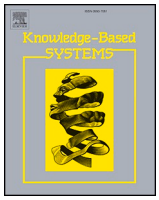
Knowledge-Based Systems 304 (2024) 112499

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



CENN: Capsule-enhanced neural network with innovative metrics for robust speech emotion recognition

Huiyun Zhang [a,*], Heming Huang [b,c], Puyang Zhao [d], Xiaojun Zhu [e], Zhenbao Yu [f]

[a] School of Software, Henan University, Kaifeng, 475004, China
[b] School of Computer Science, Qinghai Normal University, Xining, 810008, China
[c] The State Key Laboratory of Tibetan Intelligent Information Processing and Application, Xining, 810008, China
[d] Department of Biostatistics, The University of Texas Health Science Center at Houston, Houston, TX, 77030, USA
[e] School of Information Engineering, Lanzhou City University, Lanzhou, 730000, China
[f] Global Navigation Satellite System (GNSS) Research Center, Wuhan University, Wuhan 430000, China

ARTICLE INFO

Keywords:
Speech emotion recognition
Multi-head attention
Learning reproducibility
Model complexity
Overfitting

ABSTRACT

Speech emotion recognition (SER) plays a pivotal role in enhancing Human-computer interaction (HCI) systems. This paper introduces a groundbreaking Capsule-enhanced neural network (CENN) that significantly advances the state of SER through a robust and reproducible deep learning framework. The CENN architecture seamlessly integrates advanced components, including Multi-head attention (MHA), residual module, and capsule module, which collectively enhance the model's capacity to capture both global and local features essential for precise emotion classification. A key contribution of this work is the development of a comprehensive reproducibility framework, featuring novel metrics: General learning reproducibility (GLR) and Correct learning reproducibility (CLR). These metrics, alongside their fractional and perfect variants, offer a multi-dimensional evaluation of the model's consistency and correctness across multiple executions, thereby ensuring the reliability and credibility of the results. To tackle the persistent challenge of overfitting in deep learning models, we propose an innovative overfitting metric that considers the intricate relationship between training and testing errors, model complexity, and data complexity. This metric, in conjunction with the newly introduced generalization and robustness metrics, provides a holistic assessment of the model's performance, guiding the application of regularization techniques to maintain generalizability and resilience. Extensive experiments conducted on benchmark SER datasets demonstrate that the CENN model not only surpasses existing approaches in terms of accuracy but also sets a new benchmark in reproducibility. This work establishes a new paradigm for deep learning model development in SER, underscoring the vital importance of reproducibility and offering a rigorous framework for future research.

1. Introduction

Speech emotion recognition (SER) is a pivotal area of research focused on accurately identifying emotional states conveyed through speech, which is essential for a variety of applications such as Human-computer interaction (HCI), mental health monitoring, and online public opinion analysis [1]. The precision of emotion recognition is crucial in these contexts, as it directly impacts the effectiveness and user experience of the systems involved [2].

SER involves the extraction of features from spoken utterances to capture the emotional nuances expressed by the speaker [3]. These features can be broadly categorized into rhythmic prosody elements—such as pitch, rhythm, and intonation—and spectral characteristics like energy distribution and formant frequencies [4]. Key attributes, including pitch, energy, and speech rate, provide invaluable insights for identifying distinct emotions in speech [5].

Typically, SER features are divided into three categories: Low-level descriptors (LLDs) [6], High-level statistical functions (HSFs) [7], and deep learning-derived features [8]. LLDs, particularly those derived from Mel-frequency cepstral coefficients (MFCCs) and energy, effectively capture nuanced emotional content. For instance, Leem et al. demonstrated the robustness of LLD subsets in noisy conditions [10], while Lan et al. proposed an LLD-based DBLSTM approach that incorporates contextual information, outperforming spectral features [9].

* Corresponding author.
E-mail address: zhzhy@henu.edu.cn (H. Zhang).

https://doi.org/10.1016/j.knosys.2024.112499
Received 10 June 2024; Received in revised form 16 August 2024; Accepted 6 September 2024
Available online 7 September 2024

Vu et al. explored lightweight SER models, emphasizing music-related features for enhanced accuracy, efficiency, and generalizability [10].

HSFs, derived from statistical computations on LLDs, encapsulate metrics like standard deviation, skewness, and kurtosis, commonly used in SER. Zheng et al. utilized a combination of LLDs and HSFs to extract emotion-related features from speech, enhancing emotion recognition in children's reading speech [11]. Ananthakrishnan et al. proposed a novel model-based feature set that improved discrimination between emotion classes by relaxing modeling assumptions, leading to significant improvements [12]. Ntalampiras et al. integrated LLDs for emotion recognition, employing methods like short-term statistics, spectral moments, autoregressive models, and wavelet decomposition, highlighting the effectiveness of multiresolution analysis and feature merging for enhanced classification [13].

Deep learning features, autonomously extracted by Deep neural networks (DNNs), capture a wide range of latent features in speech signals. Gao et al. proposed an efficient domain adversarial training method, combining domain-adversarial and center loss to address feature distribution divergence and intra-class variation [14]. Li et al. proposed multi-source discriminant subspace alignment for cross-domain SER, leveraging Linear discriminant analysis (LDA) in a multi-source domain to enhance model robustness, outperforming state-of-the-art transfer learning algorithms [15]. Wu et al. presented a two-stage fuzzy fusion-based CNN for dynamic emotion recognition, integrating facial expression and speech modalities, achieving superior results while effectively managing imbalanced modality contributions [16].

Recent research highlights significant advancements in optimizing system performance and accuracy through sophisticated methodologies. For instance, Cao et al. explored the input-to-state stability of stochastic Markovian jump genetic regulatory networks with time-varying delays, deriving new stability conditions via Lyapunov methods and validating them [17]. Radhika et al. extend this concept to stochastic Cohen–-Grossberg BAM neural networks, focusing on stability with time-varying delays and confirming their findings numerically [18].

In the realm of optimization, Tran et al. introduced a BCMO-ANN algorithm for vibration and buckling optimization in functionally graded porous microplates, combining higher-order shear deformation and modified couple stress theories to effectively analyze material property effects [19]. Meanwhile, Ping et al. presented a hierarchical Bayesian framework for identifying non-Gaussian processes, leveraging improved orthogonal series expansion and polynomial chaos expansion to handle dimensionality and uncertainty, validated through simulations [20]. Dang et al. calibrated 2D VARANS-VOF models of wave interactions using gradient boosting decision trees, achieving high prediction accuracy and minimal error [21]. Nguyen et al. proposed a damage detection method for slab structures using 2D curvature mode shapes and Faster R-CNN, demonstrating robust classification and bounding box predictions [22]. Wang et al. developed a deep learning-based method for rail profile measurement using structured light, integrating deep learning with template-matching algorithms to enhance tracking accuracy for dynamic profiles [23]. Thendral et al. improved image encryption techniques by analyzing the synchronization of Markovian jump neural networks with additive delays, enhancing control performance and encryption effectiveness [24].

SER has seen considerable advancements, yet challenges such as overfitting and limited interpretability persist [25]. Recent approaches have employed Transformer architectures to address these issues, leveraging their capacity to model complex relationships in acoustic data. For instance, Wang et al. introduced the Swin-Transformer, which captures multi-scale emotional features through a patch-based approach [26]. Liu et al. developed Dual-TBNet, combining self-supervised learning with neural network modules to enhance feature fusion and mitigate overfitting [27]. Wagner et al. explored Transformer-based models like wav2vec 2.0 and HuBERT, demonstrating robust performance in valence prediction [28].

However, these advancements have not fully resolved issues related to robustness, generalization, and reproducibility, particularly in maintaining consistent performance across varied datasets and model executions. To address these challenges, we introduce the Capsule-enhanced neural network (CENN). This novel architecture integrates advanced deep learning techniques, including a reproducibility framework with General learning reproducibility (GLR) and Correct learning reproducibility (CLR), to ensure reliable and consistent performance. Additionally, our work introduces an innovative overfitting metric that assesses the interplay between training/testing errors, model complexity, and data complexity, alongside generalization and robustness metrics. By tackling these critical issues, our work advances SER and establishes new benchmarks for model performance and reproducibility. Our contributions are threefold:

(1) We introduce the CENN, which integrates MHA, residual module, and capsule module to improve the capture of both global and local features for accurate emotion classification in SER systems.
(2) We present a novel reproducibility framework with metrics like GLR and CLR, along with fractional and perfect variants, to evaluate and ensure the model's consistency and accuracy across multiple runs.
(3) We propose a new overfitting metric that addresses the relationship between training/testing errors, model complexity, and data complexity, complemented by generalization and robustness metrics, to provide a holistic assessment of the model's performance and guide effective regularization strategies.

The paper is structured as follows: Section 2 introduces the innovative CENN model. Section 3 presents innovative metrics in deep learning models. Section 4 details the datasets and extracted features. Section 5 presents the experimental results. Section 6 offers a discussion of the findings, while Section 7 concludes with an evaluation of the method's strengths and weaknesses, as well as ongoing and future work.

2. The Capsule-enhanced neural network

In this paper, we introduce a novel Capsule-enhanced neural network (CENN), meticulously designed to achieve robust and precise data classification. The architecture of the CENN, as depicted in Fig. 1, integrates several advanced deep learning modules, including a feature mapping module composed of dense blocks, a Multi-head attention (MHA) module [30], a residual (ResNet) module [31], and a capsule module [32]. This combination empowers the model to adeptly capture both global and local patterns within the data, rendering it highly adaptable to various input types and classification tasks. The CENN model represents a flexible and powerful framework that leverages the strengths of state-of-the-art deep learning techniques, making it suitable for tasks that demand nuanced data analysis across diverse domains.

The **feature extraction module** is fundamental to the CENN's ability to distill essential information from the input data. This module generates 107-dimensional low-level features, encapsulating critical aspects of the input, which are then processed by the feature mapping module. Within the **feature mapping module**, each dense block is structured to sequentially process the input features through a dense layer followed by Batch normalization (BN) [29], LeakyReLU activation [17][5], and dropout regularization [10]. This architecture stabilizes the learning process and improves convergence by addressing issues like vanishing gradients and overfitting issues. The sequential processing through dense layers enhances feature richness and diversity, crucial for capturing nuanced patterns in the data.

The **Multi-head attention (MHA) module** enhances the model's ability to capture relevant features across the entire input space by applying multiple attention heads simultaneously. Each attention head computes queries (Q), keys (K), and values (V), enabling the model to dynamically weigh the importance of various features. This mechanism

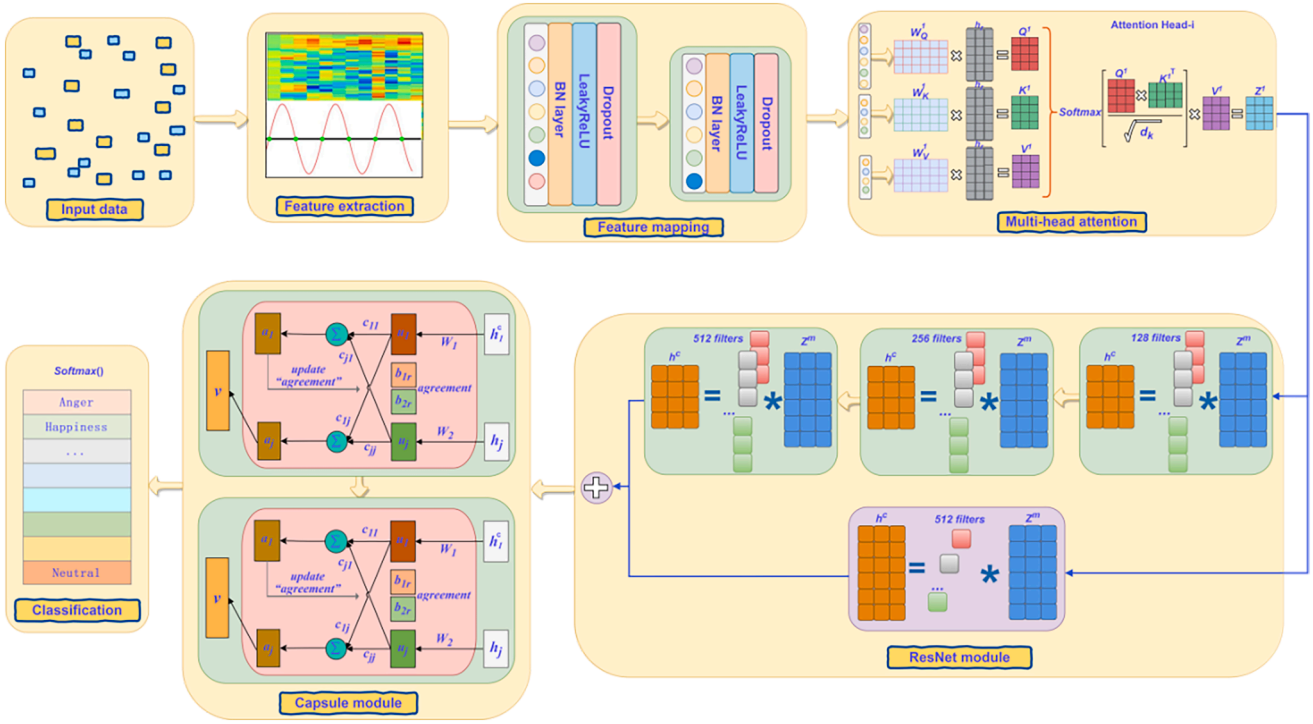


Fig. 1. Architecture of the proposed Capsule-enhanced neural network (CENN) model. The model integrates several distinct modules: an input layer for initial data processing, a feature mapping module for encoding features, a multi-head attention module for selective focus on data segments, a ResNet module for capturing local patterns and hierarchies, a capsule module for learning complex spatial relationships, and a classification layer for final output determination.

is crucial for understanding complex dependencies within the data, as it allows the model to focus on different aspects of the input during the learning process. The theoretical foundation of attention mechanisms is rooted in the Transformer architecture, which has demonstrated superior performance in various natural language processing tasks [30], and its application in CENN extends these benefits to other domains requiring intricate feature extraction.

The **ResNet module** is incorporated to capture local patterns and hierarchies within the data through residual connections. These connections are vital for addressing the vanishing gradient problem, which often hampers the training of deep networks [31]. By preserving the flow of gradients through skip connections, the ResNet module enables the construction of deeper architectures, allowing the model to learn more abstract representations of the input data. This capability is particularly important in tasks where the recognition of fine-grained details can significantly impact classification accuracy.

The **Capsule module** introduces a novel approach to encoding spatial relationships and hierarchies within the data. Unlike traditional neurons, capsules are designed to capture the orientation and relative spatial positioning of features within the input [32]. The transformation of input features into primary capsules, each representing distinct properties, allows the model to learn and represent complex spatial hierarchies more effectively [33]. This enhanced representation provides the CENN model with a significant advantage in understanding intricate relationships within the data, enabling more accurate and detailed classification.

Finally, the **classification layer** leverages the processed feature representations to assign probabilities to various classes, such as Anger, Happiness, and Neutral, through a softmax function. This output layer translates the learned representations into actionable insights, making it suitable for applications requiring precise categorization of input data.

3. Innovative metrics in deep learning models

3.1. Reproducibility

Reproducibility is a cornerstone of scientific research, particularly in deep learning, where models are often sensitive to initial conditions, stochastic processes, and data variability [34]. In this work, we define reproducibility as the model's ability to consistently generate the same results across multiple executions using identical training and testing datasets. Ensuring reproducibility is crucial for validating a model's reliability and robustness, thereby confirming its applicability in real-world scenarios.

We propose a structured framework to evaluate the reproducibility of our CENN model. This framework categorizes reproducibility into two primary dimensions: general learning reproducibility and correct learning reproducibility, each further subdivided to capture varying degrees of reproducibility.

General learning reproducibility (GLR) assesses the model's ability to generate consistent predictions across multiple executions, which is crucial for evaluating the stability of the model's performance. GLR is computed as follows:

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

where N is the total number of samples, R is the number of testing rounds conducted for each sample, and $Count_{i,r}$ denotes the count of consistent predictions for sample$i$during the r^{th}testing round. The $I(\cdot)$is the indicator function, which equals 1 if the condition inside is true and 0 otherwise. It calculates the average number of samples that have consistent predictions at least half the time across multiple testing rounds.

Fractional learning reproducibility (FLR), a specific measure under GLR, focuses on the model's ability to generate consistent results

in most but not all testing rounds:

$$FLR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{R}\sum_{r=1}^{R} I\left(\frac{R}{2} \le Count_{i,r} < R\right) \quad (2)$$

Perfect learning reproducibility (PLR) assesses whether the model produces identical results across all testing rounds, indicating maximum consistency:

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

Correct learning reproducibility (CLR) focuses on the model's ability to not only be consistent but also correct, i.e., the predicted label $\hat{y}_{i,r}$matches the true label $y_i$across multiple rounds. CLR measures the proportion of samples for which the model consistently predicts the correct label at least half the time across multiple testing rounds:

$$CLR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{R}\sum_{r=1}^{R} I\left(Count_{i,r} \ge \frac{R}{2} \ \wedge \ \forall r, \hat{y}_{i,r} = y_i\right) \quad (4)$$

where $\hat{y}_{i,r}$ is the predicted label for sample i in round r. y_i is the true label for sample i.

Fractional correct learning reproducibility (FCLR) measures the model's ability to consistently produce correct results more than half the time but not always. It is expressed as:

$$FCLR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{R}\sum_{r=1}^{R} I\left(\frac{R}{2} \le Count_{i,r} < R \ \wedge \ \forall r, \ \hat{y}_{i,r} = y_i\right) \quad (5)$$

Perfect correct learning reproducibility (PCLR) evaluates whether the model can consistently produce correct predictions across all rounds. This can be mathematically expressed as:

$$PCLR = \frac{1}{N}\sum_{i=1}^{N} I\left(Count_i = R \ \wedge \ \forall r, \hat{y}_{i,r} = y_i\right) \quad (6)$$

3.2. Overfitting metric

To effectively address the overfitting issues [35] of the proposed CENN model, we have introduced a novel overfitting metric $OF\ (\lambda, i,\ y_i,\ \hat{y}_i)$that takes into account the training error $E_{tr}(y_i,\ \hat{y}_i)$, test error $E_{te}(y_i,\ \hat{y}_i)$, model complexity $MC(\lambda,\ i,\ p)$, and data complexity $DC(x_i, \mu,\ k)$. This metric is defined as follows:

$$OF\ (\lambda, i,\ y_i,\ \hat{y}_i) = \ (E_{tr}(y_i,\ \hat{y}_i) - E_{te}(y_i,\ \hat{y}_i))^2 + \frac{DC(\mu,\ i,\ k,)}{MC(\lambda,\ i,\ p)} \quad (7)$$

For the train error $E_{tr}(y_i,\ \hat{y}_i)$ and the test error $E_{te}(y_i,\ \hat{y}_i)$, which are calculated as follows:

$$E_{tr}(y_i,\ \hat{y}_i) = \frac{1}{m}\sum_{i=1}^{m} L(y_i,\ \hat{y}_i) \quad (8)$$

$$E_{te}(y_i,\ \hat{y}_i) = \frac{1}{n}\sum_{i=1}^{n} L(y_i,\ \hat{y}_i) \quad (9)$$

where the m and n denotes the number of training samples and test samples, respectively. The $L(y_i,\ \hat{y}_i)$ is the loss function that measures the difference between the true label $y_i$and the predicted value $\hat{y}_i$ for the i^{th} sample.

For the model complexity $MC(\lambda, i, p)$, it is defined as:

$$MC(\lambda, i,\ p) = \lambda\left(\sum_{i=1}^{N} \| \ W_i \ \|_p\right) \quad (10)$$

where the λ is the regularization parameter, which is used to control the degree of penalty on model complexity. The N is the number of layers in the proposed CENN model. The W_i represents the weight matrix of the i^{th}layer, the $\| \ W_i\|_p$ represents the norm of the weight matrix W_i, where the p is the type of norm and can be the L_1norm, L_2norm, etc. [36]. Here, we evaluate the performance and reproducibility through the L_8 norm computation, validated by an ablation study as detailed below:

$$L_8(W_i) = \| \ W_i \ \|^8 = \left(\sum_{j=1}^{r}\sum_{k=1}^{c} |w_{jk}|^8\right)^{\frac{1}{8}} \quad (11)$$

where the r and the c represent the number of rows and columns in the matrix, respectively. The W_{jk}denotes the element at row j and column k of the weight matrix W_i. It illustrates the process of computing the L_8norm for a matrix W_i, which quantifies the complexity or high-order differences in its elements.

For the data complexity $DC(x_i, \mu)$, which is a metric that reflects the complexity of a dataset, this metric combines factors such as noise $N(x_i, \mu)$, data dimensionality$d$, data distribution$D(i,\ k)$, and class imbalance $CI(i)$, it is defined as:

$$DC(x_i, \mu) = \frac{ln(1 + N(x_i,\ \mu))\cdot D(i,\ k)^2}{1 + \sqrt{d}}\cdot CI(i) \quad (12)$$

where $ln(1 + N(x_i, \mu,\ i))$represents the natural logarithm transformation of noise $N(x_i, \mu)$, which is typically used to mitigate the impact of noise. This part accounts for the negative impact of noise on data complexity. The square of data distribution complexity $D(i,\ k)^2$ emphasizes the distribution characteristics of the data. The $1 + \sqrt{d}$ considers the influence of data dimensionality, with the $\sqrt{d}$ part highlighting the importance of data dimensionality on data complexity. The class imbalance $CI(i)$ measures the balance situation of different classes in the data, where a higher value indicates greater class imbalance and it is calculated: $CI(i) = 1 - max(p(c_i))$, $p(c_i)$ denotes the proportion of instances belonging to class c_i.

3.3. Generalization and robustness metrics

To further enhance the reproducibility of the proposed CENN model, we introduce two key metrics: the generalization metric [37] and the robustness metric [38]. These metrics provide a comprehensive assessment of the model's performance by considering factors such as overfitting, model complexity, data complexity, and class imbalance.

For the generalization metric $Ge(\lambda,\ \mu, y_i,\ \hat{y}_i)$, which is a measure of how well the proposed CENN model can generalize its performance to unseen or new data. It is a composite metric that takes into account various aspects of the model complexity $MC(\lambda, i,\ p)$, data complexity $DC(x_i, \mu)$, overfitting metric $OF\ (\lambda, i,\ y_i,\ \hat{y}_i)$, and the impact of class imbalance term $CI(i)$. The generalization metric $Ge(\lambda,\ \mu,\ y_i,\ \hat{y}_i)$ is defined as follows:

$$Ge(\lambda,\ \mu,\ y_i,\ \hat{y}_i) = \frac{MC(\lambda,\ i,\ p)(1 - DC(x_i, \mu)) - OF\ (\lambda,\ p,\ y_i,\ \hat{y}_i)}{1 + CI(i)} \quad (13)$$

The robustness metric $Ro(\lambda,\ \mu,\ y_i,\ \hat{y}_i)$ assesses the resilience of the CENN model by considering its performance (i.e., accuracy. Here, we use the p to represent the predicted correct probability), susceptibility to overfitting metric $OF\ (\lambda,\ p,\ y_i,\ \hat{y}_i)$, model complexity $MC(\lambda, i,\ p)$, data complexity$DC(x_i, \mu)$, and generalization capabilities $Ge(\lambda,\ \mu,\ y_i,\ \hat{y}_i)$. The robustness metric $Ro(\lambda,\ \mu,\ y_i,\ \hat{y}_i)$ensures that the model maintains high performance under various conditions and is defined as follows:

$$Ro(\lambda,\ \mu,\ y_i,\ \hat{y}_i) = \frac{p(1 - OF(\lambda,\ i,\ y_i,\ \hat{y}_i))}{MC(\lambda,\ i,\ p) + DC(x_i, \mu) + Ge(\lambda,\ \mu, y_i,\ \hat{y}_i)} \quad (14)$$

3.4. Regularization and loss function adaptation

During the training process of the proposed CENN model, we

continuously monitor the overfitting metric $OF\ (\lambda, i,\ y_i,\ \widehat{y}_i)$, robustness metric $Ro(\lambda,\ \mu, y_i,\ \widehat{y}_i)$, and generalization metric $Ge(\lambda,\ \mu,\ y_i,\ \widehat{y}_i)$. If the overfitting metric $OF\ (\lambda, i,\ y_i,\ \widehat{y}_i)$ falls below a predefined threshold θ, we adopt the cross-entropy loss function $L_{ce}\left(y_{i,k},\ \widehat{y}_{i,k}\right)$ for training, defined as follows:

$$L_{ce}\left(y_{i,k},\ \widehat{y}_{i,k}\right) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k}\cdot log\left(\widehat{y}_{i,k}\right) \tag{15}$$

If the overfitting metric exceeds the threshold ($OF\ (\lambda, i,\ y_i,\ \widehat{y}_i) \geq \theta$), we apply L_2 regularization [36] across all dense layers of the proposed model, that is, we need to regularize the sum of squares of all elements in the weight matrix w_{ij}, which is equivalent to the square of the Frobenius norm of the weight matrix. It is used to measure the size of a matrix to prevent large weights and overfitting. Specifically, we introduce a regularization term $L_2(\gamma,\ w_{ij})$ and an Improvement trade-off ratio $ITR(\lambda, y_i,\ \widehat{y}_i)$ into original loss function $L_{ce}\left(y_{i,k},\ \widehat{y}_{i,k}\right)$. The regularization term $L_2(\gamma,\ w_{ij})$ is defined as:

$$L_2\ (\gamma, w_{ij}) = \frac{\gamma}{2}\sum_{i=1}^{N}\sum_{j=1}^{M} w_{ij}^2 \tag{16}$$

The Improvement Trade-off Ratio $ITR(\lambda, y_i,\ \widehat{y}_i)$ comprises the overfitting metric $OF\ (\lambda, i,\ y_i,\ \widehat{y}_i)$, data complexity $DC(x_i, \mu)$, Wasserstein distance $W(P, Q)$, and robustness metric $Ro(\lambda,\ \mu,\ y_i,\ \widehat{y}_i)$, and is defined as:

$$ITR(\lambda, y_i,\ \widehat{y}_i) = \frac{1 - OF\ (\lambda,\ i,\ y_i,\ \widehat{y}_i)}{1 + DC(x_i, \mu)(1 - W(P, Q))}\cdot Ro(\lambda,\ \mu,\ y_i,\ \widehat{y}_i) \tag{17}$$

We iteratively calculate the data complexity $DC(x_i, \mu)$ and overfitting metric $OF\ (\lambda, i,\ y_i,\ \widehat{y}_i)$. If $OF\ (\lambda, i,\ y_i,\ \widehat{y}_i) \geq DC(x_i, \mu)$, we adopt a combined loss function $L_{c1}\left(\gamma,\ w_{ij}, y_{i,k},\ \widehat{y}_{i,k}\right)$, which incorporates the cross-entropy loss $L_{ce}\left(y_{i,k},\ \widehat{y}_{i,k}\right)$, regularization term $L_2(\gamma,\ w_{ij})$, and Improvement Trade-off Ratio $ITR(\lambda, y_i,\ \widehat{y}_i)$ across all dense layers, as detailed follows:

$$\begin{aligned} L_{c1}\left(\gamma,\ w_{ij}, y_{i,k},\ \widehat{y}_{i,k}\right) &= L_{ce}\left(y_{i,k},\ \widehat{y}_{i,k}\right) + L_2\ (\gamma, w_{ij}) + ITR(\lambda,\ y_i,\ \widehat{y}_i) \\ &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k}\cdot log\left(\widehat{y}_{i,k}\right) + \frac{\gamma}{2}\sum_{i=1}^{N}\sum_{j=1}^{M} w_{ij}^2 \\ &\quad + \frac{1 - OF\ (\lambda,\ p,\ y_i,\ \widehat{y}_i)}{1 + DC(x_i, \mu)(1 - W(P, Q))}\cdot Ro(\lambda,\ \mu,\ x_i,\ y_i,\ \widehat{y}_i) \end{aligned} \tag{18}$$

Additionally, we also iteratively compute the model complexity $MC(\lambda,\ p)$ and Simpson diversity index $S(i)$, a measure of diversity within a dataset. If $MC(\lambda,\ p) > S(i)$, we need to adopt a combined loss function $L_{c2}\left(\gamma,\ w_{ij}, y_{i,k},\ \widehat{y}_{i,k}\right)$ to train the proposed CENN model, which consists of crossentropy $L_{ce}\left(y_{i,k},\ \widehat{y}_{i,k}\right)$ and $L_2(\gamma,\ w_{ij})$ regularization, $L_2(\gamma,\ w_{ij})$ regularization is used for all layers of the model. Specifically, this loss function is defined as:

$$\begin{aligned} L_{c2}\left(\gamma,\ w_{ij}, y_{i,k},\ \widehat{y}_{i,k}\right) &= L_{ce}\left(y_{i,k},\ \widehat{y}_{i,k}\right) + L_2(\gamma,\ w_{ij}) \\ &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k}\cdot log\left(\widehat{y}_{i,k}\right) + \frac{\gamma}{2}\sum_{i=1}^{N}\sum_{j=1}^{M} w_{ij}^2 \end{aligned} \tag{19}$$

4. Datasets and feature extraction

The proposed CENN model's performance is rigorously evaluated across a diverse spectrum of benchmark datasets, including BodEMODB, EMODB, CASIA, SAVEE, IEMOCAP, and ESD [27,39]. This thorough assessment ensures a comprehensive and robust evaluation of the model's capabilities on a wide range of datasets, thereby affirming its effectiveness.

The BodEMODB Tibetan speech emotion dataset, meticulously recorded by our group at the Pattern Recognition Laboratory of Qinghai Normal University, captures the emotional expressions of 10 speakers (5 male, 5 female) across 6 distinct emotional states using 50 Tibetan texts. Each speaker contributes 300 sentences, resulting in a comprehensive dataset of 3000 samples representing emotions such as anger (A), fear (F), happiness (H), neutral (N), sadness (Sa), and surprise (Su).

The EMODB, a comprehensive German speech emotion dataset, consists of 535 sentences recorded by 10 speakers. It encompasses seven distinct emotional states: boredom (B), anger (A), fear (F), sadness (Sa), disgust (D), happiness (H), and neutral (N).

The CASIA Chinese speech emotion dataset consists of 1200 utterances contributed by 4 speakers, with each speaker providing 300 sentences. It covers 6 emotional categories: anger (A), fear (F), happiness (H), neutral (N), sadness (Sa), and surprise (Su).

The SAVEE emotion dataset includes 480 sentences from 4 speakers, representing 7 emotional states: anger (A), disgust (D), fear (F), happiness (H), sadness (Sa), surprise (Su), and neutrality (N). Notably, the neutral (N) category comprises 120 samples, whereas each of the other emotional states is represented by 60 samples.

The IEMOCAP dataset contains around 12 h of audio and video recordings from 10 actors, featuring both scripted and spontaneous interactions. It includes 1708 neutral samples, 1103 anger samples, 1084 sadness samples, and 590 happiness samples.

The ESD dataset features 350 parallel utterances from 10 Mandarin and 10 English speakers, covering 5 emotional states: neutral (N), happiness (H), anger (A), sadness (Sa), and surprise (Su). Each speaker provides 35 utterances, making it ideal for research in speech synthesis, voice conversion, and cross-lingual emotional speech generation.

To ensure transparency in our data preprocessing pipeline for the mentioned datasets, we detail the feature extraction process for our proposed CENN model. Our input comprises a comprehensive 107-dimensional fusion feature set, encompassing crucial elements such as Zero crossing rate (ZCR), MFCC, chroma, spectral contrast, first-order MFCC deltas (ΔMFCC), spectral centroid, spectral rolloff, second-order MFCC deltas (ΔΔMFCC), tonnetz, and spectral bandwidth for each frame [6,10].

To capture feature variations effectively, we segment each speech sample into frames using a 25 ms window and a 10 ms shifting step in the CENN model. Recognizing feature fluctuations across emotional states, we include a crucial preprocessing step: feature normalization. This ensures standardization and comparability, eliminating magnitude disparities and reducing prediction errors.

5. Experimental results

This section provides a comprehensive evaluation of performance and reproducibility across multiple datasets, including EMODB, CASIA, SAVEE, BodEMODB, IEMOCAP, and ESD. Datasets are randomly partitioned, with 90 % allocated to the training set and 10 % to the test set. To ensure robustness, each experiment is conducted ten times per dataset. The results are analyzed by computing both the mean and standard deviation.

5.1. The performance of the CENN model

Fig. 2 presents the training and testing accuracy curves for the CENN model across six datasets. The model demonstrates strong convergence and high accuracy on the EMODB, CASIA, BodEMODB, and ESD datasets, which can be attributed to the diverse and high-quality samples that promote robust feature learning and generalization. The model's architecture, particularly the MHA module and residual module, appears well-suited to these datasets, enhancing emotion recognition and optimizing gradient flow. Furthermore, the sequential capsule layers contribute to capturing complex spatial hierarchies and intricate feature relationships.

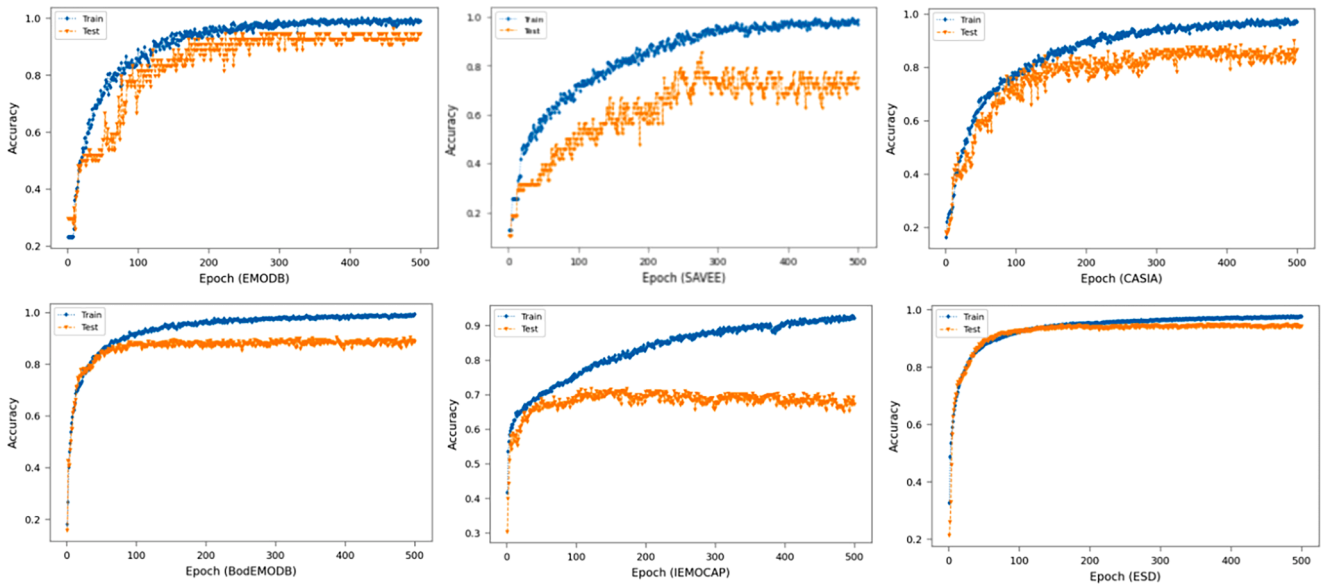


Fig. 2. Accuracy curves of the CENN model across six datasets: EMODB, SAVEE, CASIA, BodEMODB, IEMOCAP, and ESD datasets. The CENN model shows strong performance but tends to overfit on certain datasets (EMODB, SAVEE, CASIA, IEMOCAP), while BodEMODB and ESD demonstrate better generalization.

Training accuracy consistently improves across all datasets, reflecting effective learning processes. However, testing accuracy varies, highlighting differences in the model's generalization capabilities. On the EMODB and BodEMODB datasets, training accuracy approaches near-perfect levels, but testing accuracy, after an initial rise, either plateaus or fluctuates, indicating potential overfitting. In contrast, on the SAVEE, CASIA, and IEMOCAP datasets, a significant gap between training and testing accuracy emerges after a certain number of epochs, with testing accuracy plateauing around 0.7–0.75, further suggesting overfitting. The ESD dataset, however, exhibits a different trend; training accuracy nears 1.0, with testing accuracy closely following, indicative of strong generalization and minimal overfitting.

Table 1 and Fig. 3 provides a comprehensive comparison of the proposed CENN model across various datasets, emphasizing accuracy, precision, recall, and F1-score. Remarkably, the CENN model exhibits robust stability in all metrics on the ESD and BodEMODB datasets. This stability is attributed to the model's consistent and accurate predictions, showcasing its adaptability and reliability. Notably, the ESD dataset stands out as the one where the CENN model achieves its best performance.

The CENN model showcases superior accuracy, precision, recall, and F1-score on the EMODB dataset. Notably, it exhibits relatively lower stability in this context, suggesting that the model's performance on the EMODB dataset might be sensitive to specific variations or instances within the dataset.

Conversely, the CENN model faces challenges in maintaining optimal performance on the IEMOCAP and SAVEE datasets, reflecting its lowest performance among the analyzed datasets. This may be attributed to inherent complexities and imbalances within these datasets, posing difficulties for the model in generalization. These datasets exhibit higher variability in recording conditions, speaker characteristics, emotional expressions, and encompass a spectrum of acoustic variations, including differences in speech rate, intonation, accent, and background noise.

The confusion matrices in Fig. 4 depict the performance of the proposed CENN model across various datasets. The model excels in accurately distinguishing between different emotion types within the BodEMODB and ESD datasets. However, challenges arise in effectively discerning emotional types on the SAVEE and IEMOCAP datasets.

The confusion matrices in Fig. 4 depict the performance of the proposed CENN model across various datasets. The model excels in accurately distinguishing between different emotion types within the BodEMODB and ESD datasets. However, challenges arise in effectively discerning emotional types on the SAVEE and IEMOCAP datasets.

On the EMODB dataset, the model accurately identifies Anger (A), Boredom (B), Fear (F), and Sadness (Sa) but confuses Disgust (D) with Happiness (H) and Neutral (N) with Sadness (Sa). The SAVEE dataset shows significant confusion due to its small size and similar valence-arousal values, while the CASIA dataset sees high accuracy for some emotions with moderate difficulty in distinguishing others.

In datasets like IEMOCAP, imbalanced emotional instances lead to considerable confusion, particularly for the emotion Happiness/H. However, the model demonstrates excellent performance on the well-balanced ESD dataset, indicating its robust capability for accurate emotion recognition when data is sufficient.

The confusion matrices underscore the strengths and weaknesses of the CENN model across different datasets, emphasizing its effectiveness in well-balanced and sufficiently large datasets while identifying areas for improvement in smaller or imbalanced datasets.

Table 1
Results of the proposed CENN model on the different datasets: EMODB, SAVEE, CASIA, BodEMODB, IEMOCAP, and ESD.

Dataset	Metric	Avg ± Std	**Dataset**	Metric	Avg ± Std	**Dataset**	Metric	Avg ± Std
EMODB	Accuracy	0.9352 ± 0.0190	CASIA	Accuracy	0.8867 ± 0.0113	IEMOCAP	Accuracy	0.7200 ± 0.0094
	Precision	0.9061 ± 0.0172		Precision	0.8592 ± 0.0166		Precision	0.7297 ± 0.0045
	Recall	0.8944 ± 0.0249		Recall	0.8383 ± 0.0130		Recall	0.6759 ± 0.0165
	F1_score	0.9002 ± 0.0205		F1_score	0.8486 ± 0.0138		F1_score	0.7017 ± 0.0101
SAVEE	Accuracy	0.8292 ± 0.0204	BodEmoDB	Accuracy	0.9307 ± 0.0061	ESD	Accuracy	0.9683 ± 0.0016
	Precision	0.7962 ± 0.0138		Precision	0.8964 ± 0.0043		Precision	0.9601 ± 0.0056
	Recall	0.7813 ± 0.0192		Recall	0.8917 ± 0.0050		Recall	0.9616 ± 0.0082
	F1_score	0.7886 ± 0.0153		F1_score	0.8941 ± 0.0044		F1_score	0.9608 ± 0.0054

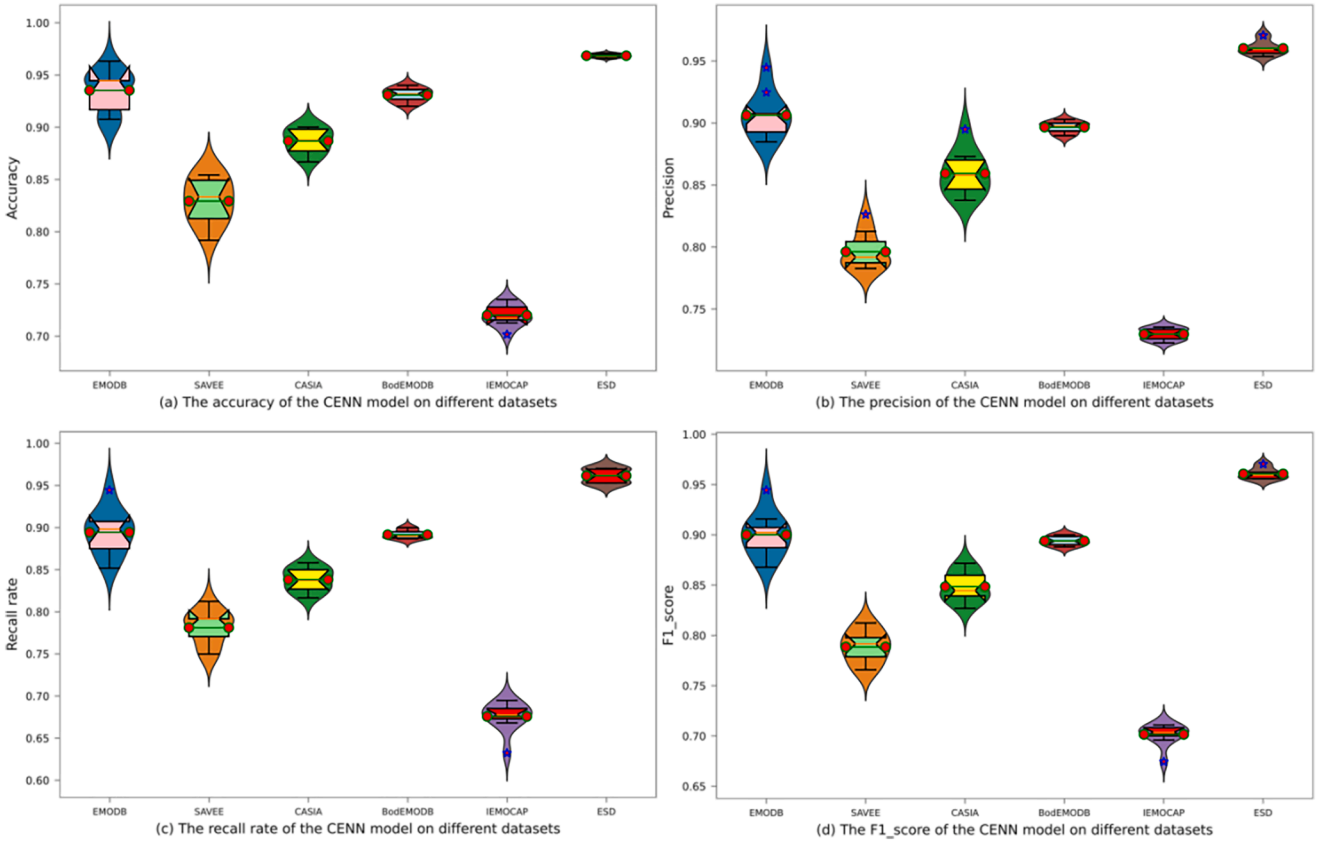


Fig. 3. The comparisons of the proposed CENN model across diverse datasets: EMODB, SAVEE, CASIA, BodEMODB, IEMOCAP, and ESD. It achieves the highest performance on the ESD, followed closely by EMODB, though with slightly lower stability. The model performs least favorably on the IEMOCAP, followed by SAVEE.

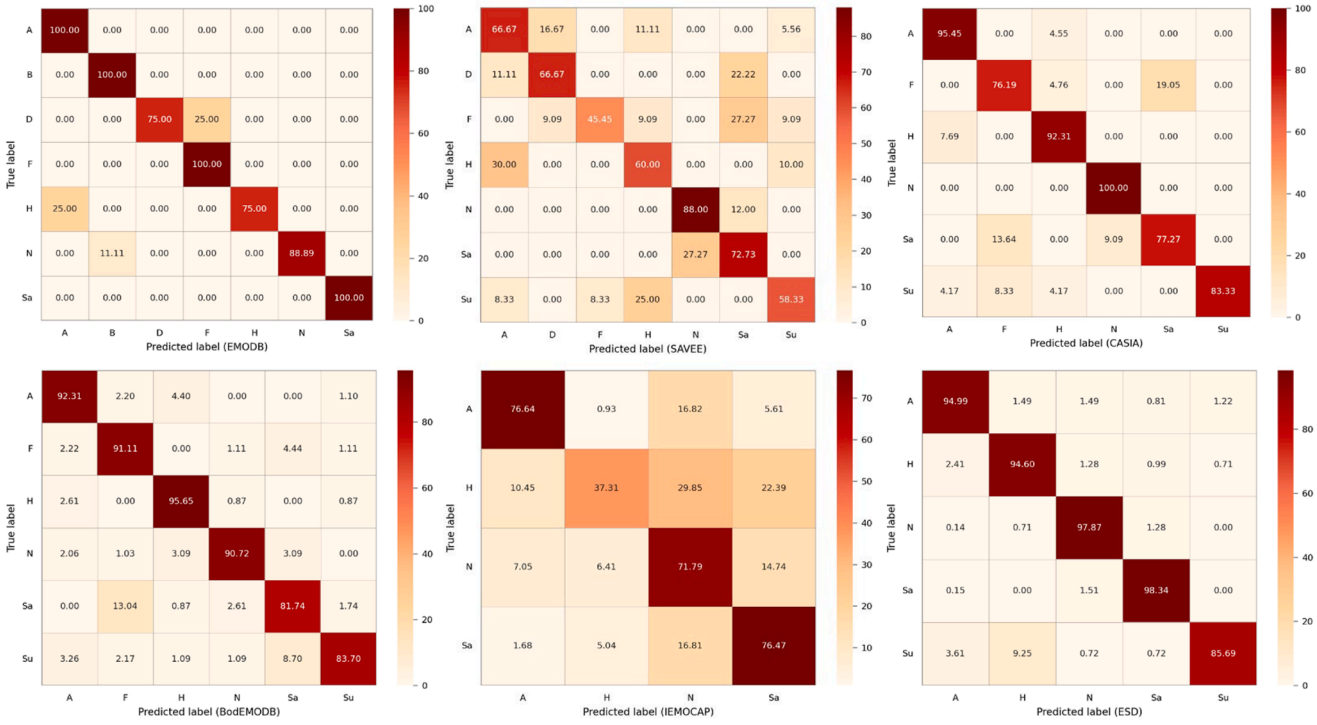


Fig. 4. The confusion matrices of the proposed CENN model across EMODB, SAVEE, CASIA, BodEMODB, IEMOCAP, and ESD datasets. The results highlight CENN's proficiency in accurately identifying various emotion types within the BodEMODB and ESD. However, challenges emerge in effectively distinguishing emotional types within the SAVEE and IEMOCAP, likely due to the small size of SAVEE and the imbalanced distribution of emotional instances in IEMOCAP.

The performance metrics—including accuracy, precision, recall, and F1-score—of the proposed CENN model were meticulously assessed across various datasets using different peer models: LSTM [40], GRU [41], CNN [42], Transformer [43], Autoencoder [44], LSM [45], TCN [46], CPAC [47], and TIM-Net [48], as depicted in Fig. 5. Each model's performance metrics are visually represented by distinct colored bars, as delineated in the legend. Remarkably, the accuracy of the CENN model consistently surpasses that of its peer models across all datasets, while its recall and F1-score demonstrate consistently high values for most datasets.

CENN consistently outperforms across various datasets. It excels on EMODB, SAVEE, CASIA, and BodEMODB, with only a few close competitors. On IEMOCAP, CENN maintains its lead, though the margin is smaller. Finally, on ESD, CENN's performance is unmatched, significantly surpassing all other models.

Consistent top-ranking or near-top-ranking across all datasets underscores CENN's robust performance. While traditional models such as LSTM [40] and GRU [41] exhibit competitive accuracy, they are not generally outperformed by the advanced architecture of CENN. Transformer-based models [43], renowned for capturing long-range dependencies, also demonstrate commendable accuracy but fall short of CENN's capabilities, likely due to CENN's enhanced feature extraction through capsules.

The superior performance of CENN across datasets like EMODB, CASIA, and ESD underscores its adaptability and effectiveness across varied scenarios. Even in datasets presenting challenges, such as the small size of SAVEE and the imbalanced distribution of IEMOCAP, CENN outperforms other models, reflecting its adept handling of complexities.

However, it is essential to note a discrepancy in precision, which tends to be relatively low across most datasets. This trade-off between recall and precision suggests that while CENN excels in correctly identifying instances of interest, it may exhibit slightly more false positive predictions. Despite the lower precision observed on certain datasets like SAVEE and CASIA, CENN consistently outperforms all other models in terms of F1-score, indicating a balanced performance between precision and recall. The use of MHA contributes significantly to CENN's exceptional performance, enabling the model to effectively capture and integrate diverse features and dependencies within input data. This approach enhances the model's ability to understand complex relationships, thus surpassing peer methods across various datasets.

5.2. The reproducibility of the CENN model

To assess reproducibility, we conducted experiments with different repetition counts (*Count* = 5, 6, 7, 8, 9, 10), examining instances where predicted labels matched true labels at least five times out of ten, each spanning 500 epochs. This analysis provides insights into the stability and consistency of CENN's predictions under varying conditions, serving as a foundation for evaluating the model's robustness.

Table 2 shows the GLR and CLR of the proposed CENN model across different datasets, including the EMODB, SAVEE, CASIA, BodEMODB, IEMOCAP, and ESD. Among these, FLR and FCLR encapsulate results from *Count*_5 to *Count*_9, while PLR and PCLR represent results from *Count*_10.

The CENN demonstrates high GLR and CLR across all datasets with initial repetition counts (*Count*_5). Both GLR and CLR decline with increasing repetition, reflecting the model's learning complexity. The high PLR and PCLR (*Count*_10) values across datasets (e.g., EMODB: 88.89 %, 85.19 %; SAVEE: 70.83 %, 66.67 %; CASIA: 81.67 %, 74.17 %; BodEMODB: 85.33 %, 82.00 %; IEMOCAP: 66.82 %, 51.89 %; ESD: 92.66 %, 91.44 %) indicate excellent reproducibility and performance, underscoring the CENN model's robustness and effectiveness in diverse

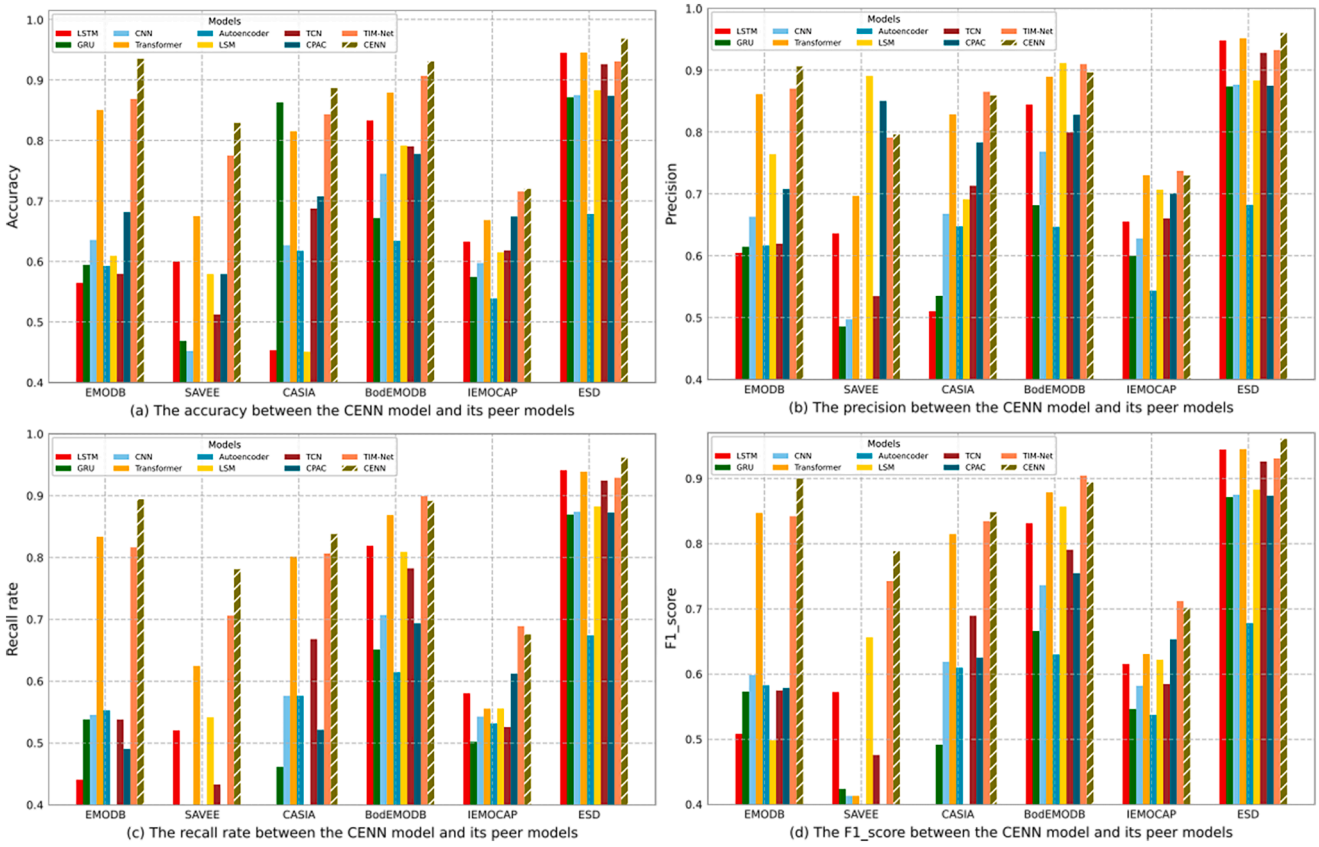


Fig. 5. Performance of the proposed CENN model across the EMODB, SAVEE, CASIA, BodEMODB, IEMOCAP, and ESD datasets. CENN consistently outperforms peers in accuracy, recall, and F1-score, despite relatively low precision, indicating a precision-recall trade-off. This underscores CENN's robustness, generalization, and effectiveness.

Table 2
General learning reproducibility (GLR) and Correct learning reproducibility (CLR) of the proposed CENN model, utilizing 10 times experimentation across diverse datasets. As reproducibility iterations increase, both GLR and CLR gradually decline, primarily due to the heightened learning difficulty within the CENN model.

Dataset	Reproducibility	Count_5	Count_6	Count_7	Count_8	Count_9	Count_10
EMODB	GLR	1.0000	0.9815	0.9630	0.9259	0.9259	0.8889
	CLR	0.9074	0.9074	0.9074	0.8704	0.8704	0.8519
SAVEE	GLR	0.9792	0.9583	0.8750	0.8542	0.7083	0.7083
	CLR	0.7292	0.7292	0.7083	0.7083	0.6667	0.6667
CASIA	GLR	1.0000	1.0000	0.9667	0.9000	0.8583	0.8167
	CLR	0.8333	0.8333	0.8333	0.7917	0.7750	0.7417
BodEMODB	GLR	0.9967	0.9833	0.9533	0.9367	0.9133	0.8533
	CLR	0.9033	0.9033	0.8933	0.8900	0.8767	0.8200
IEMOCAP	GLR	0.9866	0.9465	0.8976	0.8241	0.7439	0.6682
	CLR	0.6882	0.6637	0.6370	0.6102	0.5657	0.5189
ESD	GLR	0.9994	0.9923	0.9786	0.9649	0.9474	0.9266
	CLR	0.9673	0.9540	0.9457	0.9280	0.9159	0.9144

SER tasks.

Different datasets impact the model's performance variability. For instance, SAVEE and IEMOCAP show more significant declines, suggesting these datasets may pose unique challenges or have greater variability in data quality or characteristics. Despite some declines, the overall high GLR and CLR values, particularly in datasets like EMODB, CASIA, BodEMODB, and ESD, highlight the CENN model's robustness and its capacity to maintain reproducibility across different datasets.

Fig. 6 illustrates a comparative analysis of PLR and PCLR across various models including CENN, LSTM, GRU, CNN, Transformer, Autoencoder, TCN, TIM-Net, LSM, and CPAC. The CENN model, represented by the striped bars, consistently exhibits high PLR and PCLR values across all datasets compared to other models. This indicates its high reliability in accurately reproducing learning results. The consistently high PLR and PCLR values across different datasets suggest that CENN generalizes well and is robust against varied data distributions and characteristics. Its architecture incorporates mechanisms to capture intricate patterns and ensure stable learning, enhancing reproducibility. Additionally, CENN's design effectively handles noise and variability, making it more reliable across diverse datasets.

GRU shows better reproducibility than LSTM but still falls short of CENN, especially on CASIA and SAVEE. CNN performs well on BodEMODB and ESD but are weaker on SAVEE. Transformer models are strong on EMODB and ESD but inconsistent overall. Autoencoders perform adequately but don't match CENN's reproducibility. TCNs have mixed results, excelling on some datasets but underperforming on others. TIM-Net shows promise in specific cases but lacks CENN's robustness. CPAC is less consistent than CENN, while LSM's reproducibility varies significantly across datasets.

5.3. Comparison with previous research findings

Table 3 presents a comparative analysis of emotion recognition models across six benchmark datasets, highlighting the superior performance of the proposed CENN model. The CENN model consistently outperforms other models on most datasets, except for SAVEE and CASIA. Specifically, CENN achieves the highest accuracy on EMODB (96.30 %), IEMOCAP (72.88 %), BodEMODB (93.07 %), and ESD (96.83 %). TIM-Net [48], also introduced in 2023, demonstrates strong performance on EMODB (95.70 %), SAVEE (86.07 %), and CASIA (94.67 %). The GM-TCN model performs well on EMODB (91.39 %) and CASIA (90.17 %) [51], while CPAC achieves high accuracy on EMODB (94.95 %) and CASIA (92.75 %) [47]. Although models like TSP+INCA [49], 3D CNN [52], and DT-SVM [53] perform decently, they generally fall short of the accuracy achieved by the proposed CENN model.

The top-performing models for the EMODB dataset are CENN (96.30 %), TIM-Net (95.70 %), and CPAC (94.95 %), indicating their effectiveness in handling its features. On the SAVEE dataset, TIM-Net (86.07 %) and CPAC (83.69 %) outperform CENN (85.42 %), suggesting these models better address its specific challenges. Similarly, for the CASIA dataset, TIM-Net (94.67 %) and CPAC (92.75 %) surpass CENN (90.00 %). CENN leads the IEMOCAP dataset with 72.88 %, closely followed by TIM-Net at 72.50 %, showcasing their advanced capabilities. On the BodEMODB dataset, CENN (93.07 %) and Transformer (85.00 %) achieve the highest accuracies, highlighting their strength. For the ESD dataset, CENN (96.83 %) and Transformer (93.80 %) again lead, demonstrating their robustness and adaptability.

The newer models, such as CENN, TIM-Net, and CPAC, leverage advanced neural network architectures, contributing to their superior performance. For example, the capsule network in CENN effectively

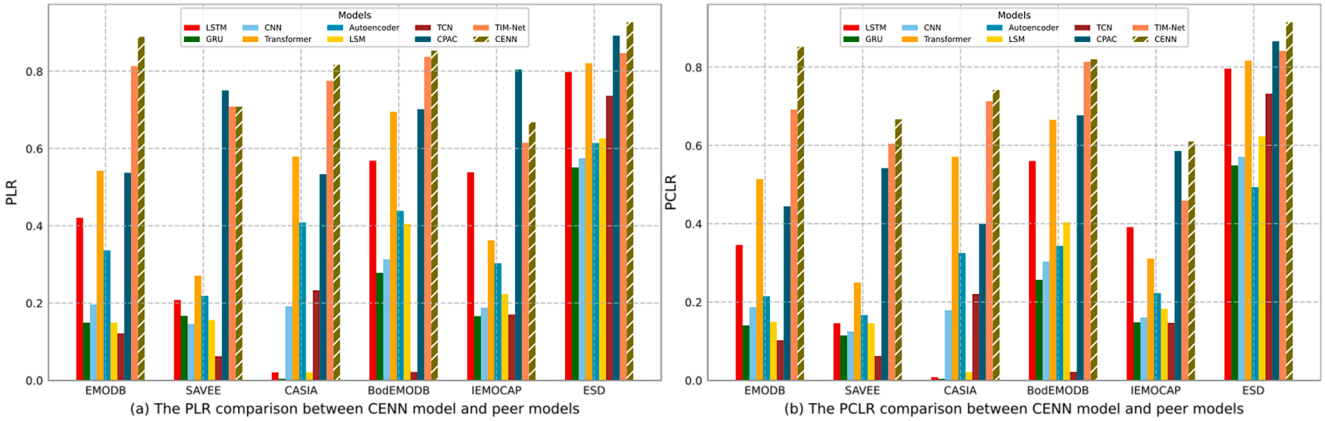


Fig. 6. The comparisons of Perfect learning reproducibility (PLR) and Perfect correct learning reproducibility (PCLR) for the proposed CENN model in relation to its peer models across six benchmark datasets: EMODB, SAVEE, CASIA, BodEMODB, IEMOCAP, and ESD. The consistently higher PCLR achieved by CENN, when contrasted with its peers, serves to accentuate its superior correctness in terms of reproducibility.

Table 3
The accuracy comparison between the proposed Capsule-enhanced neural network (CENN) model and previous models from previous research across benchmark datasets. Notably, the CENN model consistently outperforms previous research findings on most datasets, except for the SAVEE and CASIA datasets.

Model	Year	EMODB	Model	Year	SAVEE	Model	Year	CASIA
TSP+INCA [49]	2021	90.09	3D CNN [52]	2019	81.05	DT-SVM [53]	2019	85.08
QCNN [50]	2021	88.78	TSP+INCA [49]	2021	83.38	TLFMRF [54]	2020	85.83
GM-TCN [51]	2022	91.39	CPAC [47]	2022	83.69	GM-TCN [51]	2022	90.17
CPAC [47]	2022	94.95	GM-TCN [51]	2022	83.88	CPAC [47]	2022	92.75
TIM-Net [48]	2023	95.70	TIM-Net [48]	2023	86.07	TIM-Net [48]	2023	94.67
CENN (Ours)	2024	96.30	CENN (Ours)	2024	85.42	CENN (Ours)	2024	90.00
Model	Year	BodEMODB	Model	Year	IEMOCAP	Model	Year	ESD
LSTM	2023	82.33	MHA+DRN [55]	2019	67.40	LSTM	2023	93.61
GRU	2023	77.17	CNN+BiGRU [56]	2020	71.72	GRU	2023	89.04
CNN	2023	74.17	QCNN [50]	2021	70.46	CNN	2023	88.74
Transformer	2023	85.00	Light-SERNet [57]	2022	70.76	Transformer	2023	93.80
TCN	2023	78.17	TIM-Net [48]	2023	72.50	TCN	2023	92.76
CENN	2024	93.07	CENN (Ours)	2024	72.88	CENN	2024	96.83

capture spatial hierarchies in data. Additionally, certain datasets have unique features that align better with specific model architectures. For instance, SAVEE and CASIA may have characteristics that TIM-Net and CPAC exploit more effectively than CENN.

The proposed CENN model demonstrates outstanding performance across most datasets, highlighting the benefits of capsule network in SER tasks. However, TIM-Net and CPAC also show strong results, particularly on datasets where CENN does not excel. This indicates that model performance can be highly dependent on the specific characteristics of the dataset.

6. Discussion

The CENN represents a novel and effective approach to SER, leveraging a hybrid architecture that combines state-of-the-art deep learning techniques. The model's strengths are evident in its robust performance, enhanced by innovative metrics designed to address common challenges in deep learning, such as overfitting and generalization. The CENN model's reproducibility framework further solidifies its contribution to the field, offering a reliable tool for SER research and applications.

A key strength of the CENN model lies in its ability to maintain high performance across diverse datasets, demonstrating its generalization capabilities. The reproducibility-focused approach adopted in this study has also contributed to the reliability of the model's results, making it a robust tool for SER. Moreover, the inclusion of a detailed analysis of model complexity and data complexity has highlighted the importance of balancing these factors to prevent overfitting and enhance generalization.

Despite these advancements, the CENN model is not without limitations. The computational complexity associated with its architecture presents challenges for real-time applications, particularly in environments with limited processing power. Additionally, while the model has shown strong performance on benchmark datasets, its applicability to more diverse and noisy real-world datasets requires further exploration. The interpretability of the model's decisions remains another area for improvement, as understanding how the model arrives at its predictions is crucial for building trust and usability in practical applications.

7. Conclusion

The development and evaluation of the CENN mark a significant advancement in the field of SER. The integration of cutting-edge components, such as MHA, ResNet module, and Capsule module, has demonstrated substantial effectiveness in capturing both global and local patterns within speech data. This has led to notable improvements in the accuracy and robustness of emotion recognition models.

Our comprehensive evaluation across six benchmark datasets—EMODB, CASIA, SAVEE, BodEMODB, IEMOCAP, and ESD—consistently shows that the CENN model outperforms existing SER models in terms of both accuracy and reproducibility. The introduction of novel metrics, including overfitting, generalization, and robustness metrics, has provided a multidimensional framework for assessing the model's performance, ensuring its applicability to real-world scenarios.

Future research should focus on optimizing the CENN model for real-time applications, enhancing its computational efficiency without compromising performance. Expanding the model's capabilities to handle more diverse, multimodal datasets will also be critical in broadening its applicability. Moreover, improving the interpretability of the CENN model's decision-making process will be essential for fostering trust and facilitating its adoption in practical, high-stakes settings.

Beyond the domain of SER, the versatile architecture of the CENN model holds significant promise for other complex pattern recognition tasks, such as medical diagnostics and behavioral analysis. By addressing these avenues for further improvement, the CENN model is well-positioned to make substantial contributions not only to SER but also to a wide range of fields that require advanced and reliable data analysis.

CRediT authorship contribution statement

Huiyun Zhang: Writing – review & editing, Writing – original draft, Software, Methodology, Conceptualization. **Heming Huang:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Puyang Zhao:** Conceptualization, Methodology, Writing – review & editing. **Xiaojun Zhu:** Conceptualization, Methodology, Software, Writing – original draft. **Zhenbao Yu:** Conceptualization, Formal analysis, Methodology, Writing – original draft.

Declaration of competing interest

The authors declare that there is no conflict of interests regarding the publication of this paper and research funds.

Data availability

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

Acknowledgments

The authors acknowledge the Natural Science Foundation of Qinghai Province (Grant: 2022-ZJ-925) and the National Natural Science Foundation of China (Grant: 62066039).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.knosys.2024.112499.

References

[1] 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.

[2] L. Yi, M.W. Mak, Improving speech emotion recognition with adversarial data augmentation network, IEEE Trans. Neural Netw Learn Syst. 33 (1) (2022) 172–1844.

[3] Y. Lei, H. Cao, Audio-visual emotion recognition with preference learning based on intended and multi-modal perceived labels, IEEE Trans. Affect. Comput. 14 (4) (2023) 2954–2969.

[4] K. Wang, N. An, B.N. Li, Y. Zhang, L. Li, Speech emotion recognition using Fourier parameters, IEEE Trans. Affect. Comput. 6 (1) (2015) 69–75.

[5] Z. Luo, S. Lin, R. Liu, J. Baba, Y. Yoshikawa, H. Ishiguro, Decoupling speaker-independent emotions for voice conversion via source-filter networks, IEEE/ACM Trans. Audio Speech Lang Process 31 (2023) 11–24.

[6] S. Leem, D. Fulford, J. Onnela, D. Gard, C. Busso, Not all features are equal: selection of robust features for speech emotion recognition in noisy environments, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 6447–6451.

[7] W. Lin, C. Busso, Chunk-level speech emotion recognition: a general framework of sequence-to-one dynamic temporal modeling, IEEE Trans. Affect. Comput. 14 (2) (2023) 1215–1227.

[8] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, B. Schuller, Survey of deep representation learning for speech emotion recognition, IEEE Trans. Affect. Comput. 14 (2) (2023) 1634–1654.

[9] X. Lan, X. Li, Y. Ning, Z. Wu, H. Meng, J. Jia, L. Cai, Low level descriptors based DBLSTM bottleneck feature for speech driven talking avatar, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 2016, pp. 5550–5554.

[10] L. Vu, R. Phan, L. Han, D. Phung, Improved speech emotion recognition based on music-related audio features, in: European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, 2022, pp. 120–124.

[11] C. Zheng, N. Jia, W. Sun, The extraction method of emotional feature based on children's spoken speech, in: International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Hangzhou, China, 2019, pp. 165–168.

[12] S. Ananthakrishnan, A. Vembu, R. Prasad, Model-based parametric features for emotion recognition from speech, in: IEEE Workshop on Automatic Speech Recognition & Understanding, Waikoloa, HI, USA, 2011, pp. 529–534.

[13] S. Ntalampiras, N. Fakotakis, Modeling the temporal evolution of acoustic parameters for speech emotion recognition, IEEE Trans. Affect. Comput. 3 (1) (2012) 116–125.

[14] Y. Gao, S. Okada, L. Wang, J. Liu, J. Dang, Domain-invariant feature learning for cross corpus speech emotion recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 6427–6431.

[15] S. Li, P. Song, W. Zheng, Multi-source discriminant subspace alignment for cross-domain speech emotion recognition, IEEE/ACM Trans. Audio Speech Lang Process 31 (2023) 2448–2460.

[16] M. Wu, W. Su, L. Chen, W. Pedrycz, K. Hirota, Two-stage fuzzy fusion based-convolution neural network for dynamic emotion recognition, IEEE Trans. Affect. Comput. 13 (2) (2022) 805–817.

[17] Y. Cao, A. Chandrasekar, T. Radhika, V. Vijayakumar, Input-to-state stability of stochastic Markovian jump genetic regulatory networks, Math. Comput. Simul 222 (2024) 174–187.

[18] T. Radhika, A. Chandrasekar, V. Vijayakumar, et al., Analysis of Markovian jump stochastic Cohen–Grossberg BAM neural networks with time delays for exponential input-to-state stability, Neural Processing Letters 55 (2023) 11055–11072.

[19] V. Tran, T. Nguyen, H. Nguyen-Xuan, et al., Vibration and buckling optimization of functionally graded porous microplates using BCMO-ANN algorithm, Thin Walled Struct. 182 (2023) 110267. ISSN 0263-8231.

[20] X.Jia M.Ping, C. Papadimitriou, et al., A hierarchical Bayesian modeling framework for identification of Non-Gaussian processes, Mech. Syst. Signal. Process 208 (2024) 110968.

[21] B. Dang, H. Nguyen-Xuan, M. Wahab, An effective approach for VARANS-VOF modelling interactions of wave and perforated breakwater using gradient boosting decision tree algorithm, Ocean Eng. 268 (2023) 113398.

[22] D. Nguyen, M. Wahab, Damage detection in slab structures based on two-dimensional curvature mode shape method and Faster R-CNN, Adv. Eng. Software 176 (2023) 103371.

[23] S. Wang, H. Wang, Y. Zhou, et al., Automatic laser profile recognition and fast tracking for structured light measurement using deep learning and template matching, Measurement 169 (2021) 108362.

[24] T.Babu N.Thendral, A. Chandrasekar, Synchronization of Markovian jump neural networks for sampled data control systems with additive delay components: analysis of image encryption technique, Math. Methods Appl. Sci. (2022).

[25] 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.

[26] Y. Wang, C. Lu, H. Lian, Y. Zhao, B. Schuller, Y. Zong, W. Zheng, Speech Swin-Transformer: exploring a hierarchical Transformer with shifted windows for speech emotion recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, 2024, pp. 11646–11650.

[27] Z. Liu, X. Kang, F. Ren, Dual-TBNet: improving the robustness of speech features via dual-Transformer-BiLSTM for speech emotion recognition, IEEE/ACM Trans Audio Speech Lang Process 31 (2023) 2193–2203.

[28] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, Dawn of the transformer era in speech emotion recognition: closing the valence gap, IEEE Trans. Pattern Anal. Mach. Intell. 45 (9) (2023) 10745–10759.

[29] X. Li, Z. Zhang, C. Gan, Y. Xiang, Multi-label speech emotion recognition via inter-class difference loss under response residual network, IEEE Trans. Multimedia 25 (2023) 3230–3244.

[30] Y. Guo, L. Chen, Y. Chen, On connections between regularizations for improving DNN robustness, IEEE Trans. Pattern Anal. Mach. Intell. 43 (12) (2021) 4469–4476.

[31] J. Hsu, M. Su, C. Wu, Y. Chen, Speech emotion recognition considering nonverbal vocalization in affective conversations, IEEE/ACM Trans Audio Speech Lang Process 29 (2021) 1675–1686.

[32] X. Wu, Y. Cao, H. Lu, S. Liu, D. Wang, Z. Wu, X. Liu, H. Meng, Speech emotion recognition using sequential capsule networks, IEEE/ACM Trans Audio Speech Lang Process 29 (2021) 3280–3291.

[33] C. Fan, J. Wang, W. Huang, X. Yang, G. Pei, T. Li, Light-weight residual convolution-based capsule network for EEG emotion recognition, Adv. Eng. Inf. 61 (2024) 102522.

[34] S. Gawusu, A. Ahmed, Analyzing variability in urban energy poverty: a stochastic modeling and Monte Carlo simulation approach, Energy 304 (2024) 132194.

[35] C. Harvey, A. Noheria, Deep learning encoded EGG-Avoiding overfitting in EGG machining learning, J. Am. Coll. Cardiol. 83 (13) (2024).

[36] D. Wang, Some further thoughts about spectral kurtosis, spectral L2/L1 norm, spectral smoothness index and spectral Gini index for characterizing repetitive transients, Mech. Syst. Signal. Process 108 (2018) 360–368.

[37] R. Silva, O. Freitas, P. Melo-Pinto, Evaluating the generalization ability of deep learning models: an application on sugar content estimation from hyperspectral images of wine grape berries, Expert Syst. Appl. 250 (2024) 123891.

[38] Y. Zhu, H. Peng, A. Fu, W. Yang, H. Ma, Towards robustness evaluation of backdoor defense on quantized deep learning models, Expert Syst. Appl. 255 (2024) 124599.

[39] K. Zhou, B. Sisman, R. Liu, H. Li, Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 920–924.

[40] S. Mirsamadi, E. Barsoum, C. Zhang, Automatic speech emotion recognition using recurrent neural networks with local attention, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 2017, pp. 2227–2231.

[41] S.T. Rajamani, K.T. Rajamani, A. Mallol-Ragolta, S. Liu, B. Schuller, A novel attention-based gated recurrent unit and its efficacy in speech emotion recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 6294–6298.

[42] Z. Peng, Y. Lu, S. Pan, Y. Liu, Efficient speech emotion recognition using multi-scale CNN and attention, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 3020–3024.

[43] X. Wang, M. Wang, W. Qi, W. Su, X. Wang, H. Zhou, A novel end-to-end speech emotion recognition network with stacked Transformer layers, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 6289–6293.

[44] Y. Gao, J. Liu, L. Wang, J. Dang, Domain-adversarial Autoencoder with attention based feature level fusion for speech emotion recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 6314–6318.

[45] R. Lotfidereshgi, P. Gournay, Biologically inspired speech emotion recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 2017, pp. 5135–5139.

[46] Z. He, Y. Zhong, J. Pan, Joint temporal convolutional networks and adversarial discriminative domain adaptation for EEG-based cross-subject emotion recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 3214–3218.

[47] X. Wen, J. Ye and K. Liu. CTL-MTNet: a novel CapsNet and transfer learning-based mixed task net for the single-corpus and cross-corpus speech emotion recognition. arXiv preprint arXiv:2207.10644 (2022).

[48] J. Ye, X.C. Wen, Y. Wei, Y. Xu, K. Liu, Temporal modeling matters: a novel temporal emotional modeling approach for speech emotion recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1–5.

[49] T. Tuncer, S. Dogan, U. Acharya, Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques, Knowl. Based. Syst. 211 (2021) 106547.

[50] A. Muppidi, M. Radfar, Speech emotion recognition using Quaternion convolutional neural networks, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 6309–6313.

[51] J. Ye, X. Wen, X. Wang, Y. Xu, Y. Luo, C. Wu, L. Chen, K. Liu, GM-TCNet: gated multi-scale temporal convolutional network using emotion causality for speech emotion recognition, Speech Commun 145 (2022) 21–35.

[52] N. Hajarolasvadi, H. Demirel, 3D CNN-based speech emotion recognition using K-means clustering and spectrograms, Entropy 21 (5) (2019) 479.

[53] L. Sun, S. Fu, F. Wang, Decision tree SVM model with Fisher feature selection for speech emotion recognition, EURASIP Journal on Audio, Speech, and Music Processing 2019 (2019) 2.
[54] L. Chen, W. Su, Y. Feng, M. Wu, J. She, K. Hirota, Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction, Inf Sci (Ny) 509 (2020) 150–163.
[55] R. Li, Z. Wu, J. Jia, S. Zhao, H. Meng, Dilated residual network with multi-head self-attention for speech emotion recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 6675–6679.
[56] Y. Zhong, Y. Hu, H. Huang, W. Silamu, A lightweight model based on separable convolution for speech emotion recognition, Interspeech, Shanghai, China (2020) 3331–3335.
[57] A. Aftab, A. Morsali, S. Ghaemmaghami, B. Champagne, LIGHT-SERNET: a lightweight fully convolutional neural network for speech emotion recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 6912–6916.