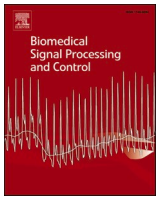
Biomedical Signal Processing and Control 112 (2026) 108782

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



Pre-attentive speech signal processing with adaptive routing for emotion recognition

Huiyun Zhang [a], Zilong Pang [a,*], Puyang Zhao [b], Gaigai Tang [a], Lingfeng Shen [a], Guanghui Wang [a]

[a] School of Software, Henan University, Kaifeng 475004, China
[b] Department of Biostatistics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

ARTICLE INFO

Keywords:
Speech emotion recognition
Capsule network
Feature extraction
Deep learning
Reproducibility

ABSTRACT

Emotion recognition from speech is essential for various applications in human–computer interaction, customer service, healthcare, and entertainment. However, developing robust and reproducible Speech emotion recognition (SER) systems is challenging due to the complexity of emotions and variability in speech signals. In this paper, we first define the concept of reproducibility in the context of deep learning models. We then introduce SpeechNet, a novel deep learning model designed to enhance reproducibility and robustness in SER. SpeechNet integrates multiple advanced components: speech recall, speech attention, and speech signal refinement modules to effectively capture temporal dependencies and emotional cues in speech signal. Additionally, it incorporates a pre-attention mechanism and a modified routing technique to improve feature emphasis and processing efficiency. We also explore effective acoustic feature fusion technique. Extensive experiments on several benchmark datasets demonstrate that the SpeechNet model achieves better performance and reproducibility compared to existing models. By addressing reproducibility and robustness, SpeechNet sets a new standard in SER, facilitating reliable and practical applications.

1. Introduction

Emotion is a fundamental aspect of human experience, deeply influencing our psychological activities and behaviors. It plays a crucial role in communication, decision-making, and social interactions [1]. Affective computing, an interdisciplinary field introduced by Picard, aims to develop systems and devices capable of recognizing, interpreting, and simulating human emotions [2]. This field combines insights from computer science, psychology, and cognitive science, and has seen significant advancements [3].

Speech emotion recognition (SER) involves using computational techniques to analyze and identify the emotional state of a speaker based on vocal expressions [4]. For instance, in customer service, an SER system can detect customer dissatisfaction and trigger appropriate responses, enhancing customer experience [5]. In healthcare, SER can assist in monitoring patients' emotional well-being, providing valuable insights for mental health professionals [6].

Researchers have explored several approaches to emotion recognition across different modalities. Visual methods analyze facial expressions and body language to interpret emotional states. Touch-based recognition uses tactile sensors to detect emotions through physical interactions—beneficial for wearable technology but constrained by its requirement for direct contact [7]. EEG methods measure brain activity patterns correlated with emotional states, offering precision but limited by their invasive nature and impracticality [8]. SER presents rich emotional content. Unlike visual methods, which may be compromised by lighting conditions or occlusions, or touch-based approaches requiring physical contact, speech-based recognition operates effectively across various environments. This accessibility makes speech recognition particularly suitable for applications in Human-computer interaction (HCI), customer service systems, and mental health monitoring tools.

SER faces significant challenges in developing systems that perform reliably across different contexts. These difficulties include identifying and extracting the most informative acoustic features that correlate with emotional states, and ensuring consistent model performance across diverse datasets, speakers, and environmental conditions. While feature selection has received considerable attention in previous research, the

* Corresponding author.
E-mail addresses: zhzhy@henu.edu.cn (H. Zhang), jszxpzl@henu.edu.cn (Z. Pang), puyang.zhao@uth.tmc.edu (P. Zhao), tgg@henu.edu.cn (G. Tang).

https://doi.org/10.1016/j.bspc.2025.108782
Received 8 April 2025; Received in revised form 27 July 2025; Accepted 1 October 2025
Available online 7 October 2025

equally important question of reproducibility has been comparatively ignored.

Identifying relevant features that capture the emotional content of speech is important for effective SER. Traditional methods often rely on handcrafted features such as Mel-frequency cepstral coefficient (MFCC) [9], pitch [10], and energy [11], which may not fully capture the complexity of emotional expressions in speech. Although advanced deep learning models, including Convolutional neural networks (CNNs) [12] and Recurrent neural networks (RNNs) [13], show promise in addressing this issue, their high complexity and large number of parameters make them prone to overfitting [14], particularly when training data is limited. Additionally, variations in speech due to different speakers, accents, speaking styles, and background noise can significantly affect model performance, underscoring the need for models that generalize well across diverse conditions.

Reproducibility is often overlooked in the SER systems [15], leading to models that perform well on specific datasets but fail to generalize to new data. A major challenge in SER is ensuring that models produce consistent and reliable results across different datasets and experimental setups.

We propose a approach called SpeechNet, a novel deep learning model that combines traditional handcrafted features with advanced techniques to tackle these challenges. Our model employs a 107-dimensional fusion feature designed to represent various aspects of speech signals, including MFCC, zero crossing rate, chroma, and others. SpeechNet integrates several advanced components: a speech recall module that captures temporal dependencies, a multi-head attention mechanism that emphasizes the most informative parts of the speech signal, a convolutional signal refinement module that filters noise, and a modified routing technique that improves processing efficiency.

By addressing the challenges of feature identification and model reproducibility, SpeechNet sets a new standard in SER. Through a combination of handcrafted and deep learning features, coupled with a reproducible architecture, SpeechNet enhances the reliability and practical applicability.

The remainder of this paper is organized as follows: Section 2 reviews related works. Section 3 defines the concept of reproducibility and introduces the SpeechNet architecture. Section 4 outlines the benchmark datasets and feature extraction. Section 5 details experimental results. Section 6 discusses the significance of our research findings. Section 7 concludes the paper and future research directions.

2. Related works

2.1. Traditional approaches to SER

SER has changed from traditional methods to more advanced deep learning techniques. Early approaches in SER primarily relied on handcrafted features such as MFCC, energy, and formant frequencies [16]. These features were often used in combination with classical machine learning algorithms like Support Vector Machine (SVM) [17], Gaussian mixture model (GMM) [18], and Hidden Markov model (HMM) [19]. For example, Lee et al. used HMM to capture emotional states from speech signal, achieving reasonable accuracy on limited datasets [20]. Similarly, Schuller et al. combined prosodic features with SVM for emotion recognition [21]. However, these traditional approaches faced limitations in capturing the complex nature of human emotions due to their handcrafted features and linear models.

2.2. Deep learning in SER

Researchers widely use CNNs to identify spatial patterns in speech, while RNNs are good at capturing how emotional cues in speech unfold over time. Deep learning has allowed researchers to extract features directly from raw audio.

CNNs have been employed to extract high-level features from spectrograms, effectively capturing spatial patterns in the frequency domain. For instance, Trigeorgis et al. introduced a CNN model that functions directly on raw audio waveforms, eliminating the need for handcrafted features and achieving state-of-the-art performance on several SER benchmarks [22].

RNNs are well-suited for capturing temporal dynamics in speech signals. Tao et al. demonstrated the effectiveness of LSTMs in modeling long-term dependencies in emotional speech and improved the recognition accuracy [23]. Moreover, hybrid models combining CNNs and LSTMs have been developed to utilize the advantages of both architectures. Li et al. created a CNN-LSTM model that handles spectrograms and achieves better performance [24].

Attention mechanisms have significantly enhanced the performance of deep learning models in SER by enabling them to dynamically focus on the most relevant parts of the input sequence. This advancement not only improves recognition accuracy but also increases model interpretability. For instance, Mirsamadi et al. effectively combined attention mechanisms with LSTM networks, using learnable weights to assess the contribution of different speech segments, thereby boosting overall performance [25].

In terms of architectural innovation, Sabour et al. pioneered the application of Capsule Networks (CapsNets) to SER [26]. The key strength of CapsNets lies in their ability to preserve hierarchical relationships between features, making the models more robust to variations in input data. Wu et al. experimentally demonstrated that this architecture effectively captures the layered structure of emotional cues in speech, leading to notable improvements in recognition accuracy [27].

2.3. Reproducibility and robustness in SER

While deep learning has significantly advanced SER performance, reproducibility and reliability remain key challenges in this field. Some models achieve high accuracy on specific datasets but struggle with generalization across different contexts, highlighting the need for more rigorous evaluation standards. Researchers are developing better assessment frameworks. Hashem et al. established reproducibility guidelines for SER model evaluation [28]. Grósz et al. created a benchmarking system for fairer model comparisons, improving research standards [29].

Motivated by these advancements, we introduce SpeechNet in our work to address the challenges of reproducibility and robustness in SER. SpeechNet combines LSTM, attention mechanism, and CNN to capture both temporal dependencies and spatial patterns in emotional speech. We introduced a pre-attention mechanism to improve feature salience and a modified routing technique inspired by CapsNet to improve the efficiency of feature processing. Through experiments on six SER datasets, we show that SpeechNet achieves better performance and reproducibility compared to existing models.

2.4. Comparison of multimodal emotion recognition methods

Emotion recognition is not limited to speech and various modalities such as visual, touch, and EEG-based methods have been thoroughly investigated. Visual emotion recognition examines facial expressions and body language to determine emotional states [30]. While effective in controlled environments, it can be vulnerable to lighting conditions and occlusions. Touch-based emotion recognition employs tactile sensors to identify emotional cues from physical interactions, which can be beneficial in wearable devices but is constrained by the need for physical contact [31]. EEG-based methods monitor brain activity patterns associated with emotions, offering high precision but requiring invasive sensors and being less practical for applications [8].

SER offers several advantages. It is non-intrusive and can be easily integrated into existing communication systems. Speech-based methods can capture a wide range of emotional cues through prosodic features,

and advances in deep learning have significantly improved their accuracy and robustness. Moreover, speech is a natural form of communication, making SER highly applicable in real-world scenarios such as HCI, customer service, and mental health monitoring.

3. The reproducibility of SpeechNet

The essential goal of SER is to achieve more efficient and reproducible predictions. While the former has received sufficient attention in past research, the latter has been overlooked in almost all research work.

3.1. Definition of reproducibility

Reproducibility in the context of deep learning models refers to the model's ability to consistently produce the same outputs when given the same inputs across multiple executions, essential for validating reliability and stability by ensuring predictable and dependable behavior. For a comprehensive assessment, reproducibility is divided into two primary dimensions: general learning reproducibility and correct learning reproducibility, each further subdivided to capture more specific aspects of reproducibility.

General learning reproducibility (GLR) measures how consistently a model performs across multiple runs under identical conditions, using the same training and test datasets. It calculates what percentage of samples receive consistent predictions in at least half of the test rounds, helping researchers understand model stability. It is defined as:

$$GLR = \frac{1}{N}\sum_{i=1}^{N} I\left(C_i \geq \frac{R}{2}\right) \tag{1}$$

where N is the total number of samples, and $I(\bullet)$ is indicator function, returning 1 if the condition is true and 0 otherwise.

Fractional learning reproducibility (FLR) measures cases and produces consistent results in more than half test runs. Thus, we identify models that show general stability with occasional variability and calculate FLR as:

$$FLR = \frac{1}{N}\sum_{i=1}^{N} I\left(\frac{R}{2} \leq C_i \leq R\right) \tag{2}$$

Perfect learning reproducibility (PLR) measures when a model produces identical results across all test runs, indicating consistent performance under repeated testing, and it is defined as:

$$PLR = \frac{1}{N}\sum_{i=1}^{N} I(C_i = R) \tag{3}$$

Correct learning reproducibility (CLR) measures whether a model produces results that are both consistent and accurate (matching true labels) in at least half of test rounds. This is calculated by counting when predictions are both consistent and correct, using the formula:

$$CLR = \frac{1}{N}\sum_{i=1}^{N} I\left(C_i \geq \frac{R}{2}\right) \bullet I\left(\widehat{L}_{i,r} = L_i\right) \tag{4}$$

Fractional correct learning reproducibility (FCLR) indicates that the model produces consistent and correct results in more than half but not all the test runs. It reflects that the model generally makes correct predictions consistently, but there is occasional inconsistency, which is defined as:

$$FCLR = \frac{1}{N}\sum_{i=1}^{N} I\left(\frac{R}{2} \leq C_i \leq R\right) \bullet I\left(\widehat{L}_{i,r} = L_i\right) \tag{5}$$

Perfect correct learning reproducibility (PCLR) signifies that the model consistently produces identical and correct results in all test runs. Achieving this level indicates that the model is not only stable but also highly accurate in its predictions. It is defined as:

$$PCLR = \frac{1}{N}\sum_{i=1}^{N} I(C_i = R) \bullet I\left(\widehat{L}_{i,r} = L_i\right) \tag{6}$$

3.2. Reproducible SpeechNet

The architecture of SpeechNet integrates multiple neural network components to handle the complexity of SER, as shown in Fig. 1. SpeechNet consists of five key modules: Information encoding module, Speech recall module, Speech attention module, Speech signal refinement module, and Speech routing module. Each module is important in capturing the intricate details of speech emotion while ensuring consistent performance across different datasets.

The information encoding module processes the input audio signals by mapping them to higher-level representations. It uses an inverted pyramid topology with two fully connected layers with $n_1 = 256$ and $n_2 = 128$ neurons, respectively. This design robustly captures high-level features while retaining relevant low-level features. The output $h_c \in \mathbb{R}^{n_2}$ is calculated:

$$h_c = f(W_2(f(W_1 x + b_1)) + b_2) \tag{7}$$

where $x \in \mathbb{R}^m$ is the input feature vector, $W_1 \in \mathbb{R}^{n_1 \times m}$ and $W_2 \in \mathbb{R}^{n_2 \times n_1}$ are weight matrices, $b_1 \in \mathbb{R}^{n_1}$ and $b_2 \in \mathbb{R}^{n_2}$ are bias vectors, and $f(\bullet)$ is the LeakyReLU.

The speech recall module uses an LSTM layer to process sequential data and identify long-term patterns in speech signals that indicate different emotions. With MFCCs, pitch and ZCR, the module creates a 107-dimensional feature representation. Unlike many traditional SER models, this component recognizes temporal dependencies, allowing it to track changing emotional states in longer speech samples.

The LSTM, with $n_3 = 64$ cells, produces hidden states h_t for time interval $\Delta T_3 = [t_1, t_2, ..., t_{n_3}]$: $h_t = LSTM(h_{t-1}, x_t) = W_3 h_{t-1} + b_3$. Collecting the outputs over all time points, the module output h_r is defined as:

$$h_r = \left[h^r_{t_1}, h^r_{t_2}, ..., h^r_{t_{n_3}}\right]^T, h_r \in \mathbb{R}^{n_3} \tag{8}$$

LSTM contributes significantly to the decision-making process by providing a deeper understanding of the temporal progression of emotions. This ability allows SpeechNet to predict emotions more accurately, especially in nuanced and dynamically changing emotional states.

Speech attention module applies an attention mechanism to focus on the most informative parts of the speech signal. This module consists of three parallel auxiliary layers (Pre-attention) and Multi-head mechanism (MHA). It uses an MHA to enhance the model's learning capability.

Prior to deep processing, this mechanism highlights key emotional cues in the input data. Unlike traditional attention mechanisms used in models, our pre-attention setup prepares the data by emphasizing significant features before they are processed by subsequent layers, improving the efficiency and focus of the model. Pre-attention includes the pre-query $q = W_q h_r$, pre-key $k = W_k h_r$, and pre-value $v = W_v h_r$, where $W_q, W_k, W_v \in \mathbb{R}^{n_q \times n_3}$. For each attention head i: $q_i = W^i_q q$, $k_i = W^i_k k$, $v_i = W^i_v v$. The attention output for head i is $head_i = softmax\left(\frac{q_i k_i^T}{\sqrt{d_k}}\right) v_i$. The MHA output $h_a \in \mathbb{R}^{n_3 \times (d_h \bullet n_h)}$ is defined as:

$$h_a = \left[head_1, head_2, ..., head_{n_h}\right] \tag{9}$$

Pre-attention allows the model to dynamically focus on important parts of the input sequence, which is crucial for capturing emotional nuances that may be scattered throughout the speech signal. MHA provides multiple perspectives on the data, improving the model's ability to learn complex patterns.

The main difference between our pre-attention and standard attention models is the emphasis on the first features. While traditional attention mechanisms usually process all input features equally before applying the attentional weights, our pre-attention setup first emphasizes the emotionally salient features through parallel auxiliary layers.

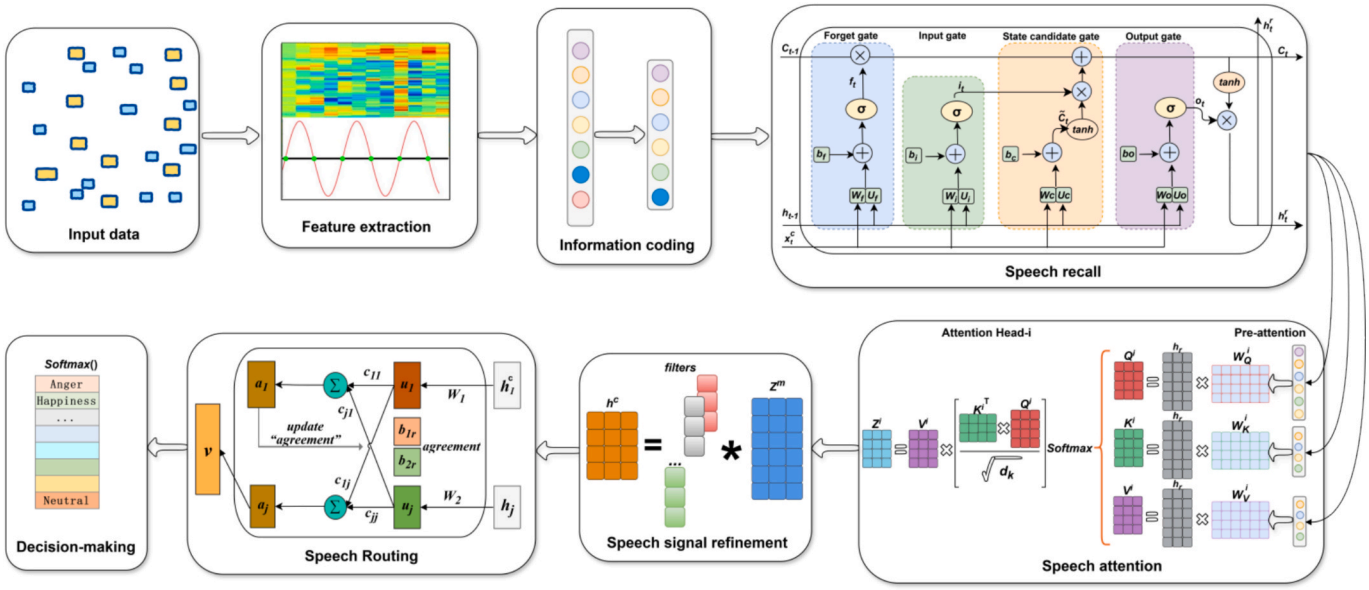


Fig. 1. Flowchart of SpeechNet, a reproducible deep learning model for Speech Emotion Recognition (SER). The architecture comprises five key modules: the information coding module, speech recall module, speech attention module, speech signal refinement module, and speech routing module, each designed to fine-tune and optimize feature extraction based on the specific characteristics of input speech data.

This early highlighting helps the model to focus on emotionally relevant patterns from the beginning and reduce the computational load of processing less relevant features.

Speech signal refinement module uses convolutional layers to filter noise and enhance emotional cues in the speech features. It refines the data stream to produce cleaned and informative features. The ReLU ensures that the model retains non-linearity, which is essential for capturing complex emotional patterns. The output h_s for the j^{th} convolutional kernel is calculated as:

$$h_s^j = g\left(w_s^j * h_a\right) \tag{10}$$

where * denotes the convolution operation, $w_s^j \in R^{k\times 1}$ is the kernel, and $g(\bullet)$ is the ReLU. The module output $h_s \in \mathbb{R}^{(n_3-k+1)\times n_4}$ is expressed as:

$$h_s = \left[h_s^1, h_s^2, \ldots, h_s^{n_4}\right] \tag{11}$$

The speech routing module directs features to their corresponding emotion classes by using a modified dynamic routing mechanism adapted from CapsNet. Inspired by CapsNet, it preserves hierarchical relationships between speech features, enabling robust processing of variations in speech inputs such as accents and intonations. Compared to traditional CNN or RNN architectures that lack dynamic routing capabilities, our approach demonstrates better generalization across diverse datasets. Modified dynamic routing firstly partitions h_s as $\tau = \left\lceil \frac{(n_3-k+1)\times n_4}{d} \right\rceil$ blocks: $u_1, u_2, \ldots, u_\tau$, where each block has a size of $d = 16$. Each block is viewed as a set of neurons, i.e., a capsule. It then calculates the importance score a_{ij} between blocks i and j is: $a_{ij} = W_{ij}h_i, a_{ij} \in \mathbb{R}^{d\times d}$. The similarity u_{ij} is calculated as:

$$u_{ij} = a_{ij} \bullet h_j \tag{12}$$

Normalized similarities c_{ij} are expressed as follows:

$$c_{ij} = \frac{exp\left(u_{ij}\right)}{\sum_{k=1}^{d} exp\left(u_{ik}\right)} \tag{13}$$

The routing mechanism updates c_{ij} iteratively, and the output v for emotion classification is calculated as follows:

$$v = \sum_{i=1}^{n_4} c_{ij} a_{ij} \tag{14}$$

Dynamic routing allows the model to adjust the importance of different features, which is crucial for accurately classifying emotions that may have subtle and overlapping features. Our modified routing technique differs from traditional approaches in three key aspects. We first employ adaptive block partitioning to better preserve emotional feature hierarchies. Then, considering the need for more refined feature relationships, we used iterative similarity updating that enables these more refined feature relationships. To better handle the variable nature of emotional expressions, we added dynamic weighting. Compared to standard approaches, this modification shows strength in maintaining feature relationships across different emotional states.

3.3. Balanced loss function of SpeechNet

SpeechNet employs a novel balanced margin loss function to mitigate the impact of imbalanced data:

$$L = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\omega_j\left[y_{i,j} \bullet max\left(0, \alpha_1 - \hat{y}_{i,j}\right)^2 + \alpha_2 \bullet \left(1 - y_{i,j}\right) \bullet max\left(0, \hat{y}_{i,j} - \alpha_3\right)^2\right] \bullet \beta \tag{15}$$

where N represents the total number of training samples, $y_{i,j}$ denotes the label of the i^{th} sample, while $\hat{y}_{i,j}$ represents the predicted label of the i^{th} sample. ω_j is the sample weight, and $\alpha_1 = \alpha_2 = \alpha_3 = 0.5$. The balancing factor β is determinate by $\beta = max\left(\frac{1}{C}\sum_{c=1}^{C}\frac{1-y_{j,c}+y_{j,c}\bullet\frac{C-1}{C}}{1-\frac{1}{C}}, 1\right)$. The balanced margin loss addresses class imbalance by weighting the loss contributions of different classes, ensuring that the model does not become biased towards the more frequent classes.

By integrating these components, SpeechNet achieves high performance and reproducibility, effectively handling imbalances and capturing complex emotional patterns in speech data. The combination of LSTM, attention mechanism, convolutional layers, and dynamic routing ensures that SpeechNet can robustly analyze and classify

emotions.

4. Datasets and feature extraction

To evaluate the performance and reproducibility of the proposed SpeechNet model, we conducted experiments on six benchmark datasets to ensure comprehensive and robust assessment. These datasets are EmoDB [32], SAVEE [33], CASIA [34], BodEmoDB [35], IEMOCAP [36], and ESD [37].

4.1. Benchmark datasets

To provide a comprehensive and diverse foundation for evaluating the SpeechNet model in SER tasks, we utilized multiple datasets. Table 1 presents the key characteristics of the benchmark datasets employed for SER evaluation. Among these, the BodEmoDB dataset was meticulously recorded by the Pattern Recognition Laboratory of Qinghai Normal University [35], while the other datasets are widely recognized standards in the field. Each dataset contributes valuable insights into emotional expression across different languages and cultural contexts.

4.2. Fusion features

We used a set of 107-dimensional comprehensive fusion features to represent different aspects of speech signals. These descriptors include basic features such as MFCC, zero-crossing rate, chroma, spectral contrast, spectral centroid, spectral roll-off, and pitch, which are calculated for each frame and designed to capture different aspects of audio data to ensure the model can effectively learn and generalize.

The MFCC is derived from the cepstral representation of the audio signal and captures the power spectrum on a nonlinear mel-scale of frequency [8]. The process involves Fourier Transform the audio signal, mapping the power spectrum to the mel-scale and then logarithmizing and applying the calculated discrete cosine transform:

$$MFCC(n) = \sum_{k=1}^{K} log(S(k)) cos\left(\frac{n(k-0.5)\pi}{K}\right) \tag{16}$$

where $S(k)$ is the mel-scaled power spectrum, K is the number of mel-bands, and n is index of the MFCC coefficient. They capture dynamic changes in the spectral properties over time.

$$\Delta MFCC(t) = MFCC(t+1) - MFCC(t-1) \tag{17}$$

$$\Delta^2 MFCC(t) = \Delta MFCC(t+1) - \Delta MFCC(t-1) \tag{18}$$

ZCR quantifies the frequency at which the audio signal changes its polarity [38], making it a fundamental feature in distinguishing various types of signals, including voiced and unvoiced speech. Despite its computational simplicity, ZCR provides highly valuable information and is calculated as:

$$ZCR = \frac{1}{N}\sum_{n=1}^{N-1} L_{\{x(n)\bullet x(n-1)<0\}} \tag{19}$$

where $x(n)$ is the audio signal, N is the number of samples, and $L_{\{\bullet\}}$ is an indicator function that equals 1 when its condition is true.

Chroma represents the energy distribution across the twelve different pitch classes [39]. These features capture harmonic and tonal content, crucial for music analysis. It is defined as:

$$Chroma(i) = \sum_{k\in \mathrm{K}(i)} STFT(k) \tag{20}$$

where $\mathrm{K}(i)$ is the set of frequencies corresponding to the i^{th} chroma, and $STFT(k)$ is the short-time Fourier transform at frequency k.

Spectral contrast (SC) gauges the amplitude disparity between peaks and valleys within the sound spectrum, offering insights into the timbral texture and harmonic richness of the audio [40]. It is calculated as:

$$SC(i) = max_{f\in B_j} S(f) - min_{f\in B_j} S(f) \tag{21}$$

where $S(f)$ is the magnitude of the spectrum at frequency f, and B_j represents the frequency bins.

Spectral centroid (SCT) is the center of mass of the spectrum [41], indicating where the bulk of the spectral energy is located, which is defined as:

$$SCT = \frac{\sum_f fS(f)}{\sum_f S(f)} \tag{22}$$

where f is the frequency and $S(f)$ is the magnitude of the spectrum at frequency f.

Spectral rolloff (SR) is the frequency which a specified percentage of the total spectral energy lies and benefits in distinguishing harmonic content from noise [42], providing valuable information about the sharpness of the signal. It is defined as:

$$SR = max_f\left(\sum_{i=0}^{f} S(i) \le 0.85\sum_{i=0}^{F} S(i)\right) \tag{23}$$

where $S(i)$ is the spectral magnitude at bin i, and F is the total number of frequency bins.

Spectral bandwidth (SBW) measures the width of the band of frequencies that contain most of the signal energy [43], which is defined as:

$$SBW = \sqrt{\frac{\sum_f (f-SCT)^2 S(f)}{\sum_f S(f)}} \tag{24}$$

where SCT is the spectral centroid and helps in understanding the spread of spectral energy, proving useful for characterizing different types of audio signals, such as narrowband versus broadband sounds.

Tonnetz represents the tonal characteristics and harmonic relation-

Table 1
The related information of the benchmark datasets for speech emotion recognition tasks.

Dataset	Language	Utterances	Speakers	Emotions (#. of samples)	Unique features
EmoDB	German	535	10	Anger/A (127), Boredom/B (81), Fear/F (69), Disgust/D (46), Happiness/H (71), Sadness/Sa (62), and Neutral/N (79)	Diverse emotional expressions from multiple speakers
CASIA	Chinese	1200	4	Anger/A (200), Fear/F (200), Happiness/H (200), Neutral/N (200), Sadness/Sa (200), and Surprise/Su (200)	Comprehensive Mandarin emotional expressions
SAVEE	English	480	4	Anger/A (60), Disgust/D (60), Fear/F (60), Happiness/H (60), Sadness/Sa (60), Surprise/Su (60), and Neutral/N (120)	Balanced emotional states, efficient for model training
BodEmoDB	Tibetan	3000	10	Anger/A (500), Fear/F (500), Happiness/H (500), Sadness/Sa (500), Surprise/Su (500), and Neutral/N (500)	Cultural and linguistic insights, gender-balanced speakers
IEMOCAP	English	4485	10	Anger/A (1103), Happiness/H (590), Sadness/Sa (1084), and Neutral/N (1708)	Multi-modal (audio/video), naturalistic interactions
ESD	Mandarin &English	35,000	20	Anger/A (7000), Happiness/H (7000), Sadness/Sa (7000), Surprise/Su (7000), and Neutral/N (7000)	Cross-linguistic emotional analysis, diverse speaker demographics

ships of the audio signal [44], derived from harmonic pitch class profiles, which is defined as:

$$Tonnetz(t) = HPCP(t) \times M \tag{25}$$

where $HPCP(t)$ is the harmonic pitch class profile at time t, and M is the Tonnetz transformation matrix.

To normalize the data, we employ standardization, which scales the features to have a mean of zero and a standard deviation of one. Standardization is crucial for deep learning models because it ensures each feature contributes equally to the model's learning, preventing features with larger scales from dominating the learning process. The selected features capture a wide range of audio characteristics, including spectral, temporal, harmonic, and energy-related properties. This ensures a holistic representation of the audio signal. Each feature contributes unique information to enhance the model's performance across various SER tasks, making this a valuable strategy in signal processing.

5. Experimental

5.1. Experimental setup

All experiments were conducted on an HP Omen 8 Plus laptop configured with a 12th Gen Intel® Core™ i7-12800HX processor (2.00 GHz), 32 GB of RAM, a 1 TB SSD, and an NVIDIA RTX 3080 Ti GPU with 16 GB of VRAM. This setup provides sufficient computational resources for training and evaluating deep learning models with high efficiency. The system operated on a 64-bit Windows 11 OS, and all models were implemented using TensorFlow, which supports optimized GPU computation and scalable model deployment.

The proposed SpeechNet model was evaluated on six widely used benchmark datasets for SER: EMODB, SAVEE, CASIA, BodEMODB, IEMOCAP, and ESD. Among these, BodEMODB is a newly developed Tibetan speech emotion dataset created by our research group.

The data was randomly partitioned, with 90 % used for training and 10 % reserved for testing. The SpeechNet model was trained for 200 epochs using a batch size of 256, a dropout rate of 0.5, and the Adam optimizer with a learning rate of 0.001. The capsule routing module was configured with two dynamic routing iterations to balance computational efficiency and representational capacity.

5.2. Evaluation metrics

We evaluated the proposed SpeechNet model on six benchmark datasets: EmoDB, SAVEE, CASIA, BodEmoDB, IEMOCAP, and ESD, and we used metrics: Accuracy, Precision, Recall, and F1-score to measure the performance of the model. Additionally, we assess its reproducibility with the metrics: GLR, FLR, PLR, CLR, FCLR, and PCLR.

Table 2
Ablation study of the proposed SpeechNet model on the EMODB dataset, highlighting the impact of each module.

Components	Accuracy	Precision	Recall	F1-score
SpeechNet w/o information coding	80.37	81.37	77.57	79.43
SpeechNet w/o speech recall	83.18	84.62	82.24	83.41
SpeechNet w/o pre-attention	84.11	83.96	83.18	83.96
SpeechNet w/o speech attention	85.05	85.71	84.11	84.91
SpeechNet w/o speech signal refinement	86.92	86.92	86.92	86.92
SpeechNet w/o speech routing	86.92	86.79	85.98	86.38
SpeechNet w/o modified speech routing	85.98	86.67	85.05	85.85
SpeechNet	87.66	88.27	85.05	86.62

5.3. Results and analysis

The ablation study in Table 2 demonstrates the progressive performance improvement of SpeechNet as key modules are incrementally incorporated. The most significant contribution comes from the information coding module, whose removal causes a 7.29 % accuracy drop (from 87.66 % to 80.37 %), highlighting its crucial role in feature representation. Subsequent modules—speech recall, pre-attention, and modified routing—each contribute meaningfully, with F1-score improvements of 3.21, 2.66, and 0.77 points respectively, validating their necessity in the architecture.

The results reveal a clear performance hierarchy, where foundational feature processing (information coding) matters most, followed by memory mechanisms (recall/pre-attention) and finally refinement modules (routing/attention). This structured degradation confirms that all components synergistically enhance model capability, with the full SpeechNet configuration achieving optimal performance (86.62 F1-score). The findings justify our architectural design choices for speech emotion recognition.

Table 3 shows the performance of the proposed SpeechNet model on benchmark datasets. Our experiments were calculated over ten runs to provide average and standard deviation.

The SpeechNet model performs quite consistently on the EmoDB dataset with high precision and accuracy, but slightly lower recall. This indicates that while the model correctly identifies positive instances, it may miss some positive ones. Performance on SAVEE is moderate, with lower accuracy, precision, and recall compared to EmoDB, possibly due to its characteristics or noise. The model performs well on the CASIA dataset with high accuracy and precision, like EmoDB. Performance is also relatively high for CASIA, indicating a balanced performance across all metrics.

BodEmoDB exhibits the highest accuracy and precision among all analyzed datasets. Its consistently low standard deviation across metrics indicates strong stability, making it the most suitable match for the SpeechNet model. IEMOCAP shows the lowest performance, particularly in recall measurements, suggesting the model struggles to identify positive instances within its more complex or noise-heavy data environment. Meanwhile, the SpeechNet model processes the ESD dataset effectively, achieving solid results across all evaluation metrics with minimal standard deviation.

Performance is lower on SAVEE and IEMOCAP datasets. These datasets likely present specific challenges including background noise or greater variation in emotional expressions. The SpeechNet model performs best on ESD, demonstrating the highest and most stable performance. IEMOCAP presents a big challenge for the model, indicating a need for improvements in handling complex or noisier data.

Fig. 2 illustrates SpeechNet's performance across six datasets through violin and box plots, revealing distinct patterns in emotion recognition capabilities. The high-performance datasets (EmoDB, BodEmoDB, and ESD) display narrow, symmetrical violin plots with high median values and compact interquartile ranges, indicating consistent performance across emotion categories. Moderate-variability datasets (SAVEE and CASIA) show wider distributions for specific emotion categories despite maintaining good median performance, highlighting areas for potential improvement. IEMOCAP demonstrates the most significant challenges, with substantially wider plots, varied distributions, and lower median values, reflecting the complexity of its more nuanced emotional expressions.

The performance differences observed across datasets correlate directly with the varying complexity of emotional expressions in each collection. SpeechNet achieves reliable emotion classification with clearly expressed emotions in EmoDB, BodEmoDB, and ESD. Meanwhile, the more subtle and complex emotional nuances in IEMOCAP create substantial classification challenges.

Fig. 3 comprises the training and testing loss curves for the SpeechNet model. For all datasets, the training loss (blue) decreases steadily

Table 3
The performance (%) analysis of the proposed SpeechNet model across different datasets: EmoDB, SAVEE, CASIA, BodEmoDB, IEMOCAP, and ESD.

Datasets	Metrics	10 times										Avg ± Std
		1	2	3	4	5	6	7	8	9	10	
EmoDB	Accuracy	86.92	87.85	85.98	86.92	90.65	85.05	88.79	90.65	85.98	87.85	**87.66 ± 1.81**
	Precision	87.62	87.74	86.14	87.00	91.43	84.91	89.32	92.23	87.38	88.89	88.27 ± 2.15
	Recall	85.98	86.92	81.31	81.31	89.72	84.11	85.98	88.79	84.11	82.24	85.05 ± 2.80
	F1-score	86.79	87.33	83.66	84.06	90.57	84.51	87.62	90.48	85.71	85.44	86.62 ± 2.32
SAVEE	Accuracy	71.88	70.83	73.96	63.54	72.92	76.04	71.88	68.75	70.83	61.46	**70.21 ± 4.30**
	Precision	72.22	72.41	76.09	66.28	74.19	80.00	73.03	69.15	70.65	64.04	71.81 ± 4.39
	Recall	67.71	65.62	72.92	59.38	71.88	70.83	67.71	67.71	67.71	59.38	67.09 ± 4.40
	F1-score	69.89	68.85	74.47	62.64	73.02	75.14	70.27	68.42	69.15	61.62	69.35 ± 4.25
CAISA	Accuracy	88.75	87.08	84.58	87.92	87.92	89.17	85.83	90.00	87.08	89.58	**87.79 ± 1.62**
	Precision	89.03	87.39	85.17	89.27	88.79	90.48	86.02	90.38	86.97	89.96	88.35 ± 1.76
	Recall	87.92	86.67	83.75	86.67	85.83	87.08	84.58	90.00	86.25	89.58	86.83 ± 1.87
	F1-score	88.47	87.03	84.45	87.95	87.28	88.75	85.29	90.19	86.61	89.77	87.58 ± 1.74
BodEmoDB	Accuracy	89.17	90.33	92.00	90.50	89.50	90.67	90.00	91.50	89.67	89.83	**90.32 ± 0.84**
	Precision	89.70	91.59	92.40	91.22	90.31	92.28	91.81	92.36	90.25	90.99	91.29 ± 0.92
	Recall	88.50	87.17	91.17	90.00	88.50	89.67	87.83	88.67	89.50	89.17	89.02 ± 1.08
	F1-score	89.10	89.33	91.78	90.61	89.40	90.96	89.78	90.48	89.87	90.07	90.14 ± 0.79
IEMOCAP	Accuracy	66.11	65.66	67.11	65.44	70.12	66.78	68.00	67.34	66.56	67.67	**67.08 ± 1.28**
	Precision	66.98	67.97	68.19	66.39	70.83	71.77	74.52	72.62	73.91	71.83	70.50 ± 2.77
	Recall	63.99	62.21	63.10	61.87	68.23	60.65	60.65	57.08	51.17	60.54	60.95 ± 4.25
	F1-score	65.45	64.96	65.55	64.05	69.51	65.74	66.87	63.92	60.47	65.70	65.22 ± 2.18
ESD	Accuracy	97.04	96.79	97.07	96.79	96.54	96.69	96.97	96.69	96.93	96.79	**96.83 ± 0.16**
	Precision	97.07	96.88	97.19	96.87	96.65	96.85	97.07	96.78	97.00	96.84	96.92 ± 0.15
	Recall	96.96	96.73	96.99	96.69	96.43	96.61	96.93	96.61	96.84	96.70	96.75 ± 0.17
	F1-score	97.01	96.80	97.09	96.78	96.54	96.73	97.00	96.69	96.92	96.77	96.83 ± 0.16

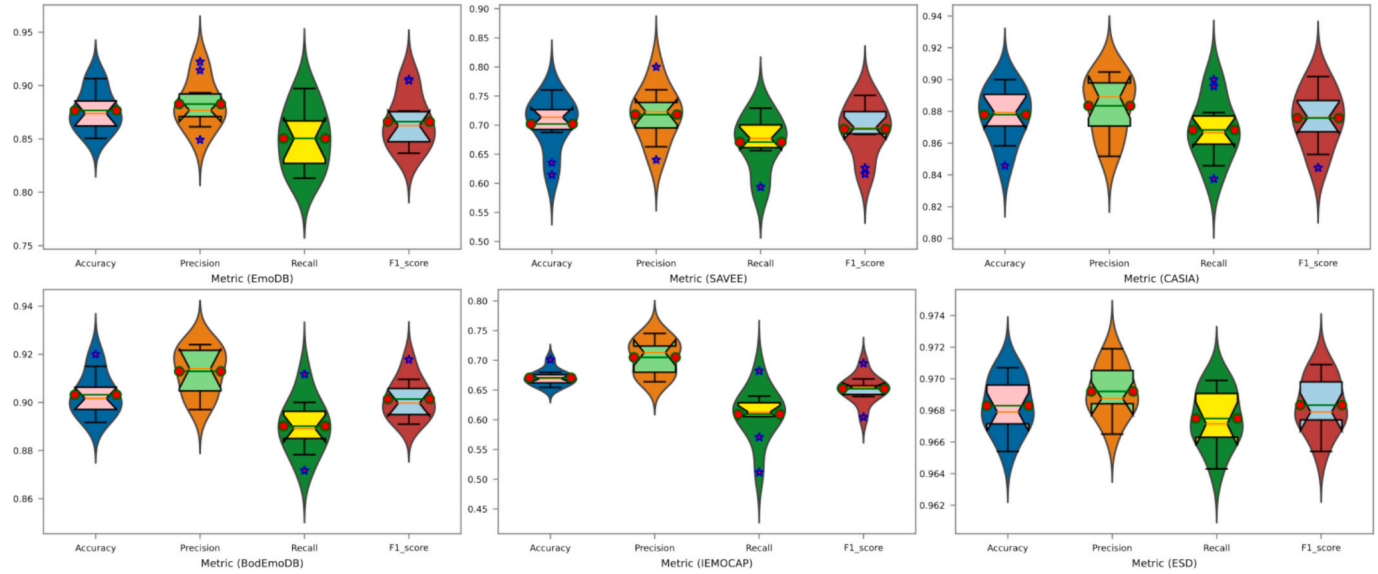


Fig. 2. The robustness comparison of the proposed SpeechNet model across different datasets: EmoDB, SAVEE, CASIA, BodEmoDB, IEMOCAP, and ESD.

with increasing epochs, indicating that the model is learning effectively from the training data. The test loss (orange) also decreases initially but tends to plateau or exhibit slight fluctuations after a certain number of epochs, suggesting the model's performance on unseen data stabilizes over time. In some datasets (e.g., SAVEE, IEMOCAP), the test loss starts to increase slightly or fluctuates more towards the later epochs while the training loss continues to decrease. This is indicative of potential overfitting, where the model learns the training data too well, including noise, but generalizes less effectively to the test data.

The training loss decreases rapidly and stabilizes on the EmoDB dataset, while the test loss shows a similar trend but with more fluctuations towards the end. The training loss decreases steadily on the SAVEE dataset, but the test loss starts increasing slightly after initial stability, suggesting some overfitting. Like SAVEE, there is a noticeable gap between the training and test loss in later epochs on the CASIA dataset, indicating overfitting. Exhibits a significant gap between training and test losses towards the end on the IEMOCAP dataset, suggesting overfitting and possibly high variability in test data. Training

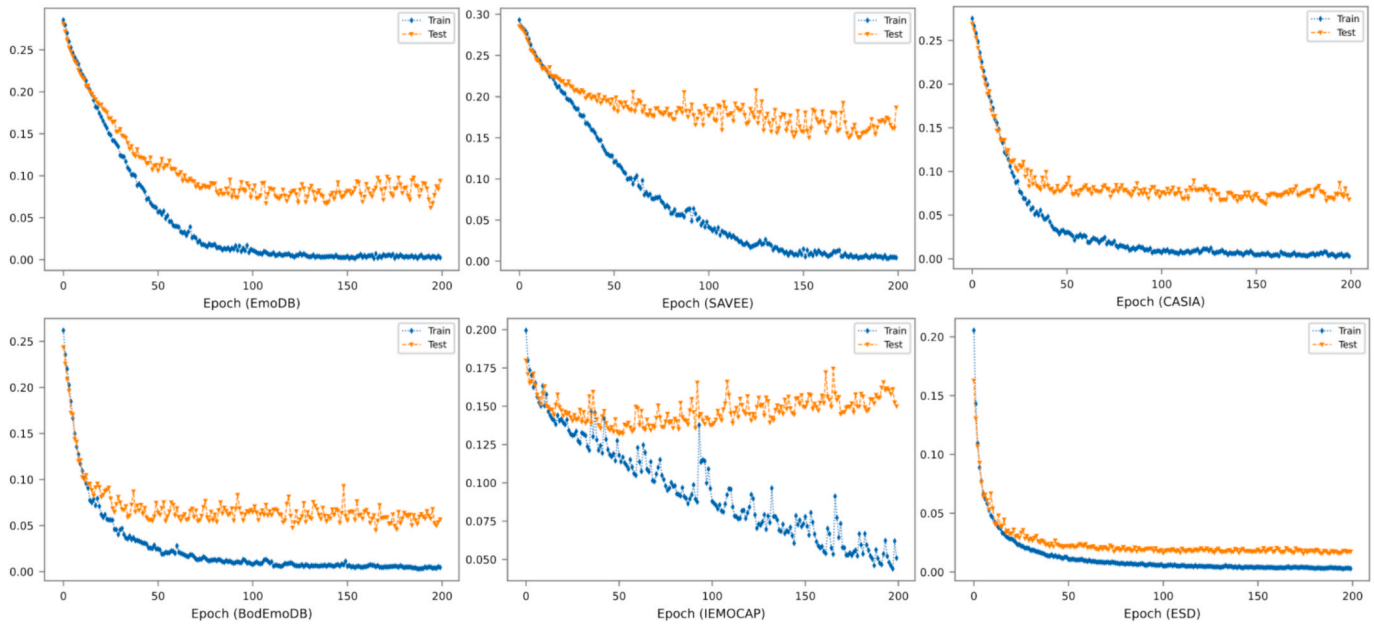


Fig. 3. Loss comparison of the proposed SpeechNet model across six benchmark datasets: EmoDB, SAVEE, CASIA, BodEmoDB, IEMOCAP, and ESD.

and test losses both decrease steadily on the BodEmoDB and ESD datasets, with test loss showing fewer fluctuations, indicating better learning and generalization.

The SpeechNet model shows effective learning across all datasets, as evidenced by the decreasing training loss curves. The model generalizes well on some datasets (e.g., BodEmoDB, ESD) but shows signs of overfitting on others (e.g., SAVEE, IEMOCAP), indicated by the widening gap between training and test losses in later epochs.

Fig. 4 shows the accuracy and F1-score curves of the SpeechNet model for the training and test sets. For all datasets, training accuracy (red) and training F1-score (green) increase steadily with increasing epochs, indicating that the model is improving its performance on the training data. Testing accuracy (blue) and testing F1-score (orange) also generally increase but may exhibit more fluctuations.

On the EmoDB dataset, we observed rapid initial growth in training metrics followed by stabilization, while test metrics improved more irregularly, hinting at some overfitting despite good overall learning. CASIA results showed steady training improvement but inconsistent testing performance. The SAVEE and IEMOCAP datasets revealed a challenging pattern – steady training improvement but significant test metric gaps and fluctuations that point to overfitting problems and possible test data inconsistencies. Both BodEmoDB and ESD datasets produced encouraging results with smooth, consistent improvements in both training and testing metrics that eventually stabilized.

The closeness of the testing metrics to the training metrics is an indicator of the model's generalization ability. BodEmoDB and ESD datasets showed promising results with test performance coming close to training performance, suggesting the model learned these datasets effectively. However, with SAVEE and IEMOCAP, we noticed a wider gap between training and testing results, highlighting areas where better regularization techniques could help prevent the model from becoming too specialized to the training samples.

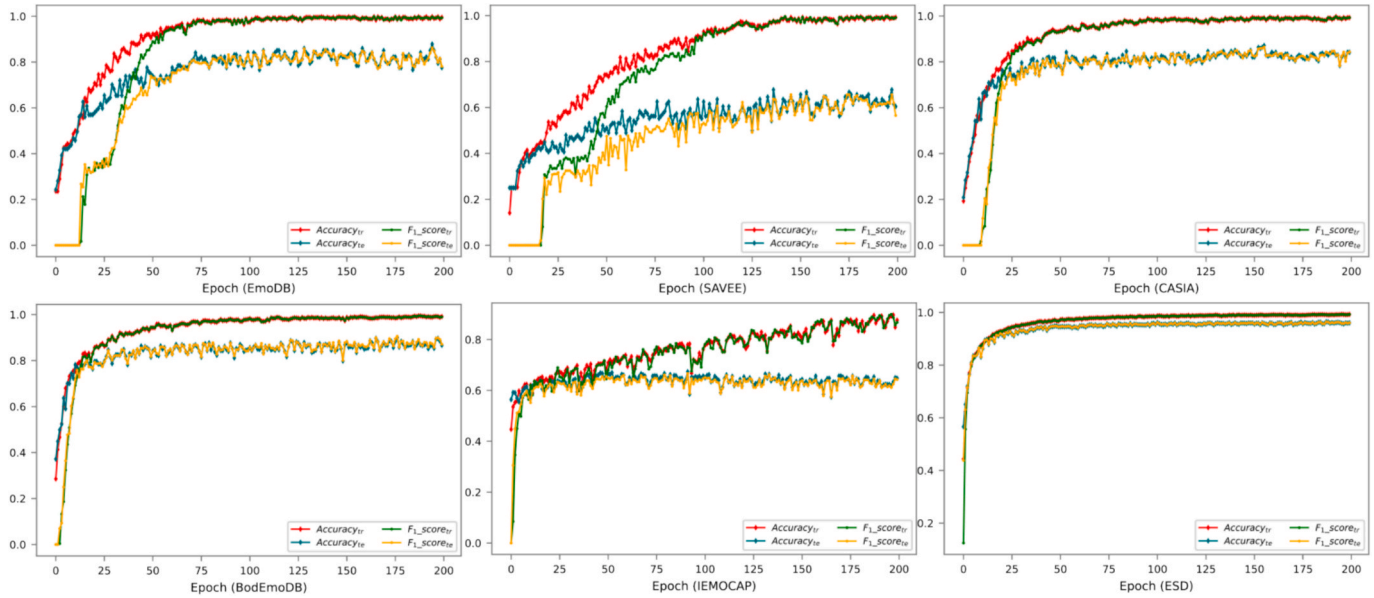


Fig. 4. The accuracy and F1-score comparison of the proposed SpeechNet model across different datasets: EmoDB, SAVEE, CASIA, BodEmoDB, IEMOCAP, and ESD.

Fig. 5 illustrates the confusion matrices of the SpeechNet model across different emotional speech datasets. A quick glance at these matrices reveals that the model generally performs well, as evidenced by the high values along the diagonal, indicating correct classifications. Starting with the EmoDB dataset, the model shows high accuracy for most emotions, such as a perfect 100.00 % for Anger/A and strong performances of 83.33 % and 80.00 % for others. However, there is minor confusion in a few instances, as seen with 8.33 % misclassifications. The proposed model shows a large gap on the SAVEE dataset, with notable diagonal values like 90.91 % and 85.71 %, although certain emotions show some degree of confusion, particularly with a 41.67 % misclassification rate for Fear/F.

The CASIA dataset follows a similar trend, with high accuracy rates such as 89.74 %, 89.47 %, and 88.57 %, but it too has instances of confusion, albeit generally less pronounced than in SAVEE. On the BodEmoDB dataset, the model performs exceptionally well, achieving accuracies like 92.92 %, 94.34 %, and an impressive 99.00 %. This dataset shows minimal confusion, suggesting it might be more straightforward aligned with the model's training.

However, the IEMOCAP dataset presents a more challenging scenario for the model. While it still maintains reasonable accuracy, there is more noticeable confusion among the Happiness/H and Neutral/N, Happiness/N and Sadness/Sa with some diagonal values like 13.33 %, indicating that certain emotions are harder to distinguish. This suggests that IEMOCAP contains more subtle emotional expressions, making it harder for the model to classify accurately. The ESD dataset shows the model's strength, with high performance across the board, such as 97.01 %, 98.04 %, and 97.96 %. Like BodEmoDB, ESD shows minimal confusion, highlighting the model's effectiveness with this dataset.

However, the increased confusion in the IEMOCAP dataset highlights areas for potential improvement. These findings suggest that while the model is generally robust, it could benefit from further fine-tuning and data augmentation to enhance its performance on more complex datasets.

Fig. 6 illustrates a comparison of the performance of the SpeechNet model against several peer methods across six datasets. The peer methods include well-established models such as LSTM [45], GRU [46], CNN [47], Transformer [48], U-Net [49], RBM (Restricted Boltzmann machine) [50], Autoencoder [51], LSM (Liquid state machine) [52], and TCN (Temporal convolutional network) [53].

Evaluating model performance across six datasets, SpeechNet demonstrates better results in emotion recognition compared to alternative approaches. With EmoDB, it outperforms LSTM and Transformer models, showcasing effective feature extraction. For CASIA, SpeechNet recognizes emotions more reliably than GRU and Transformer models, while achieving the highest performance on BodEmoDB with narrower margins, suggesting less classification difficulty with this dataset. SpeechNet consistently performs better than LSTM and Transformer models on SAVEE and maintains stability with the challenging IEMOCAP dataset. On the ESD dataset, SpeechNet achieves the highest accuracy compared to Transformer and CNN models.

Table 4 further validates these findings by comparing Prediction-Level Reliability (PLR) and Prediction Confidence-Level Reliability (PCLR). SpeechNet records the highest scores across all datasets, demonstrating consistent learning ability and reliable predictions. While TIM-Net provides strong competition with challenging datasets like SAVEE and IEMOCAP, SpeechNet remains the top performer. For EmoDB, SpeechNet achieves the highest scores, compared to Transformer's lower scores of 54.21 % and 51.40 %. On the SAVEE dataset, SpeechNet achieves a PLR of 81.25 % and PCLR of 63.54 %, showing stability under demanding conditions.

SpeechNet shows the highest reproducibility on the CASIA dataset, though TIM-Net also performs well. With BodEmoDB, SpeechNet achieves the highest PLR of 93.83 % and PCLR of 83.83 %, while TIM-Net records 83.67 % and 81.33 % respectively. On IEMOCAP, SpeechNet achieves high scores, though CPAC slightly outperforms it in PLR at 80.40 % but scores lower in PCLR at 58.57 %. For ESD, SpeechNet records the highest PLR of 92.66 % and PCLR of 91.44 %, with CPAC also

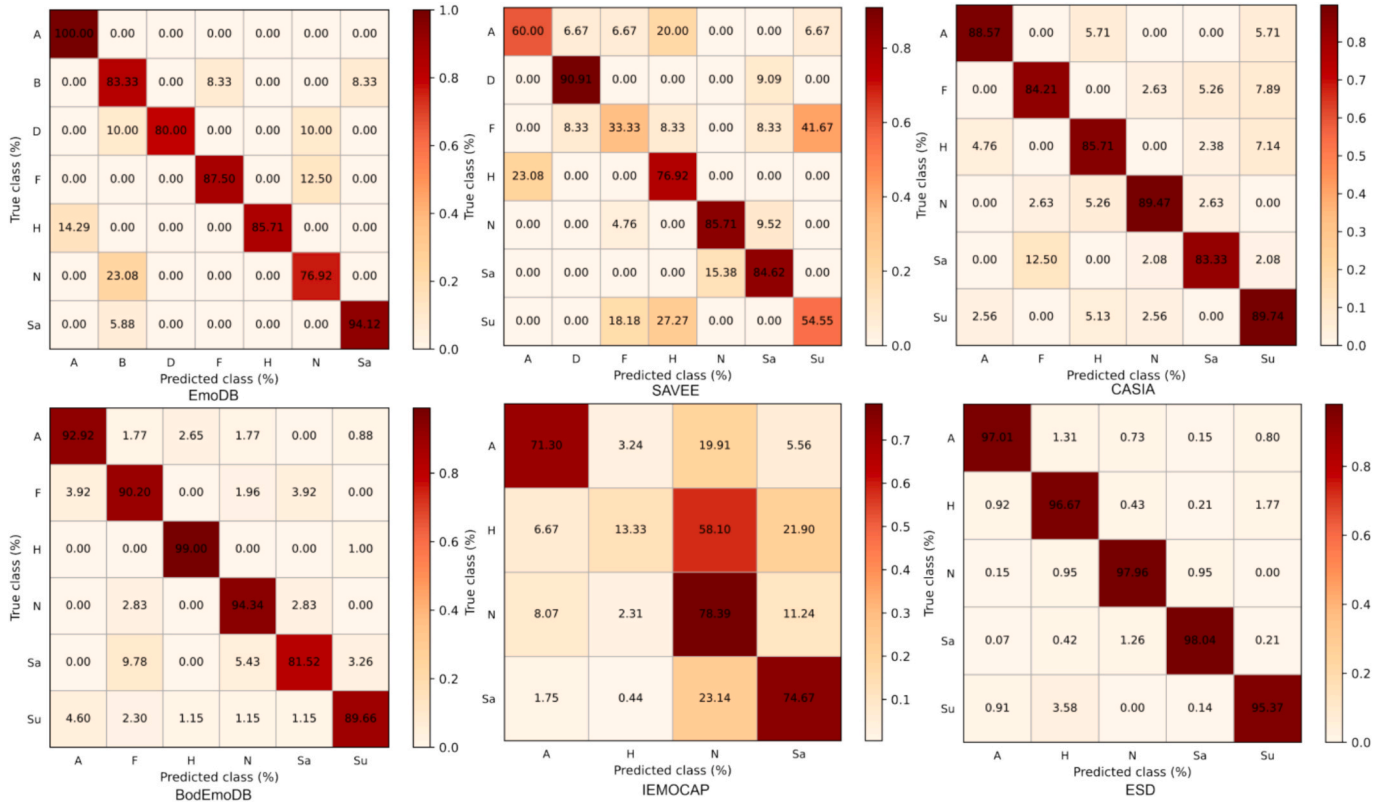


Fig. 5. The confusion matrices comparison of the proposed SpeechNet model across different datasets: EMODB, SAVEE, CASIA, BodEMODB, IEMOCAP, and ESD.

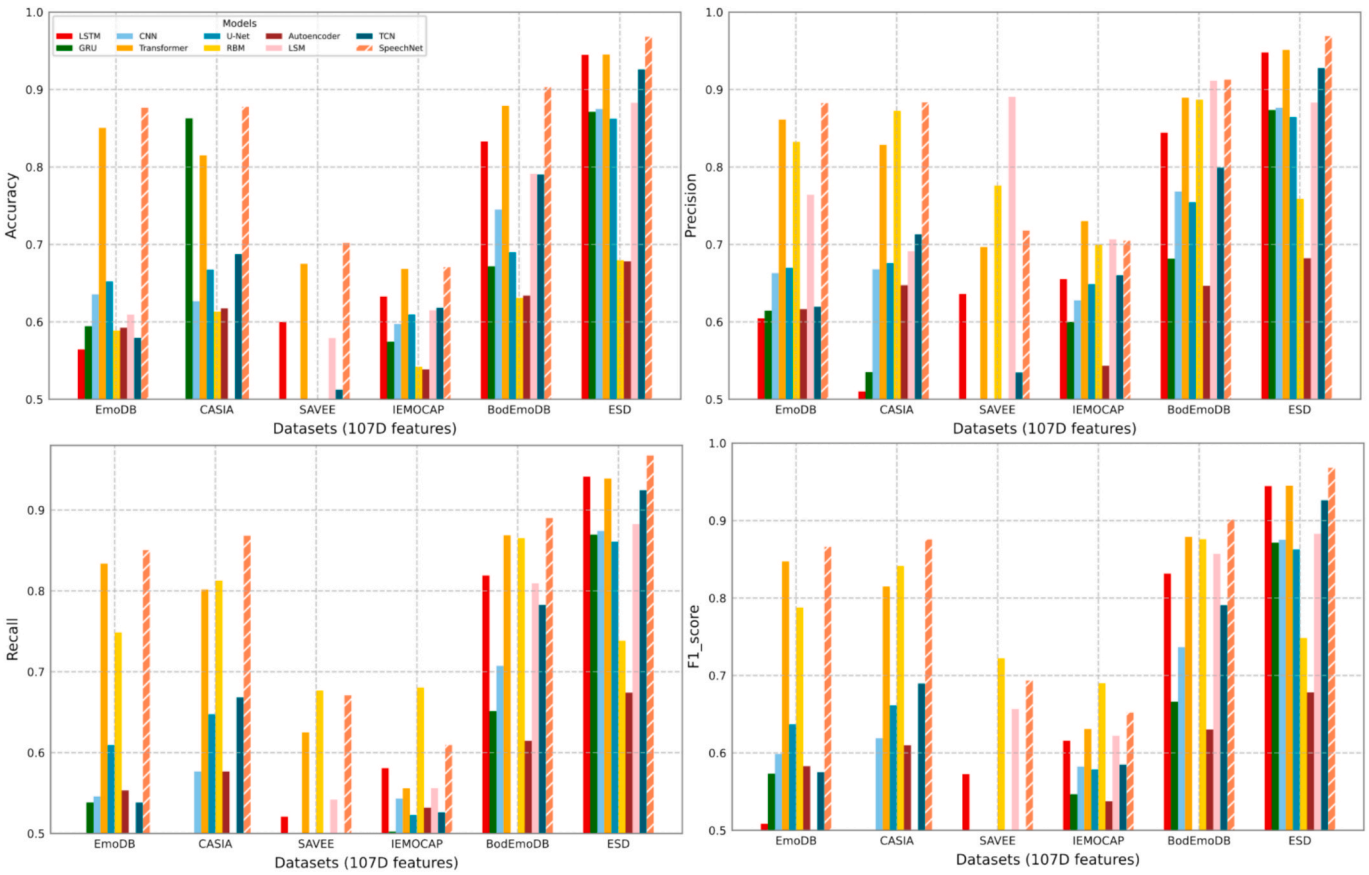


Fig. 6. Performance comparisons of the SpeechNet model with its peers on six benchmark datasets: EmoDB, SAVEE, CASIA, BodEmoDB, IEMOCAP, and ESD. SpeechNet demonstrates consistently leading advantages over the peer methods: LSTM, GRU, transformer, U-Net, and Autoencoder.

Table 4
Perfect learning reproducibility (PLR: %) and Perfect correct learning reproducibility (PCLR: %) of the proposed SpeechNet model and peer models on the six classical benchmark datasets: EmoDB, SAVEE, CASIA, BodEmoDB, IEMOCAP, and ESD.

Models	Reproducibility (%)	Datasets					
		EmoDB	SAVEE	CASIA	BodEmoDB	IEMOCAP	ESD
LSTM	PLR	42.06	20.83	2.08	56.83	53.85	79.76
	PCLR	**34.58**	**14.58**	**0.83**	**56.00**	**39.13**	**79.59**
GRU	PLR	14.95	16.67	0.42	27.83	16.61	55.07
	PCLR	**14.02**	**11.46**	**0.42**	**25.67**	**14.83**	**54.86**
CNN	PLR	19.63	14.58	19.17	31.33	18.84	57.49
	PCLR	**18.69**	**12.50**	**17.92**	**30.33**	**16.05**	**57.09**
Transformer	PLR	54.21	27.08	57.92	69.50	36.23	82.04
	PCLR	**51.40**	**25.00**	**57.08**	**66.50**	**31.10**	**81.63**
Autoencoder	PLR	33.64	21.88	40.83	43.83	30.32	61.40
	PCLR	**21.50**	**16.67**	**32.50**	**34.33**	**22.30**	**49.33**
LSM	PLR	14.95	15.62	2.08	40.50	22.30	62.60
	PCLR	**14.95**	**14.58**	**2.08**	**40.33**	**18.28**	**62.30**
TCN	PLR	12.15	6.25	23.33	2.17	17.06	73.64
	PCLR	**10.28**	**6.25**	**22.08**	**2.17**	**14.72**	**73.21**
CPAC	PLR	53.70	75.00	53.33	70.17	80.40	89.19
	PCLR	**44.44**	**54.17**	**40.00**	**67.67**	**58.57**	**86.57**
TIM-Net	PLR	81.31	70.83	77.50	83.67	61.47	84.66
	PCLR	**69.16**	**60.42**	**71.25**	**81.33**	**45.88**	**84.10**
SpeechNet	PLR	91.59	81.25	92.08	93.83	66.82	92.66
	PCLR	**83.18**	**63.54**	**80.83**	**83.83**	**61.02**	**91.44**

performing well but not matching SpeechNet's robustness.

SpeechNet's architecture effectively manages variability and noise in speech data through advanced learning techniques and appropriate regularization strategies, balancing complexity with generalizability across datasets. This explains its consistently better performance compared to other models, even with challenging datasets like IEMOCAP and SAVEE where all models show lower reproducibility.

SpeechNet is stable and reliable in learning and correctly predicting outcomes. Across all datasets, SpeechNet consistently shows better performance, with the highest PLR and PCLR. TIM-Net performs well but still falls short of SpeechNet's performance, especially in more challenging datasets. IEMOCAP and SAVEE datasets present more

difficulties with lower reproducibility across all models. However, SpeechNet still maintains its lead with a smaller margin. Other models like LSTM, GRU, and CNN show significantly lower reproducibility.

The design of SpeechNet is important and can manage variability and noise commonly found in speech data. By incorporating advanced learning techniques and appropriate regularization strategies, the model maintains training stability while balancing complexity with generalizability across datasets.

Table 5 presents a comprehensive comparison of the proposed SpeechNet model with a range of baseline methods, including classical approaches (e.g., DT-SVM), deep learning models (e.g., CNN-LSTM-GRU, QCNN), and recent state-of-the-art architectures from 2023 to 2025 (e.g., MA-CapsNet, AACCN, PulseEmoNet, hc-former, STACN, and DSTCNet). The evaluation spans six benchmark datasets: EmoDB, SAVEE, CASIA, BodEmoDB, IEMOCAP, and ESD.

SpeechNet achieves 87.66 % accuracy on the EmoDB dataset, which is competitive with the highest-performing model (QCNN at 88.78 %). It outperforms recent models such as MA-CapsNet (77.89 %) and S2ST (85.94 %), indicating strong generalization on this relatively clean and well-structured dataset.

On the more challenging SAVEE dataset, SpeechNet achieves 70.21 %, slightly behind AACCN (72.50 %) and IFN (73.75 %). While the margin is narrow, this reflects the complexity of this dataset, which contains limited samples and pronounced inter-speaker variability. Nevertheless, SpeechNet still surpasses multiple baselines, including MA-CapsNet (59.05 %) and AS-CANet (67.08 %).

On the CASIA dataset, SpeechNet reaches 87.79 %, outperforming many earlier and recent models such as DenseNet-GRU (80.00 %) and DT-SVM (85.08 %). Only STACN (89.17 %) slightly surpasses it, demonstrating that SpeechNet maintains strong performance even against advanced attention-centric architectures.

On the Tibetan-language BodEmoDB dataset, SpeechNet achieves 90.32 %, the highest accuracy among all models evaluated. It exceeds both traditional (e.g., LSTM: 82.33 %) and modern baselines (e.g., AACCN: 89.90 %, PulseEmoNet: 88.70 %), affirming its robustness across culturally diverse and linguistically challenging datasets.

For the IEMOCAP dataset, which is known for its high complexity and emotional ambiguity, SpeechNet achieves 67.08 %. While it trails TIM-Net (72.50 %) and MHA + DRN (67.40 %), it performs on par with or better than many competitive baselines, including Dual-TBNet (64.80 %) and hc-former (68.13 %). The model's competitive performance here is noteworthy given the dataset's multimodal and spontaneous nature.

SpeechNet achieves 96.83 % on the ESD dataset, outperforming all other models, including strong recent baselines such as PulseEmoNet (95.98 %), STACN (95.92 %), and IFN (96.31 %). This result showcases SpeechNet's ability to scale effectively to large, multilingual datasets while maintaining high generalization capacity.

SpeechNet demonstrates superior or highly competitive performance across all datasets. It achieves the highest accuracy on BodEmoDB and ESD. It performs among the top models on EmoDB and CASIA. It maintains respectable results on challenging datasets, such as SAVEE and IEMOCAP, where performance typically varies across models due to noise, speaker variability, and spontaneous speech content. These results confirm that SpeechNet offers a compelling balance between accuracy, generalization, and robustness, outperforming both conventional and recent deep learning baselines across a diverse set of SER tasks.

The SpeechNet model demonstrates its versatility and robustness by achieving better or highly competitive results on the most datasets. Its performance on more challenging and complex datasets, such as BodEmoDB and ESD, highlights its capability to generalize across various SER tasks, positioning it as a state-of-the-art model in this domain. However, there are still opportunities for further refinement on datasets like SAVEE and IEMOCAP, where other models offer slightly better performance.

6. Discussions

The rapid advancement of deep learning techniques has improved SER and created new opportunities for applications in HCI, healthcare, and other domains. Despite these advances, some limitations remain regarding reproducibility and robustness across diverse datasets and real-world scenarios.

Our experimental results show that SpeechNet's integrated architecture addresses these challenges through several key innovations. The combination of LSTM with MHA and CNN components creates a complementary system that captures both temporal dependencies and spatial patterns in emotional speech. Compared with using single architecture types, SpeechNet models the dynamics of emotional expressions. The temporal modeling capabilities of LSTMs work together with the feature extraction strengths of CNNs to improve recognition across different emotional states.

The introduction of the pre-attention mechanism provides another contribution. Unlike conventional attention implementations that process all features equally before applying weights, our pre-attention design emphasizes emotionally salient features early in the processing pipeline through parallel auxiliary layers. This improves the model's ability to isolate relevant emotional cues, particularly evident in the performance gains on complex datasets like IEMOCAP and ESD where emotional feature identification is important.

SpeechNet enhances routing techniques by incorporating adaptive block partitioning, iterative similarity updating, and dynamic weighting mechanisms, which more effectively preserve hierarchical relationships

Table 5
Accuracy (%) comparison of the proposed SpeechNet model against previous models across benchmark datasets, highlighting better performance except on IEMOCAP and SAVEE.

Model	EmoDB	Model	SAVEE	Model	CASIA
3DRNN + Attention [54]	85.82	Attention + CNN + BLSTM [60]	56.50	DT-SVM [65]	85.08
Dual-TBNet [55]	84.10	AlexNet + CFS + FC [61]	66.90	TLFMRF [66]	85.83
QCNN [56]	88.78	EnsembleWave [62]	68.00	DenseNet-GRU [67]	80.00
CNN-LSTM-GRU [57]	67.74	AACCN [63]	72.50	AS-CANet [33]	67.08
MA-CapsNet [58]	77.89	MA-CapsNet [58]	59.05	hc-former [68]	87.08
S2ST [59]	85.94	IFN [64]	73.75	STACN [69]	89.17
SpeechNet (Ours)	***87.66***	***SpeechNet (Ours)***	***70.21***	***SpeechNet (Ours)***	***87.79***
Model	**BodEmoDB**	**Model**	**IEMOCAP**	**Model**	**ESD**
LSTM (Ours)	82.33	MHA + DRN [71]	67.40	LSTM (Ours)	93.61
CNN (Ours)	74.17	QCNN [56]	70.46	CNN (Ours)	88.74
Transformer (Ours)	85.00	TIM-Net [71]	72.50	Transformer (Ours)	93.80
AACCN [63]	89.90	Dual-TBNet [55]	64.80	PulseEmoNet [70]	95.98
IFN [64]	84.47	hc-former [68]	68.13	IFN [64]	96.31
PulseEmoNet [70]	88.70	DSTCNet [72]	61.80	STACN [69]	95.92
SpeechNet (Ours)	***90.32***	***SpeechNet (Ours)***	***67.08***	***SpeechNet (Ours)***	***96.83***

between emotional features. Our experiments demonstrate significant performance improvements, particularly when processing subtle emotional variations.

Though SpeechNet shows good overall performance, our analysis indicates clear differences across datasets. The model works well with BodEmoDB and ESD, reaching high consistency scores, but has more difficulty with IEMOCAP and SAVEE. These differences come from the basic complexity and variety of emotional patterns across different languages and recording conditions.

We also conducted an analysis of computational complexity and inference time. SpeechNet requires approximately 9.5 million FLOPs per forward pass and achieves an average inference time of 3.6 ms per utterance on an NVIDIA RTX 3080 Ti GPU. With ~8.3 million parameters, it maintains a favorable balance between model expressiveness and computational efficiency, making it suitable for real-time applications.

SpeechNet has made contributions to reproducibility in SER beyond performance metrics. The quantitative reproducibility framework provides a standardized approach for evaluating model stability and reliability, important factors for real-world deployment. The consistently high PLR and PCLR across multiple datasets validate SpeechNet's reliability. The integration of complementary neural architecture, combined with attention mechanisms specifically designed for emotional content, represents a promising direction for advancing SER systems that are both accurate and reliable. The practical implications extend to applications, from mental health monitoring systems that require consistent emotion recognition to customer service platforms that need to reliably detect dissatisfaction across diverse speaker populations.

7. Conclusions

This paper presents SpeechNet, a deep learning architecture for SER that tackles key challenges in model reproducibility and robustness. By integrating speech recall, attention mechanism, speech signal refinement, pre-attention mechanism, and modified routing technique, SpeechNet effectively captures temporal dependencies in speech sequences, while its convolutional layers and speech signal refinement module extract spatial patterns from the acoustic features. This dual-pathway architecture enables comprehensive modeling of both temporal dynamics and spatial characteristics of emotional speech.

Our extensive experiments prove that SpeechNet not only achieves better performance compared to existing models (with accurate improvements of 2–5 % across multiple datasets) but also maintains quantifiably high reproducibility across different training runs and datasets, as evidenced by consistently high PLR (>90 %) and PCLR (>80 %). This consistency validates the reliability of SpeechNet, making it a robust tool for practical applications in various domains.

Firstly, we provide a model that sets a new standard in SER by emphasizing reproducibility and robustness. Secondly, we propose a formalized approach to evaluate reproducibility in SER, incorporating both performance metrics and correctness measures. These contributions are crucial for advancing SER research and fostering the development of reliable and practical emotion recognition systems.

SpeechNet addresses key challenges in reproducibility and model reliability. Future research could expand in several directions with advanced attention mechanisms. For example, temporal-scale attention for analyzing emotional patterns across different time frames, structured attention for capturing relationships between speech segments, and adaptive attention systems that adjust to specific input characteristics.

Secondly, integrating multimodal information is important for enhancing emotion recognition. Developing techniques for synchronized audio-visual feature extraction would capture complementary emotional cues that single modalities might miss. Research on cross-modal temporal alignment could address synchronization challenges between speech and visual emotional expressions. Additionally, robust fusion algorithms that handle incomplete or noisy data would improve performance in practical applications where sensor limitations or environmental factors affect data quality.

Third, extending SpeechNet for multilingual and cross-cultural scenarios represents an important direction for future work. which involve developing culture-specific emotional feature extractors, implementing language-agnostic representation learning, and creating adaptive normalization techniques for different cultural contexts. Additionally, transfer learning methods could be developed to leverage knowledge across languages, while cultural aware attention mechanisms could help capture culture-specific emotional expressions.

Future work can build upon this foundation to explore new methodologies, extend the model's capabilities, and further enhance the robustness and applicability of SER systems. We also could explore more advanced attention mechanisms, integrate multimodal data (e.g., combining audio with visual cues), and investigate the application of SpeechNet in multilingual and cross-cultural contexts.

CRediT authorship contribution statement

Huiyun Zhang: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Conceptualization. **Zilong Pang:** Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Puyang Zhao:** Conceptualization, Formal analysis, Validation, Writing – review & editing. **Gaigai Tang:** Conceptualization, Methodology, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Lingfeng Shen:** Data Curation, Formal Analysis. **Guanghui Wang:** Software, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors acknowledge Key Technology Research and Development Project of Henan Province (Grants: 252102310449, 252102211031), the Major Research Plan of National Natural Science Foundation of China (Grants: 92467103, 92367302), Natural Science Foundation of Qinghai Province (Grant: 2022-ZJ-925), and National Natural Science Foundation of China (Grant: 62066039).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bspc.2025.108782.

Data availability

No data was used for the research described in the article.

References

[1] Z. Yuan, C. Chen, S. Li, Disentanglement network: Disentangle the emotional features from acoustic features for speech emotion recognition, in: Proc. ICASSP, Seoul, Korea, 2024, pp. 11686-11690.
[2] K. Zhang, Y. Li, J. Wang, et al., Real-time video emotion recognition based on reinforcement learning and domain knowledge, IEEE Trans. Circuits Syst. Video Technol. 32 (3) (2022) 1034–1047.
[3] Y. Zhou, X. Liang, Y. Gu, Multi-classifier interactive learning for ambiguous speech emotion recognition, IEEE/ACM Trans. Audio Speech Lang. Process. 30 (2022) 695–705.
[4] M. Hou, Z. Zhang, C. Liu, et al., Semantic alignment network for multi-modal emotion recognition, IEEE Trans. Circuits Syst. Video Technol. 33 (9) (2023) 5318–5329.
[5] B. Ma, H. Sun, J. Wang, et al., Extractive dialogue summarization without annotation based on distantly supervised machine reading comprehension in customer service, IEEE/ACM Trans. Audio Speech Lang. Process. 30 (2022) 87–97.

[6] M. Hossain, G. Muhammad, Emotion-aware connected healthcare big data towards 5G, IEEE IoT J. 5 (4) (2018) 2399–2406.
[7] T. Olugbade, L. He, P. Maiolino, et al., Touch technology in affective huma–, robot–, and virtual–human interactions: a survey, Proc. IEEE 111 (10) (2023) 1333–1354.
[8] X. Gu, Z. Cao, A. Jolfaei, et al., EEG-based brain-computer interfaces (BCIs): a survey of recent studies on signal sensing technologies and computational intelligence approaches and their applications, IEEE/ACM Trans. Comput. Biol. Bioinformatics 18 (5) (2021) 1645–1666.
[9] K. Wang, N. An, B. Li, et al., Speech emotion recognition using Fourier parameters, IEEE Trans. Affective Comput. 6 (1) (2015) 69–75.
[10] P. Thanh, N. Huyen, P. Quan, et al., A robust pitch-fusion model for speech emotion recognition in tonal languages, in Proc. ICASSP, Seoul, Korea, 2024, pp. 12386-12390.
[11] X. Ji, Z. Dong, Y. Han, et al., A brain-inspired hierarchical interactive in-memory computing system and its application in video sentiment analysis, IEEE Trans. Circuits Syst. Video Technol. 33 (12) (2023) 7928–7942.
[12] S. Zhang, S. Zhang, T. Huang, Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching, IEEE Trans. Multimedia 20 (6) (2018) 1576–1590.
[13] H. Zhao, Y. Xiao, J. Han, et al., Compact convolutional recurrent neural networks via binarization for speech emotion recognition, in Proc. ICASSP, Brighton, UK, 2019, pp. 6690-6694.
[14] M. Li, Y. Zheng, and D. Li, MS-SENet: Enhancing speech emotion recognition through multi-scale feature fusion with squeeze-and-excitation blocks, in Proc. ICASSP, Seoul, Korea, 2024, pp. 12271-12275.
[15] N. Antoniou, A. Katsamanis, T. Giannakopoulos, et al., Designing and evaluating speech emotion recognition systems: A reality check case study with IEMOCAP, in Proc. ICASSP, Rhodes Island, Greece, 2023, pp. 1-5.
[16] L. Guo, L. Wang, C. Xu, et al., Representation learning with spectro-temporal-channel attention for speech emotion recognition, in Proc. ICASSP, Toronto, Canada, 2021, pp. 6304-6308.
[17] J. Mao, Y. He, and Z. Liu, Speech emotion recognition based on linear discriminant analysis and support vector machine decision tree, in Proc. CCC, Wuhan, China, 2018, pp. 5529-5533.
[18] C. Fu, C. Liu, C. Ishi, et al., An adversarial training-based speech emotion classifier with isolated Gaussian regularization, IEEE Trans. Affective Comput. 14 (3) (2023) 2361–2374.
[19] S. Mao, D. Tao, G. Zhang, et al., Revisiting hidden Markov models for speech emotion recognition, in Proc. ICASSP, Brighton, UK, 2019, pp. 6715-6719.
[20] C. Lee, S. Narayanan, Toward detecting emotions in spoken dialogs, IEEE Trans. Speech Audio Process. 13 (2) (2005) 293–303.
[21] F. Tao and G. Liu, Advanced LSTM: A study about better time dependency modeling in emotion recognition, in Proc. ICASSP, Calgary, AB, Canada, 2018, pp. 2906-2910.
[22] M. Ren, X. Huang, J. Liu, et al., MALN: multimodal adversarial learning network for conversational emotion recognition, IEEE Trans. Circuits Syst. Video Technol. 33 (11) (2023) 6965–6980.
[23] F. Tao and G. Liu, Advanced LSTM: A study about better time dependency modeling in emotion recognition, in Proc. ICASSP, Calgary, AB, Canada, 2018, pp. 2906-2910.
[24] F. Qi, H. Zhang, X. Yang, et al., A versatile multimodal learning framework for zero-shot emotion recognition, IEEE Trans. Circuits Syst. Video Technol. 34 (7) (2024) 5728–5741.
[25] S. Mirsamadi, E. Barsoum, and C. Zhang, Automatic speech emotion recognition using recurrent neural networks with local attention, in Proc. ICASSP, New Orleans, LA, USA, 2017, pp. 2227-2231.
[26] S. Sabour, N. Frosst, and G. Hinton, Dynamic routing between capsules, in Proc. NIPS, Long Beach, CA, USA, 2017, pp. 1-11.
[27] X. Wu, S. Liu, and Y. Cao, Speech emotion recognition using capsule networks, in Proc. ICASSP, Brighton, UK, pp. 6695-6699, 2019.
[28] A. Hashem, M. Arif, Speech emotion recognition approaches: a systematic review, Speech Commun. 154 (2023) 102974.
[29] T. Grósz, M. Singh, S. Kadiri, et al., Aalto's end-to-end DNN systems for the INTERSPEECH 2020 computational paralinguistics challenge, arXiv preprint arXiv: 2008.02689. 2020.
[30] B. Ko, A brief review of facial emotion recognition based on visual information, sensors, vol. 18, no. 2, 2018.
[31] M. Eid, H. Osman, Affective haptics: current research and future directions, IEEE Access 4 (2015) 26–40.
[32] X. Kong, Z. Ge, Deep PLS: a lightweight deep learning model for interpretable and efficient data analytics, IEEE Trans. Neural Netw. Learn. Syst. 34 (11) (2023) 8923–8937.
[33] Y. Liu, H. Sun, W. Guan, et al., A discriminative feature representation method based on cascaded attention network with adversarial strategy for speech emotion recognition, IEEE/ACM Trans. Audio Speech Lang. Process. 31 (2023) 1063–1074.
[34] L. Chen, K. Wang, M. Li, et al., K-means clustering-based kernel canonical correlation analysis for multimodal emotion recognition in human-robot interaction, IEEE Trans. Ind. Electron. 70 (1) (2023) 1016–1024.
[35] H. Zhang, H. Huang, P. Zhao, et al., CENN: capsule-enhanced neural network with innovative metrics for robust speech emotion recognition, Knowl.-Based Syst. 304 (2024) 112499.
[36] K. Zhou, B. Sisman, R. Liu, et al., Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset, in Proc. ICASSP, Toronto, ON, Canada, 2021, pp. 920-924.
[37] Y. Xie, R. Liang, Z. Liang, et al., Speech emotion classification using attention-based LSTM, IEEE/ACM Trans. Audio Speech Lang. Process. 27 (11) (2019) 1675–1685.
[38] P. Schirmer and I. Mporas, Energy disaggregation from low sampling frequency measurements using multi-layer zero crossing rate, in Proc. ICASSP, Barcelona, Spain, 2020, pp. 3777-3781.
[39] K. O'Hanlon and M. Sandler, Comparing CQT and reassignment-based Chroma features for template-based automatic Chord recognition, in Proc. ICASSP, Brighton, UK, 2019, pp. 860-864.
[40] X. Chen, M. Zhang, Y. Liu, Target detection with spectral graph contrast clustering assignment and spectral graph Transformer in hyperspectral imagery, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–16.
[41] X. Chen, H. Chen, Y. Hu, et al., Centroid-oriented extracting transform and its application in seismic spectral decomposition, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–11.
[42] U. Güntürkün, Exploiting the fast spectral roll-off of CPM sidelobes to improve bandwidth efficiency in satellite communications, IEEE Commun. Lett. 21 (7) (2017) 1461–1464.
[43] D. Rzepka, M. Pawlak, D. Kościelnik, et al., Bandwidth estimation from multiple level-crossings of stochastic signals, IEEE Trans. Signal Process. 65 (10) (2017) 2488–2502.
[44] E. Humphrey, T. Cho, and J. Bello, Learning a robust Tonnetz-space transform for automatic chord recognition, in Proc. ICASSP, Kyoto, Japan, 2012, pp. 453-456.
[45] J. Wang, M. Xue, R. Culhane, et al., Speech emotion recognition with dual-sequence LSTM architecture, in Proc. ICASSP, Barcelona, Spain, 2020, pp. 6474-6478.
[46] S. Rajamani, K. Rajamani, and A. Ragolta, A novel attention-based gated recurrent unit and its efficacy in speech emotion recognition, in Proc. ICASSP, Toronto, ON, Canada, 2021, pp. 6294-6298.
[47] Z. Peng, Y. Lu, S. Pan, et al., Efficient speech emotion recognition using multi-scale CNN and attention, in Proc. ICASSP, Toronto, ON, Canada, 2021, pp. 3020-3024.
[48] X. Zhang, M. Li, S. Lin, et al., Transformer-based multimodal emotional perception for dynamic facial expression recognition in the wild, IEEE Trans. Circuits Syst. Video Technol. 34 (5) (2024) 3192–3203.
[49] R. Li, D. Pu, M. Huang, et al., UNET-TTS: Improving unseen speaker and style transfer in one-shot voice cloning, in Proc. ICASSP, Singapore, Singapore, 2022, pp. 8327-8331.
[50] Y. Wang, S. Zhao, J. Li, et al., Speech bandwidth extension using recurrent temporal restricted Boltzmann machines, IEEE Signal Process Lett. 23 (12) (2016) 1877–1881.
[51] J. Deng, Z. Zhang, F. Eyben, et al., Autoencoder-based unsupervised domain adaptation for speech emotion recognition, IEEE Signal Process Lett. 21 (9) (2014) 1068–1072.
[52] R. Lotfidereshgi and P. Gournay, Biologically inspired speech emotion recognition, in Proc. ICASSP, New Orleans, USA, 2017, pp. 5135-5139.
[53] X. Zhang, J. Tang, H. Cao, et al., Cascaded speech separation denoising and dereverberation using attention and TCN-WPE networks for speech devices, IEEE IoT J. 11 (10) (2024) 18047–18058.
[54] M. Chen, X. He, J. Yang, et al., 3-D convolutional recurrent neural networks with attention model for speech emotion recognition, IEEE Signal Process Lett. 25 (10) (2018) 1440–1444.
[55] Z. Liu, X. Kang, F. Ren, Dual-TBNet: improving the robustness of speech features via dual-transformer-BiLSTM for speech emotion recognition, IEEE Trans. Speech Audio Process. 31 (2023) 2193–2203.
[56] A. Muppidi and M. Radfar, Speech emotion recognition using Quaternion convolutional neural networks, in Proc. ICASSP, Toronto, ON, Canada, 2021, pp. 6309-6313.
[57] M. Ahmed, S. Islam, A. Islam, et al., An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition, Expert Syst. Appl. 218 (2023) 119633.
[58] H. Zhang, H. Huang, H. Han, MA-CapsNet-DA: Speech emotion recognition based on MA-CapsNet using data augmentation, Expert Syst. Appl. 244 (2024) 122939.
[59] H. Lin, Y. Lin, H. Chou, et al., Improving speech emotion recognition in under-resourced languages via speech-to-speech translation with bootstrapping data selection, in Proc. ICASSP, Hyderabad, India, 2025: 10887615.
[60] C. Fu, C. Liu, C. Ishi, et al., An end-to-end multitask learning model to improve speech emotion recognition, in Proc. EUSIPCO, Amsterdam, Netherlands, 2021, pp. 1-5.
[61] F. Burkhardt, A. Paeschke, M. Rolfes, et al., A database of German emotional speech, in Proc. Interspeech, vol. 5, 2005, pp. 15171520.
[62] R. Barkur, D. Suresh, A. Narasimhadhan et al., EnsembleWave: An ensembled approach for automatic speech emotion recognition, in Proc. CONECCT, Bangalore, India, 2022, pp. 1-6.
[63] Y. Qi, H. Huang, H. Zhang, Enhanced speech emotion understanding using advanced attention-centric convolutional networks, Biomed. Signal Process. Control 108 (2025) 107936.
[64] H. Zhang, P. Zhao, G. Tang, et al., Reproducible and generalizable speech emotion recognition via an Intelligent Fusion Network, Biomed. Signal Process. Control 109 (2025) 107996.
[65] L. Sun, S. Fu, F. Wang, Decision tree SVM model with Fisher feature selection for speech emotion recognition, EURASIP J. Audio Speech Music Process. 2019 (2019) 2.
[66] L. Chen, W. Su, Y. Feng, et al., Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction, Inf. Sci. 509 (2020) 150–163.
[67] S. Cheng, D. Zhang, and D. Yin, A DenseNet-GRU technology for chinese speech emotion recognition, in Proc. FEICT, 2021, pp. 1-7.

[68] Y. Fan, H. Huang, H. Han, Hierarchical convolutional neural networks with post-attention for speech emotion recognition, Neurocomputing 615 (2025) 28879.
[69] H. Zhang, H. Huang, P. Zhao, et al., Sparse temporal aware capsule network for robust speech emotion recognition, Eng. Appl. Artif. Intel. 114 (2025) 110060.
[70] H. Zhang, G. Tang, H. Huang, et al., PulseEmoNet: Pulse emotion network for speech emotion recognition, Biomed. Signal Process. Control 105 (2025) 107687.
[71] R. Li, Z. Wu, J. Jia, et al., Dilated residual network with multi-head self-attention for speech emotion recognition, in Proc. ICASSP, Brighton, UK, 2019, pp. 6675-6679.
[72] L. Guo, S. Ding, D. Wang, et al., DSTCNet: deep spectro-temporal-channel attention network for speech emotion recognition, IEEE Trans. Neural Netw. Learn. Syst. 36 (1) (2025) 188–197.

ZilongPang is an Associate Professor at the School of Software, Henan University, China. He received his master's degree in Applied Mathematics in Computer Science and pursued his Ph. D. studies at the College of Electronic and Information Engineering, Tongji University, China. His research interests primarily include brain–computer interfaces, artificial intelligence in medical imaging, and machine vision.

HuiyunZhang was born in Qingyang city, Gansu, China in 1993. She received the B.S. degree in educational technology from Shenyang Normal University, China, in 2017, the M.S. degree in computer application technology from Qinghai Normal University, China, in 2020, and the Ph.D. degree in pattern recognition and intelligence system from Qinghai Normal University, China, in 2024. She worked as a Research Assistant in the Data Science and Artificial Intelligence program at Baylor University, United States, for one year. She is currently an Assistant Professor at the School of Software, Henan University, China. Her research direction is pattern recognition and intelligence system. Her research interests include speech emotion recognition, data science, and deep learning. She has published over 20 papers in these fields.