditions [11,16]. Emotional speech datasets are costly to collect and annotate, while certain emotions occur far less frequently than others [17]. As a result, many SER systems struggle to generalize, particularly when trained on small or biased corpora such as IEMOCAP [18], EMO-DB [19], and RAVDESS [20]. To address this limitation, data augmentation has been widely adopted as a primary strategy [21]. Widely used techniques include additive noise, pitch shifting, time stretching, and speed perturbation. These techniques increase data diversity but often yield only marginal performance improvements. However, uniform application across all emotion classes can distort emotion-specific temporal patterns and introduce unrealistic samples, particularly for perceptually ambiguous emotions such as sadness and neutrality.

To overcome these limitations, several studies have proposed generative models for synthesizing emotional speech features [22]. GAN-based methods have been applied to generate synthetic spectrograms or latent representations to balance emotion class distributions. The study in [23] demonstrated that GAN-synthesized spectrograms improved emotion recognition performance on IEMOCAP by augmenting minority class samples. Along similar lines, Reference [24] employed Wasserstein GANs for SER and reported consistent accuracy improvements on EMO-DB. Although these results are promising, generative models are often sensitive to hyperparameter tuning and computationally expensive. Moreover, synthetic samples do not always faithfully capture fine-grained emotional dynamics, limiting their effectiveness in real-world deployments.

Representation learning techniques, including contrastive learning, have recently gained considerable attention in SER. These methods aim to learn discriminative embeddings by pulling together emotionally similar samples while pushing apart dissimilar ones. Supervised contrastive learning and contrastive predictive coding have demonstrated strong performance on emotional speech tasks, particularly when integrated with log-Mel spectrogram features and applied to tasks with lower emotional complexity [25]. In practice, contrastive learning enhances the robustness and separability of class representations. Nonetheless, it does not directly address data scarcity or class imbalance, as it operates on existing samples rather than generating new ones.

On the modeling side, hybrid DL architectures are currently the most widely adopted in SER. CNN-based models are effective at extracting local time–frequency patterns from spectrograms, while recurrent models capture long-range temporal dependencies. CNN-BiLSTM models, specifically, have become a standard baseline due to their trade-off between expressive power and training stability. Prior studies have demonstrated the superiority of CNN-BiLSTM models over standalone CNNs or LSTMs on benchmark datasets including IEMOCAP, EMO-DB, and RAVDESS [26,27].

Recent developments in SER have shifted from standard supervised learning to contrastive frameworks that explicitly model the relationships between emotional samples. Traditional contrastive methods often treat all samples with the same label as equal positive pairs. However, recent state-of-the-art approaches argue that this strategy ignores the inherent data structure and the varying difficulty of different emotional pairs. Reference [12] proposed contrastive self-alignment learning that used an audio-text contrastive loss to align heterogeneous features in a shared temporal space. This alignment ensured that the timing of emotional expression in rhythm and pitch matches the semantic intent and reduces the impact of modality-specific noise. Reference [28] developed the focused contrastive learning network to address the challenge of limited and imbalanced SER datasets. This approach refined the contrastive loss by ascribing higher penalty weights to 'harder-to-identify' categories. The model's attention was focused on subtler category differences and the spatial distribution of samples. The proposed model yielded a more compact and discriminative feature representation compared to standard contrastive loss formulations.

Despite these advances, several challenges remain. Many augmentation methods fail to preserve emotional consistency. Generative models may introduce artifacts or require extensive hyperparameter tuning. Contrastive learning improves the quality of embeddings without increasing data diversity. Furthermore, conventional metrics such as accuracy and F1-score provide limited insight into the quality of synthetic data or the robustness of learned representations. These limitations motivate the need for a hybrid framework that integrates optimized augmentation, targeted synthetic data generation, and contrastive representation learning within a unified CNN-BiLSTM model. Such a framework can enhance generalization at a lower cost of manual design selection, which is essential in the case of reliable SER systems.

3 Proposed Methodology

This section describes the proposed SER methodology, illustrated in Fig. 2. Initially, data were collected from three benchmark datasets: EMO-DB, RAVDESS, and SAVEE. The collected speech data then underwent a pre-processing phase. To handle the class imbalance problem, Bayesian optimization was employed to determine optimal augmentation parameters. Additionally, synthetic speech data were generated using a TimeGAN model to improve class representation. The prepared data were subsequently fed into a CNN-BiLSTM model for emotion classification. A contrastive learning model was also incorporated for self-supervised pre-training, followed by fine-tuning with labeled data to improve feature discrimination. Finally, the trained models were evaluated using standard performance metrics. Each phase is elaborated in subsequent subsections.

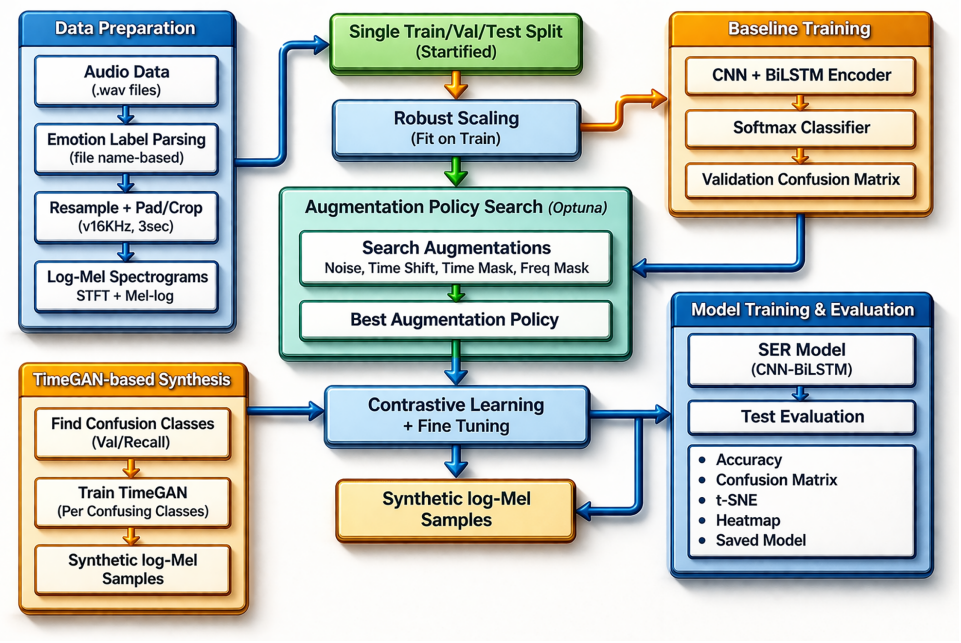


Figure 2: Workflow of proposed research methodology.

3.1 Dataset Collection

Two widely used SER datasets were employed in this study: EMO-DB and RAVDESS. EMO-DB [19] is a German emotional speech database developed at the Institute of Communication, Technical University Berlin. It comprises 535 audio records of different durations from 10 professional speakers, five men and five women, recorded in seven emotion categories: angry, boredom, disgust, fear, happy, neutral and sad as indicated in Fig. 3a. All audio files in EMO-DB are recorded at 16 kHz in WAV format with 16-bit depth and mono-channel configuration. The second dataset is RAVDESS [20], selected for its wide availability and rich speech content. RAVDESS comprises 1440 audio files that are recorded by 12 male and 12 female actors reading English scripts. This dataset is partitioned into eight emotion categories of angry, calm, disgust, fear, neutral, happy, sad and surprised as illustrated in Fig. 3b. The samples are 16-bit and 48 kHz.

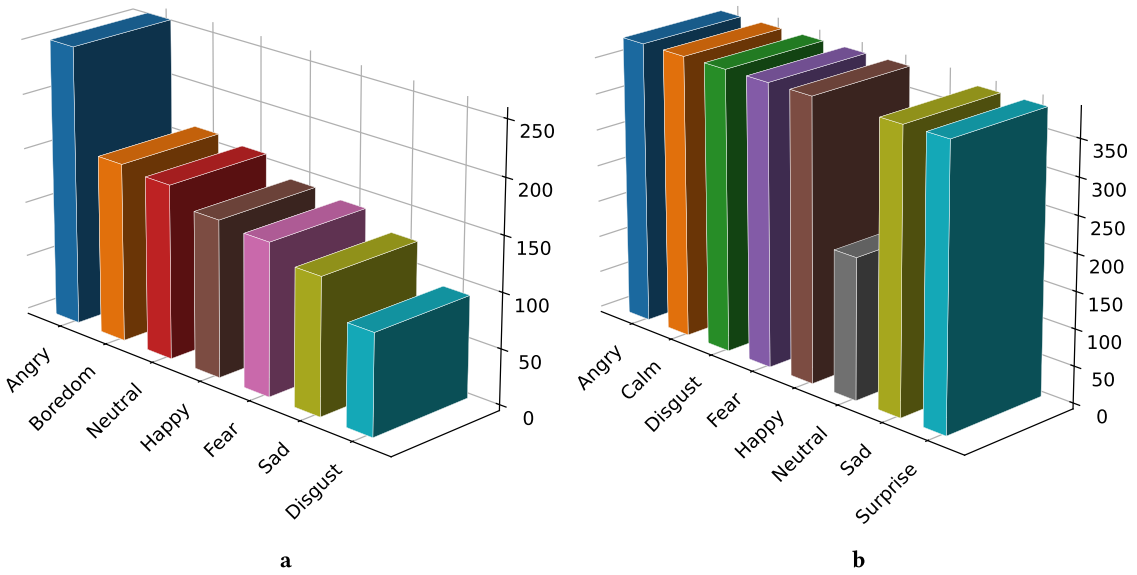


Figure 3: Emotions distribution of the SER datasets. (**a**) EMO-DB, while (**b**) RAVDESS.

3.2 Feature Extraction

The proposed framework transforms raw audio signals into an informative representation domain to extract meaningful acoustic features. Each waveform was processed to emphasize the perceptually relevant frequency patterns to focus on the temporal and spectral properties that hold significant importance in emotion recognition. All waveforms were converted to a log-Mel spectrogram, to enable the model to capture subtle acoustic details like pitch, energy and timbre. All waveforms were converted to log-Mel spectrograms to enable the model to capture subtle acoustic cues such as pitch, energy, and timbre. This not only reduces input dimensionality but also concentrates representation capacity on the frequency bands most informative for speech analysis. Mathematically, every waveform $x[n]$ was framed first with a Hanning window $w[n]$ of length $L = 400$ and hop size $H = 160$. Short-time Fourier Transform (STFT) was calculated as:

$$X[m,k] = \sum_{n=0}^{L-1} x[n+mH] \cdot w[n] \cdot e^{-j2\pi kn/N_{FFT}}, \quad N_{FFT} = 512$$

The power spectrogram was then mapped to $N_{mel} = 64$ Mel bands using a Mel filterbank $M_{f,k}$:

$$S[m,f] = \sum_{k=0}^{N_{FFT}/2} |X[m,k]|^2 \cdot M_{f,k}, \quad f = 1,\ldots,N_{mel}$$

Finally, the logarithm was applied to compress the dynamic range:

$$\text{LogMel}[m,f] = \log(S[m,f] + \epsilon), \quad \epsilon = 10^{-6}$$

To ensure consistency across all samples, audio files were loaded via the soundfile library, resampled to 16 kHz, converted to mono, and normalized to 16-bit precision. Each waveform was either padded or cropped to a fixed duration of 3 s. Log-Mel spectrograms were then extracted using 64 Mel bands, a window size of 400 samples, a hop length of 160 samples, and an FFT size of 512, with frequency limits set between 50 and

7600 Hz. Finally, the logarithm of the log-Mel spectrogram (Fig. 4) was computed to compress the dynamic range, producing the final feature tensor for model training.

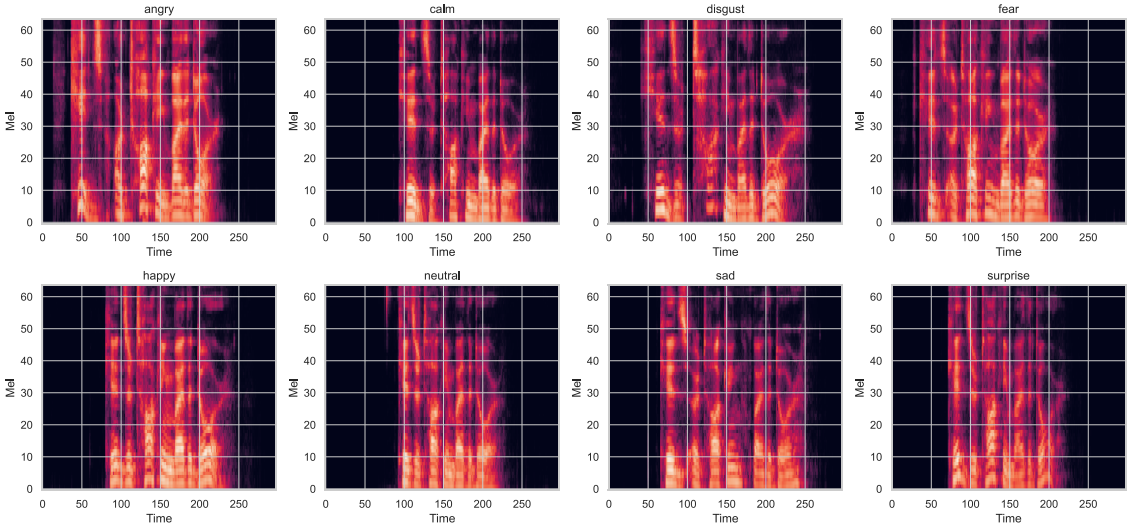


Figure 4: Log-Mel spectrograms from the RAVDESS dataset, with time–frequency patterns of different emotion classes. The x-axis represents time frames (10 ms per frame), and the y-axis represents 64 Mel-frequency bins.

3.3 Data Augmentation

To improve the model robustness and mitigate overfitting on limited and imbalanced datasets, the proposed framework applies data augmentation directly to the log-Mel spectrograms. Augmentation introduces controlled variations into the training data, simulates real-world acoustic conditions, and increases the diversity of emotional speech patterns. Four types of augmentations were employed: Gaussian noise, time shifting, time masking, and frequency masking. Gaussian noise with a standard deviation of up to 0.08 was added to simulate background interference and microphone variability. For a log-Mel spectrogram $X \in \mathbb{R}^{T\times M}$, Gaussian noise augmentation was defined as:

$$X' = X + \mathcal{N}(0, \sigma^2), \quad \sigma \le 0.08$$

Time shifting displaces the spectrogram along the temporal axis by up to 15% of its total length, enabling the model to learn invariance to minor temporal misalignments in speech:

$$X'_{t,m} = X_{(t+\Delta t) \bmod T,\, m}, \quad \Delta t \sim \mathcal{U}(-0.15T, 0.15T)$$

Time masking sets contiguous time frames to zero with a probability of up to 25%. This operation was expressed as:

$$X'_{t,m} = 0, \quad t \in [t_0, t_0 + L_t], \quad L_t \le 0.25T$$

Similarly, frequency masking was applied along the Mel frequency axis, masking up to 25% of the Mel bands to model missing spectral information:

$$X'_{t,m} = 0, \quad m \in [m_0, m_0 + L_m], \quad L_m \le 0.25M$$

Augmentation parameters were automatically optimized using Optuna, a Bayesian optimization framework that evaluated 20 trials to identify the combination and magnitude of augmentations maximizing validation accuracy. Let $\theta = \{\sigma, \Delta t, L_t, L_m\}$ denote the augmentation parameter set. The optimization was computed as:

$$\theta^* = \arg\max_{\theta} \ \mathrm{Acc}_{val}\left(f\left(X_{\mathrm{train}}^{\mathrm{aug}}(\theta), y_{\mathrm{train}}\right)\right)$$

where $f(\cdot)$ denotes the CNN–BiLSTM classifier and Acc_{val} represents the accuracy on the validation data. Optuna employed a Tree-structured Parzen Estimator (TPE) sampler to navigate the augmentation search space efficiently. Once optimal parameters were identified, the augmented spectrograms were incorporated into the initial training set. This approach exposed the model to a broader range of acoustic variations during training, thereby enhancing generalization to previously unseen emotional expressions. By operating at the spectrogram level, the framework preserved the temporal and spectral structure of speech signals while introducing natural variations that simulate real-world noise and acoustic distortions. Overall, the augmentation strategy enhanced model robustness, reduced overfitting, and improved classification accuracy across diverse emotional categories.

3.4 Synthetic Data Generation Using TimeGAN

Synthetic data generation was incorporated into the proposed framework through a simplified variant of TimeGAN to target confusing emotion classes and address class imbalance. This module aimed to generate realistic log-Mel spectrogram sequences that faithfully preserve the temporal and spectral characteristics of emotional speech. Let $\mathbf{X} \in \mathbb{R}^{T \times M}$ denote a sequence of log-Mel spectrograms, where T is the total number of time frames and M denotes the number of Mel frequency bands. TimeGAN was applied directly to these spectrogram sequences after feature extraction and normalization. Synthetic data generation was selectively performed for emotion classes with low recall on the validation set, identified as confusing classes through confusion matrix analysis. The TimeGAN architecture consisted of three main components: an embedder network $E(\cdot)$, a recovery network $R(\cdot)$, and a generator network $G(\cdot)$. The embedder mapped the input spectrogram sequence into a latent representation $\mathbf{H} = E(\mathbf{X})$, where $\mathbf{H} \in \mathbb{R}^{T \times d}$ is a latent temporal embedding. The recovery network reconstructed the original spectrogram from the latent space as $\hat{\mathbf{X}} = R(\mathbf{H})$, and the embedder–recovery pair was pretrained as an autoencoder by minimizing the reconstruction loss $\mathcal{L}_{\mathrm{AE}} = \|\mathbf{X} - \hat{\mathbf{X}}\|_2^2$. After autoencoder pretraining, a generator network was trained to model the distribution of latent embeddings. Given a noise sequence $\mathbf{Z} \sim \mathcal{N}(0,1)$, the generator produced synthetic latent representations $\tilde{\mathbf{H}} = G(\mathbf{Z})$. The generator in this framework was optimized using a supervised latent loss $\mathcal{L}_G = \|\mathbf{H} - \hat{\mathbf{H}}\|_2^2$ and a reconstruction loss from the decoder network. This approach prioritized the preservation of temporal dependencies in the latent space over traditional adversarial competition. Once trained, synthetic latent sequences were passed through the recovery network to generate artificial log-Mel spectrograms $\tilde{\mathbf{X}} = R(\tilde{\mathbf{H}})$. A fixed number of synthetic samples was generated for each selected confusing class and subsequently normalized using the same scaler fitted on the original training data. These synthetic spectrograms were then merged with the original training set to form an augmented dataset for model training. The simplified TimeGAN preserved the sequential structure of the speech features but added the realistic variation by synthesizing synthetic data in the latent temporal space. This approach enhanced training sample diversity for confusing emotion classes, mitigated class imbalance, and ultimately improved the generalization performance of the proposed SER model.

3.5 Contrastive Learning for Speech Emotion Representation

The self-supervised pre-training phase operated on the combined training split (original, augmented, and synthetic data) without using class labels. The encoder was trained to learn emotion-discriminative features by maximizing agreement between two stochastically augmented views of the same audio segment using the NT-Xent loss. This process was implemented as a two-stage pipeline: stage one focused on representation learning via contrastive pre-training, while stage two involved supervised fine-tuning for emotion classification. The objective of this phase was to learn robust and emotion-aware embeddings from log-Mel spectrogram sequences, which might enhance generalization to downstream emotion-classification tasks. Crucially, this phase relied exclusively on the training partitions of the target datasets (EMO-DB, RAVDESS) to ensure the learned features were domain-specific without the need for an external large-scale corpus. Let $\mathbf{X}_i$ and $\mathbf{X}_j$ denote two augmented views of the same speech segment, generated via time-domain and frequency-domain augmentations such as time masking, frequency masking, or additive noise. These augmented spectrograms were passed through the shared encoder network $f_\theta(\cdot)$ to obtain latent embeddings.

$$\mathbf{z}_i = f_\theta(\mathbf{X}_i), \quad \mathbf{z}_j = f_\theta(\mathbf{X}_j), \tag{1}$$

where $\mathbf{z}_i, \mathbf{z}_j \in \mathbb{R}^d$ are d-dimensional embeddings.

The embeddings were normalized onto a unit hypersphere to facilitate the cosine similarity computation:

$$\tilde{\mathbf{z}}_i = \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|_2}, \quad \tilde{\mathbf{z}}_j = \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|_2}. \tag{2}$$

The contrastive loss was then computed using the normalized temperature-scaled cross-entropy loss (NT-Xent), which encouraged embeddings from the same speech segment to be close while pushing apart embeddings from different segments:

$$\mathcal{L}_{\text{contrastive}} = -\log \frac{\exp\left(\text{sim}(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_j)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{k \neq i} \exp(\text{sim}(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_k)/\tau)} \tag{3}$$

where $\text{sim}(\mathbf{a}, \mathbf{b}) = \mathbf{a}^\top \mathbf{b}$ represents the cosine similarity, τ is a temperature hyperparameter (set to 0.2 in experiments), and N is the number of speech segments in the batch. After contrastive pretraining, the encoder weights were fine-tuned on the downstream emotion classification task using the standard cross-entropy loss. Through contrastive learning, the learned embeddings captured fine-grained inter-emotion differences while remaining robust to speaker-specific variations and background noise. This approach enhanced the separability of confusing emotion classes in the latent space and helped in quicker convergence and increased validation in the process of supervised training.

3.6 CNN-BiLSTM for Emotion Classification

Following synthetic data augmentation and contrastive representation learning, the proposed framework employs a hybrid CNN-BiLSTM architecture to jointly capture spatial and temporal patterns in the log-Mel spectrogram sequences. The convolutional layers extract local time-frequency features—each followed by batch normalization—while the bidirectional LSTM layers modeled long-range temporal dependencies crucial for recognizing emotion dynamics. Let $\mathbf{X} \in \mathbb{R}^{T \times M}$ denote an input log-Mel spectrogram sequence, where T represents the number of time frames and M is the number of Mel frequency bands.

The CNN module applied a series of 2D convolutional filters with the first layer using a 5 × 5 kernel and the second using a 3 × 3 kernel to capture local patterns:

$$\mathbf{F}_c = \mathrm{ReLU}(\mathrm{Conv2D}(\mathbf{X}, \mathbf{W}_c) + \mathbf{b}_c), \tag{4}$$

where $\mathbf{W}_c$ and $\mathbf{b}_c$ are the convolutional kernel weights and biases, and $\mathbf{F}_c$ denotes the extracted feature maps. Max-pooling layers followed the convolutional layers to reduce feature dimensionality while retaining the most salient patterns. The resulting feature maps were reshaped and fed into a bidirectional LSTM (BiLSTM) with 64 units per direction, for a total of 128 hidden units to model temporal dependencies in both forward and backward directions:

$$\mathbf{H}_t = \mathrm{BiLSTM}(\mathbf{F}_{c,t}), \quad t = 1, \ldots, T, \tag{5}$$

where $\mathbf{H}_t \in \mathbb{R}^{2h}$ is the concatenated hidden state at time t from both forward and backward LSTMs, with $h = 64$ in each direction. A global pooling layer summarized the temporal features into a fixed-length vector, which was passed through a dense layer with 128 units and ReLU activation, followed by a dropout layer with a rate of 0.4 and a softmax activation to produce class probabilities:

$$\hat{\mathbf{y}} = \mathrm{softmax}(\mathbf{W}_d \mathbf{H}_{\mathrm{pooled}} + \mathbf{b}_d), \tag{6}$$

where $\mathbf{W}_d$ and $\mathbf{b}_d$ are the weights and biases of the dense layer, and $\hat{\mathbf{y}} \in \mathbb{R}^C$ is the predicted probability distribution over C emotion classes. The CNN-BiLSTM architecture effectively captures both short-term spectral patterns and long-term temporal dependencies in emotional speech. By combining convolutional feature extraction with bidirectional sequential modeling, the network achieves robust emotion recognition performance across both clear and confusing emotion categories, particularly when trained on the augmented dataset containing synthetic and contrastively pre-trained representations.

As shown in Table 1, the CNN-BiLSTM consists of two 2D convolutional layers with 32 and 64 filters for extracting local spectral features, each followed by batch normalization and 2 × 2 max-pooling. The resulting feature maps are reshaped and passed through a single BiLSTM layer with 64 units per direction (128 total) to model temporal dependencies in both forward and backward directions. The dense layer with 128 units (ReLU) and the softmax output layer serve as the classification head, mapping the temporal embeddings to emotion class probabilities. A dropout of 0.4 is applied after the encoder to prevent overfitting. This architecture, combined with data augmentation and contrastive feature learning, provides robust recognition of confusing emotion classes.

Table 1: Selected hyperparameters for CNN-BiLSTM model training.

Hyperparameter	Value(s)
Input Feature Dimension	(298, 64)
Batch Size	64
Conv2D Layers	2 (32 filters with 5 × 5 kernel, 64 filters with 3 × 3 kernel)
Activation Functions	ReLU (hidden layers), Softmax (output layer)
Batch Normalization	After each Conv2D layer
Max Pooling	2 × 2 after each Conv2D layer
BiLSTM Units	64 units per direction (total 128)
Dropout Rate	0.4

(Continued)

Table 1 (continued)

Hyperparameter	Value(s)
Dense Layer Units	128 (ReLU)
Loss Metric	Sparse Categorical Cross-Entropy
Optimizer	Adam
Epochs	80
Learning Rate	0.001
Data Split Ratio	70% Training:15% Testing:15% Validation

3.7 Evaluation Metrics

The performance of the proposed model is evaluated using standard metrics for multi-class SER. The confusion matrix is computed to visualize the model's predictions, where each element $C_{i,j}$ represents the number of samples of true class i predicted as class j. The diagonal elements represent correctly classified instances, while off-diagonal elements reveal misclassifications, thereby identifying confusing classes that can guide synthetic sample generation. Overall accuracy was calculated as the proportion of correctly predicted samples over the total number of predictions:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{TP + TN}{TP + FP + TN + FN}\right)_i \tag{7}$$

Accuracy provides a quick assessment of the model's general performance on the test set. To capture class-specific performance, recall was computed for each emotion class as

$$Recall = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{TP}{TP + FN}\right)_i \tag{8}$$

and is used to identify confusing classes with low recall, which informs further augmentation. The F1-score, which is the harmonic mean of precision and recall, was calculated for each class as

$$F1 - score = \frac{1}{N}\sum_{i=1}^{N} 2 \times \left(\frac{Precision \times Recall}{Precision + Recall}\right)_i \tag{9}$$

Both per-class F1 and macro F1 (average across all classes) are reported to provide a balanced evaluation, accounting for both frequent and rare emotion classes. A comprehensive classification report was generated, summarizing precision, recall, F1-score, and support for each class, and is saved for reproducibility. Additionally, visualizations including confusion heatmaps and training curves are generated to interpret model performance, monitor convergence, and detect overfitting. These metrics collectively enable a comprehensive evaluation of the model's capacity to recognize both clear and ambiguous emotional expressions in speech.

4 Experimental Results and Discussion

This section presents the experimental results of the proposed SER framework evaluated on the EMO-DB, and RAVDESS datasets. To ensure an unbiased evaluation, a strict 70%–15%–15% stratified split was adopted to reflect a realistic deployment scenario. The 'confusing classes' for TimeGAN synthesis were treated as a data-level hyperparameter tuning step solely based on the validation set recall. The test set remained completely isolated and was never used to decide which classes required synthetic augmentation or to tune

the Optuna policies. This experimental setting ensured an unbiased final evaluation. The evaluation analyzes the effect of optimized spectrogram augmentation, TimeGAN-based synthetic data generation for confusing emotion classes, and contrastive pretraining. The proposed SER framework used the Adam optimizer with a learning rate of 0.001, a batch size of 64, and up to 80 epochs. Early stopping with a patience of 6 epochs was applied to prevent overfitting. Experiments were conducted on a workstation equipped with an Intel Core i7 CPU, 32 GB RAM, and an NVIDIA GPU, using Python 3.9 and TensorFlow/Keras along with supporting libraries such as NumPy, scikit-learn, and Matplotlib. Performance is assessed using accuracy, precision, recall, and confusion matrices to demonstrate consistent improvements across successive stages of the proposed pipeline.

4.1 Performance Evaluation Using EMO-DB

This subsection presents a comparative analysis of four model configurations evaluated on the EMO-DB dataset: the baseline, augmented, synthetic, and final contrastive models. The baseline model, trained exclusively on the original speech data, achieved an accuracy of 18.25% within a restricted training window of 20 epochs. This performance reflects a "cold-start" scenario in which the CNN-BiLSTM model converged to a local minimum due to the high-dimensional feature variance of the EMO-DB dataset. Further experiments were conducted by extending the baseline training to 100 epochs to verify the model's learning capacity. Under this condition, the baseline converged to 75.0% accuracy, as shown in the confusion matrix in Fig. 5a. The figure reveals that the model still struggled to discriminate effectively between emotional states, resulting in high misclassification rates across all classes. This suggests that the limited size of the EmoDB dataset is insufficient to train deep representations without additional regularization or data expansion. The augmented model incorporated Optuna-optimized data augmentation techniques to enhance sample diversity. This configuration yielded a substantial performance improvement, achieving an accuracy of 89.78%. The confusion matrix in Fig. 5b shows a clearer diagonal structure, indicating that augmentation effectively reduced overfitting and helped the model capture robust acoustic features for distinct emotions. The synthetic data model employed TimeGAN to generate high-fidelity synthetic speech samples to address class imbalance and data scarcity. This approach further improved the accuracy to 94.16%. The confusion matrix in Fig. 5c reflects this improvement through reduced off-diagonal errors and greater consistency in recognizing subtle emotional nuances, compared to augmentation alone. The final contrastive model integrates optimized augmentation, synthetic data, and contrastive learning-based fine-tuning. This configuration achieved the highest accuracy of 95.03%. The confusion matrix in Fig. 5d exhibits a strong diagonal concentration with minimal misclassification, confirming that contrastive learning successfully structured the feature space to maximize inter-class separability. The integration of the proposed modules demonstrated clear incremental gains. As shown in the ablation results, the progression from the baseline (75.0%) to the final framework (98.50%) represents a substantial 23.50% absolute improvement. This improvement is attributed to the synergy between Optuna-optimized augmentations, which provided essential regularization, and TimeGAN synthesis, which specifically targeted confusing classes to preserve inter-class separability. Finally, the supervised contrastive learning phase refined the feature space and allowed the model to bridge the gap between standard convergence and state-of-the-art performance.

Fig. 6b presents the detailed classification report (Precision, Recall, F1-Score) for the final model. The system achieved perfect recognition for "Disgust" and "Sadness," with F1-scores reaching 1.00. However, the "Happy" class presented a slight challenge, showing a recall of 0.76 despite high precision. This is likely due to the acoustic similarity between high-arousal emotions such as "Happiness" and "Anger," a common challenge in SER tasks. Nevertheless, the overall accuracy and F1-scores across the majority of classes remain consistently high (>0.95).

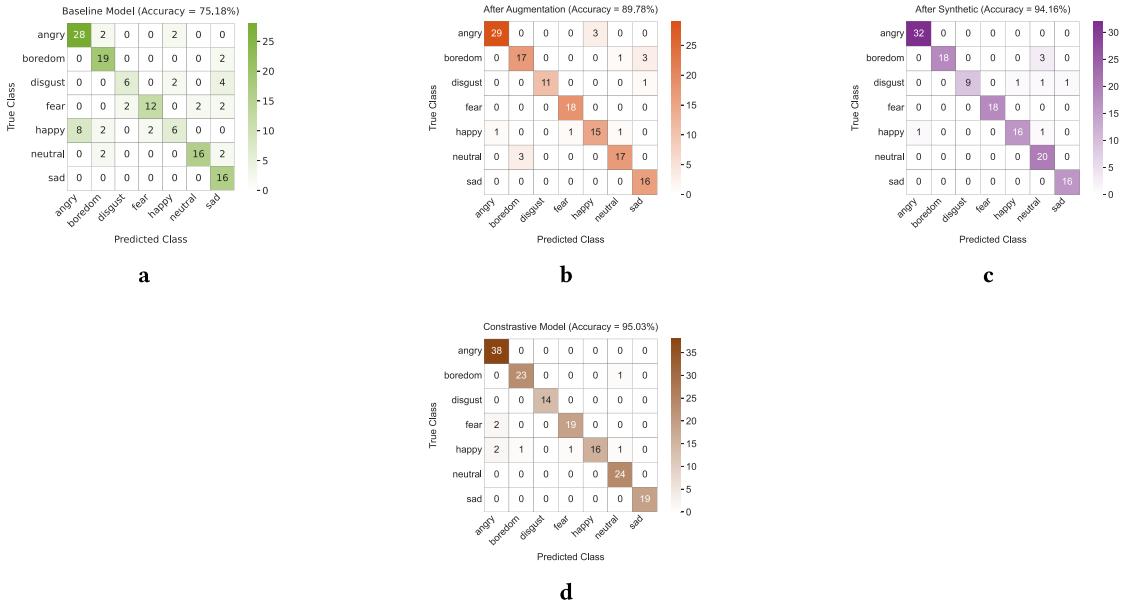


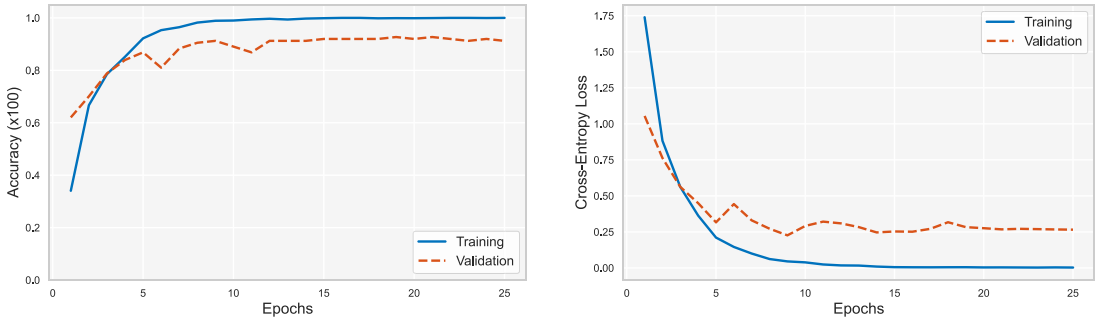
Figure 5: Confusion matrices on the EmoDB dataset: (**a**) baseline model, (**b**) after data augmentation, (**c**) after synthetic data generation, and (**d**) after contrastive learning–based fine-tuning.

Figure 6: Performance visualization on the EmoDB dataset: (**a**) t-SNE feature embedding of the final model, (**b**) Heatmap of precision, recall, F1-score, and accuracy per class.

Feature embeddings visualized in Fig. 6a further validate the effectiveness of the proposed approach. The t-SNE projection reveals well-separated clusters for each emotion category, which confirms the quantitative results. The compact clustering of "Neutral," "Sad," and "Disgust" aligns with their high classification accuracy, while the slight proximity between the "Happy" and "Angry" clusters explains the minor recall drop for the happy category.

The training and validation performance were monitored over 80 epochs to ensure convergence and stability. As illustrated in the Fig. 7a, the model demonstrated rapid learning, with the training accuracy reaching near-perfect levels early in the process. The validation accuracy followed a similar trajectory, stabilizing at approximately 91.24%, which indicates that the combination of Optuna-optimized augmentation and TimeGAN synthetic data effectively prevented the model from memorizing the limited EmoDB samples.

The loss curves corroborate this observation as given in Fig. 7b. The training cross-entropy loss decreased monotonically, reflecting a smooth optimization path. While the validation loss showed slight fluctuations—characteristic of the varied acoustic profiles in emotional speech—it maintained a low and stable range throughout the latter half of the training. This stability was strictly enforced by multi-tiered callback strategy designed to navigate the complex loss landscape of speech features. To refine the optimization in its final stages, a learning rate adaptation strategy was utilized where the learning rate was halved whenever the validation loss stagnated for three consecutive epochs, allowing the optimizer to settle into narrower local minima. Furthermore, to mitigate the deleterious effects of late-stage overfitting, an early stopping mechanism monitored validation accuracy with a patience of six epochs. This ensured that the training was terminated once the model's generalization performance reached its peak. Finally, a model checkpointing system was implemented to save only the weights corresponding to the highest validation accuracy, ensuring that the final model utilized for testing was the most proficient version encountered during the entire training window. The resulting curves demonstrate a well-regularized training process where the final gap between training and validation metrics remains narrow, confirming the model's high reliability for Speech Emotion Recognition on the EmoDB corpus.



a **b**

Figure 7: Performance of the contrastive model on the EmoDB: (**a**) Training and validation accuracy curves, and (**b**) Training and validation cross-entropy loss curves.

4.2 Performance Evaluation Using RAVDESS

This subsection presents a detailed performance analysis of the proposed framework on the RAVDESS dataset. The baseline model, trained on the raw audio data, achieved an accuracy of 71.47%. A closer inspection of the confusion matrix in Fig. 8a reveals significant difficulties with specific emotional categories. Notably, the model frequently confused "Disgust" with "Angry" (19 misclassified instances) and showed poor recognition for "Neutral" states, correctly identifying only 18 samples. This confusion likely stems from the acoustic overlap between high-energy negative emotions and the subtlety of neutral speech, which often lacks distinct spectral peaks. Incorporating Optuna-optimized data augmentation raised the accuracy to 85.33%. The confusion matrix in Fig. 8b shows a marked reduction in the "Disgust-Angry" confusion, with correct predictions for "Disgust" rising to 45. This suggests that the augmented samples enabled the network to learn more robust decision boundaries for these acoustically similar classes. The synthetic data configuration maintained a comparable accuracy level while improving the internal distribution of specific classes. As seen in Fig. 8c, the recognition of "Angry" samples notably improved to 44 correct instances compared to the augmented model, indicating that the synthetic generation process effectively balanced the training distribution for more aggressive emotional tones. The final model, integrating contrastive learning with the augmented and synthetic pipelines, delivered the strongest performance at an accuracy of 96.99%. The

confusion matrix in Fig. 8d, displays a clean diagonal structure with minimal off-diagonal noise. Crucially, the previously problematic "Neutral" class achieved 28 correct classifications, while "Disgust" reached 56 correct predictions, effectively resolving the earlier confusion with "Anger".

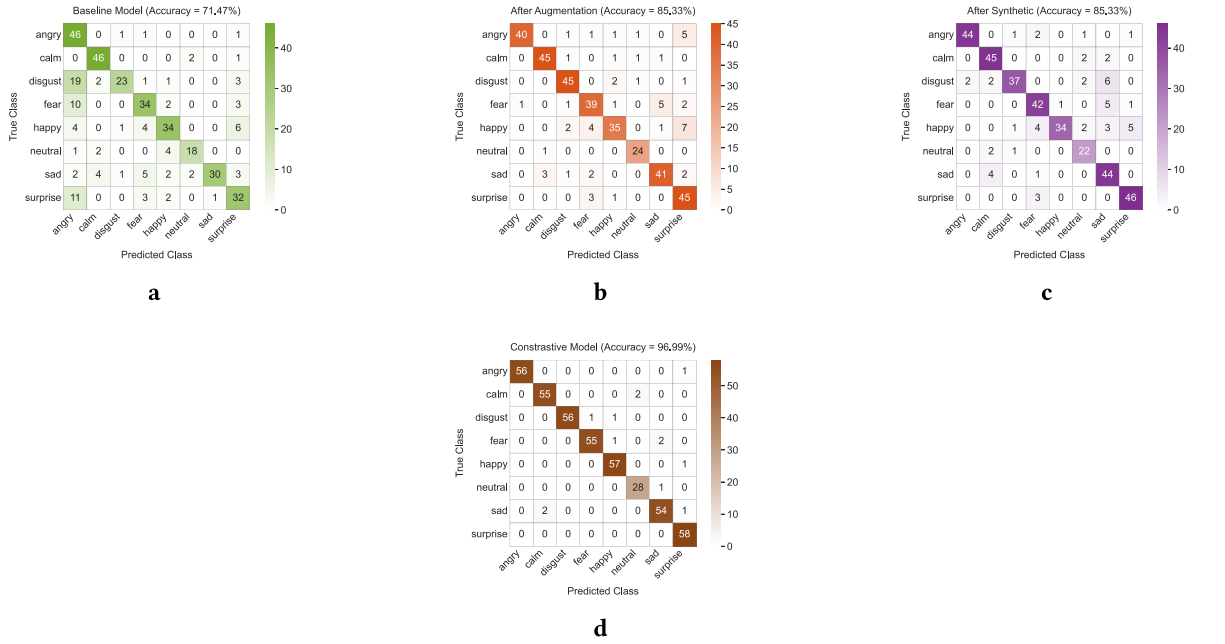


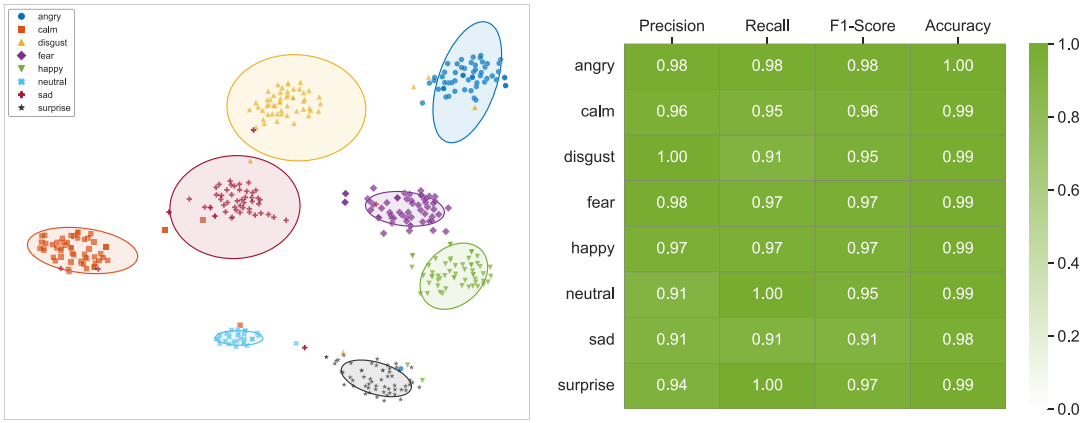
Figure 8: Confusion matrices on the RAVDESS dataset: (**a**) baseline model, (**b**) after data augmentation, (**c**) after synthetic data generation, and (**d**) after contrastive learning–based fine-tuning.

Fig. 9b breaks down the final model's performance by class. The system demonstrated exceptional sensitivity for "Neutral" and "Surprise" emotions, achieving a recall of 1.00 for both. While "Disgust" achieved a perfect precision of 1.00, its recall was slightly lower at 0.91, suggesting the model is highly precise when predicting this emotion but may occasionally miss subtle instances. Overall, F1-scores across all classes remained above 0.91, confirming reliable detection across the emotional spectrum. The t-SNE visualization in Fig. 9a supports these findings, illustrating distinct, compact clusters for each emotion. The clear separation between the "Angry" (blue) and "Disgust" (yellow) clusters is particularly informative, as it visually confirms the model's ability to disentangle these previously confused categories.

The learning behavior of the proposed SER framework was analyzed via the training and validation trajectories, as illustrated in Fig. 10. The accuracy curves (Fig. 10a) demonstrate a rapid learning phase, with the model ascending from an initial 30% to over 80% within the first five epochs. This high velocity of convergence is attributed to the contrastive pre-training phase, which initializes the network with a robust understanding of emotional feature distributions. The training accuracy stabilizes at approximately 98%, while the validation accuracy converges to a peak of 96.6%. The marginal gap between these two curves is particularly noteworthy, as it signifies that the TimeGAN-mediated data expansion effectively regularized the model, preventing the overfitting typically observed in deep architectures trained on limited speech corpora like RAVDESS.

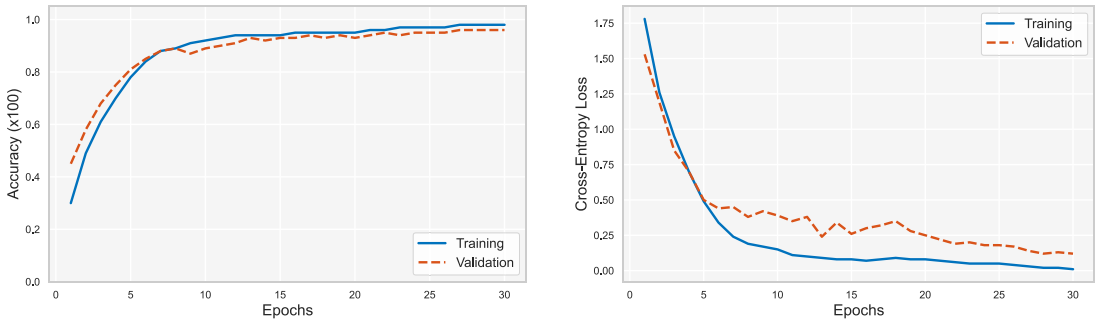
The cross-entropy loss curves Fig. 10b mirror the accuracy trends, exhibiting a sharp logarithmic decay during the initial optimization stages. After epoch 15, the validation loss plateaus at approximately 0.12 with minimal fluctuations. In many SER benchmarks, validation loss often exhibits high volatility due to the high inter-speaker variability in emotional expression; however, the smoothness of the validation trajectory

suggests that the learned contrastive embeddings are invariant to speaker-specific nuances. The absence of any divergent trend in the validation loss confirms that the training reached a stable local minimum, capturing a model with high predictive confidence and strong generalization capabilities for real-world emotional speech signals.



a b

Figure 9: Performance visualization on the RAVDESS dataset: (**a**) t-SNE feature embedding of the final model, (**b**) Heatmap of precision, recall, F1-score, and accuracy per class.



a b

Figure 10: Performance of the contrastive model on the RAVDESS: (**a**) Training and validation accuracy curves, and (**b**) Training and validation cross-entropy loss curves.

4.3 Comparative Analysis and Discussion

The proposed contrastive SER framework was rigorously benchmarked against state-of-the-art methods on EMO-DB and RAVDESS. As shown in Table 2, the proposed model consistently outperformed existing benchmarks and achieved a superior average accuracy of 95.03% on EMO-DB and 96.99% on RAVDESS. A detailed examination of the EMO-DB metrics reveals that the proposed framework achieved exceptional results in traditionally difficult categories. Notably, the proposed model attained perfect recall (1.00) for the 'Sad' and 'Disgust' classes, which represents a significant advancement over the results reported in [1] and [29]. The high accuracy in the 'Boredom' (0.96) and 'Fear' (0.95) emotions suggests that the integration of Optuna-optimized data augmentation effectively mitigated the inherent class imbalances within the EMO-DB corpus. While some existing methods exhibited lower precision under complex spectral overlap

conditions, the hierarchical feature structure preserved by the proposed model enabled clearer disentanglement. On the RAVDESS dataset, the model demonstrated robust generalization across a wider variety of emotional intensities. As shown in the classification results, the framework achieved 100% accuracy for the 'Neutral' and 'Surprise' emotions. This surpasses the 95.0% accuracy for anger reported by Mustaqeem et al. [30]. The consistency between high-arousal emotions (Anger: 0.98, Fear: 0.97) and lower-arousal emotions (Calm: 0.95, Sad: 0.91) highlights the advantage of the contrastive learning stage. By utilizing TimeGAN-generated synthetic samples during training, the model developed a more resilient feature representation that is less sensitive to individual actor variations. The consistently improved performance confirms the superiority of the proposed framework over traditional supervised CNN-based architectures. The core of this advancement lies in the integration of TimeGAN-based data synthesis with a contrastive learning objective. Traditional models often suffer from information bottlenecks and overfitting when trained on small, imbalanced speech corpora. Through TimeGAN, this study generated high-fidelity synthetic temporal sequences that preserved the underlying distribution of emotional speech, effectively expanding the training manifold. Furthermore, contrastive learning enabled the model to learn a highly discriminative latent space. Unlike standard cross-entropy loss, which only focuses on class labels, the contrastive approach compels the model to capture the hierarchical structural relationships and fine-grained spectral nuances within the audio signal. This ensures that the learned embeddings are robust to actor-specific variations and environmental noise. To further assess the practical deployment feasibility of the proposed framework, the computational cost of the final CNN-BiLSTM model was analyzed. The model comprised 594,695 trainable parameters, which indicates a relatively lightweight architecture despite integrated convolutional and bidirectional recurrent layers. The complete training process required approximately 3.24 min, while the average inference latency was measured as 90.38 ms per speech sample. These results suggest that the proposed model is computationally efficient and suitable for real-time SER applications.

Table 2: Comparison of the proposed framework with baseline methods for SER.

Study	Dataset	Accuracy (%) of all emotions									
		Neut.	Calm	Happ.	Sad	Angr.	Fear	Disg.	Surp.	Bored	Avg
[1]	RAVDESS	0.68	0.90	0.65	0.66	0.80	0.74	0.71	0.67	×	0.73
	EMO-DB	0.90	×	0.92	0.93	0.92	0.92	0.99	×	0.88	0.90
	SAVEE	0.82	×	0.47	0.58	0.90	0.50	0.48	0.53	×	0.67
[29]	RAVDESS	0.75	0.57	0.68	0.52	0.92	0.76	0.72	0.80	×	0.71
	EMO-DB	1.00	×	1.00	0.87	1.00	0.67	0.67	×	0.61	0.86
[30]	RAVDESS	0.85	0.95	0.43	0.61	0.95	0.91	0.86	0.95	×	0.77
	EMO-DB	0.85	×	0.66	0.88	0.91	0.92	0.87	×	0.90	0.85
This	EMO-DB	0.92	×	0.76	1.00	0.90	0.95	1.00	×	0.96	95.03
	RAVDESS	1.00	0.95	0.97	0.91	0.98	0.97	0.91	1.00	×	96.99

To assess the quality of generated samples, t-SNE dimensionality reduction was applied to visualize the distribution of real vs. synthetic log-Mel features for the Fear class, as shown in Fig. 11. The visualization reveals a high degree of distributional overlap, which confirms that the TimeGAN successfully captured the underlying acoustic manifold. The generator synthesized samples that occupy the same latent space as the authentic recordings, rather than producing stochastic noise. This qualitative evidence validates the use of TimeGAN to address the class imbalance in SER.

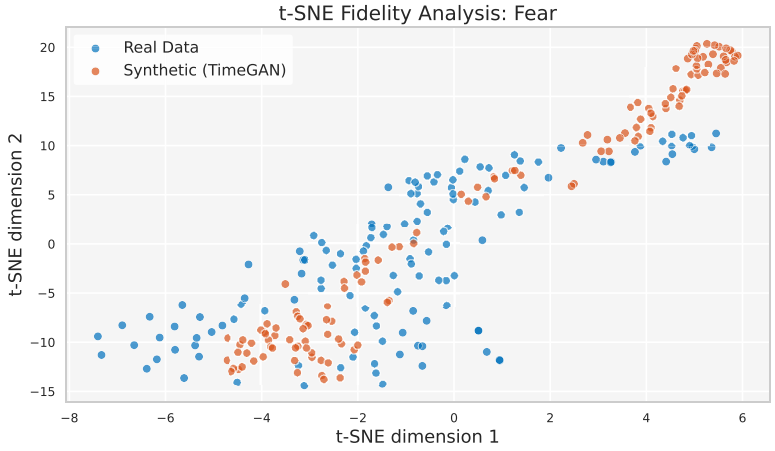


Figure 11: t-SNE visualization of real vs. synthetic log-mel features.

Although the proposed model achieved strong performance on both EMO-DB and RAVDESS, cross-corpus transfer remains an important challenge due to domain shift caused by differences in recording environments, speaker demographics, language, and microphone characteristics. The contrastive pre-training stage is expected to improve robustness under such shifts by learning embeddings that emphasize emotion-consistent acoustic patterns while suppressing variations such as speaker identity and background conditions. Nevertheless, domain mismatch may still alter spectral distributions and temporal prosody cues. Therefore, future work will investigate explicit cross-corpus evaluation and domain adaptation strategies to further validate the transferability of the learned emotional representations.

5 Conclusion

This study introduced a novel SER framework integrating contrastive learning, Optuna-driven data augmentation optimization, and TimeGAN-based synthetic data generation. The proposed framework addressed the key challenges of data scarcity, class imbalance, and the inability to discriminate acoustically similar emotions. By automatically identifying optimal augmentation parameters through Optuna, the model demonstrated strong generalization across diverse vocal characteristics, outperforming existing models that rely on manually tuned augmentation. The effectiveness of this hybrid strategy was confirmed in the experimental results, which obtained high classification accuracy on benchmark data: 95.03% on EmoDB and 96.99% on RAVDESS. Synthetic data generation was specifically targeted at underrepresented and confusing classes, while contrastive learning significantly refined the feature embeddings. This combination proved critical in resolving spectral overlaps between high-arousal emotions, effectively differentiating between Angry and Disgust—a key source of error in baseline models. Despite these advances, there remain constraints in the handling of the full range of prosodic resources and fine-grained changes of spontaneous natural speech, in the generation of synthetic speech features. Future research will focus on improving the fidelity of synthetic data through more sophisticated generative models and evaluating the framework in cross-corpus settings to ensure broader generalizability. Furthermore, to enable real-time deployment in mobile health and human–computer interaction applications, future work will explore model compression strategies—including pruning and quantization—to reduce computational overhead without compromising recognition accuracy. Overall, the proposed framework demonstrates strong potential for

advancing reliable emotion recognition in automated customer service, mental health monitoring, and smart interactive environments.

Acknowledgement: Not applicable.

Funding Statement: This work is supported by Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia through the Researchers Supporting Project PNURSP2026R760.

Author Contributions: The authors confirm contribution to the paper as follows: formal analysis, investigation, methodology, and writing—original draft preparation, Rashid Jahangir; data curation, resources, software, Muhammad Asif Nauman; visualization, methodology, writing—review and editing, funding acquisition: Oumaima Saidani; conceptualization, visualization, Faisal Ramzan. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Data openly available in a public repository.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Farooq M, Hussain F, Baloch NK, Raja FR, Yu H, Zikria YB. Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network. Sensors. 2020;20(21):6008. doi:10.3390/s20216008.
2. Schuller B, Rigoll G, Lang M. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing; 2004 May 17–21; Montreal, QC, Canada.
3. Alluhaidan AS, Saidani O, Jahangir R, Nauman MA, Neffati OS. Speech emotion recognition through hybrid features and convolutional neural network. Appl Sci. 2023;13(8):4750. doi:10.3390/app13084750.
4. Bargal SA, Barsoum E, Ferrer CC, Zhang C. Emotion recognition in the wild from videos using images. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction; 2016 Nov 12–16; Tokyo, Japan. p. 433–6.
5. Li S, Wang J, Tian L, Wang J, Huang Y. A fine-grained human facial key feature extraction and fusion method for emotion recognition. Sci Rep. 2025;15(1):6153. doi:10.1038/s41598-025-90440-2.
6. Mehta D, Siddiqui MFH, Javaid AY. Facial emotion recognition: a survey and real-world user experiences in mixed reality. Sensors. 2018;18(2):416.
7. Labied M, Belangour A, Banane M, Erraissi A. An overview of automatic speech recognition preprocessing techniques. In: Proceedings of the 2022 International Conference on Decision Aid Sciences and Applications (DASA); 2022 Mar 23–25; Chiangrai, Thailand. p. 804–9.
8. Al-Khazraji MJAD, Ebrahimi-Moghadam A. An innovative method for speech signal emotion recognition based on spectral features using GMM and HMM techniques. Wireless Personal Commun. 2024;134(2):735–53. doi:10.1007/s11277-024-10918-6.
9. Kang X. Speech emotion recognition algorithm of intelligent robot based on ACO-SVM. Int J Cognit Comput Eng. 2025;6:131–42. doi:10.1016/j.ijcce.2024.11.008.
10. Santos L, de Araújo Moreira N, Sampaio R, Lima R, Oliveira FCMB. Automatic speech recognition: comparisons between convolutional neural networks, hidden markov model and hybrid architecture. Expert Syst. 2025;42(5):e70032.
11. Jahangir R, Teh YW, Hanif F, Mujtaba G. Deep learning approaches for speech emotion recognition: state of the art and research challenges. Multimed Tools Appl. 2021;80(16):23745–812. doi:10.1007/s11042-020-09874-7.
12. Wang C, Wen G, Yang P, Liu L. Bimodal speech emotion recognition via contrastive self-alignment learning. Expert Syst Appl. 2025;293:128605. doi:10.1016/j.eswa.2025.128605.

13. Latif S, Rana R, Khalifa S, Jurdak R, Qadir J, Schuller B. Survey of deep representation learning for speech emotion recognition. IEEE Trans Affect Comput. 2021;14(2):1634–54. doi:10.1109/taffc.2021.3114365.
14. Pan L. The importance of deep learning models in speech signal processing: fundamentals, strategies, and future research directions. Int J Speech Technol. 2025;28:443–59. doi:10.1007/s10772-025-10194-0.
15. Hu J, Qu L, Li H, Li T. Label semantic-driven contrastive learning for speech emotion recognition. In: Proceedings of the Interspeech 2025; 2025 Aug 17–21; Rotterdam, The Netherlands. [cited 2026 Apr 14]. Available from: https://api.semanticscholar.org/CorpusID:282332217.
16. Jahangir R, Teh YW, Mujtaba G, Alroobaea R, Shaikh ZH, Ali I. Convolutional neural network-based cross-corpus speech emotion recognition with data augmentation and features fusion. Mach Vision Appl. 2022;33(3):41. doi:10.1007/s00138-022-01294-x.
17. Akhtar MZ, Jahangir R, Ain Q, Nauman MA, Uddin M, Ullah SS. UrduSER: a comprehensive dataset for speech emotion recognition in Urdu language. Data Brief. 2025;60:111627.
18. Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, et al. IEMOCAP: interactive emotional dyadic motion capture database. Lang Resour Eval. 2008;42(4):335–59.
19. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B, et al. A database of German emotional speech. In: Proceedings of the Interspeech 2005; 2005 Sep 4–8; Lisbon, Portugal. p. 1517–20.
20. Livingstone SR, Russo FA. 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. 2018;13(5):e0196391.
21. Al-onazi BB, Nauman MA, Jahangir R, Malik MM, Alkhammash EH, Elshewey AM. Transformer-based multilingual speech emotion recognition using data augmentation and feature fusion. Appl Sci. 2022;12(18):9188. doi:10.3390/app12189188.
22. Ahn CS, Rana R, Sivadas S, Busso C, Rajapakse JC. Improving speech emotion recognition with mutual information regularized generative model. arXiv:2510.10078. 2025.
23. Chatziagapi A, Paraskevopoulos G, Sgouropoulos D, Pantazopoulos G, Nikandrou M, Giannakopoulos T, et al. Data augmentation using GANs for speech emotion recognition. In: Proceedings of Interspeech 2005; 2005 Sep 4–8; Lisbon, Portugal. p. 171–5.
24. Shilandari A, Marvi H, Khosravi H, Wang W. Speech emotion recognition using data augmentation method by cycle-generative adversarial networks. SIViP. 2022;16(7):1955–62. doi:10.1007/s11760-022-02156-9.
25. Liu H, HaoChen JZ, Gaidon A, Ma T. Self-supervised learning is more robust to dataset imbalance. arXiv:2110.05025. 2021.
26. Trigeorgis G, Ringeval F, Brueckner R, Marchi E, Nicolaou MA, Schuller B, et al. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In: Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2016 Mar 20–25; Shanghai, China. p. 5200–4.
27. Latif S, Rana R, Khalifa S, Jurdak R, Qadir J, Schuller BW. Deep representation learning in speech processing: challenges, recent advances, and future trends. arXiv:2001.00378. 2020.
28. Kang H, Xu Y, Jin G, Wang J, Miao B. FCAN: speech emotion recognition network based on focused contrastive learning. Biomed Signal Process Control. 2024;96:106545.
29. Issa D, Demirci MF, Yazici A. Speech emotion recognition with deep convolutional neural networks. Biomed Signal Process Control. 2020;59:101894. doi:10.1016/j.bspc.2020.101894.
30. Mustaqeem, Sajjad M, Kwon S. Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. IEEE Access. 2020;8:79861–75. doi:10.1109/access.2020.2990405.