ICT Express 11 (2025) 657–665

Contents lists available at ScienceDirect

ICT Express

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

TMNet: Transformer-fused multimodal framework for emotion recognition via EEG and speech

Md Mahinur Alam [a], Mohamed A. Dini [a], Dong-Seong Kim [a], Taesoo Jun [b,*]

[a] Department of IT Convergence Engineering, Kumoh National Institute of Technology, Gumi 39177, South Korea
[b] Department of Computer Software Engineering, Kumoh National Institute of Technology, Gumi 39177, South Korea

ARTICLE INFO

Keywords:
BiLSTM
BiGRU
EEG signal
Emotion recognition
Multimodal
Speech recognition
Transformer

ABSTRACT

In the evolving field of emotion recognition, which intersects psychology, human–computer interaction, and social robotics, there is a growing demand for more advanced and accurate frameworks. The traditional reliance on single-modal approaches has given way to a focus on multimodal emotion recognition, which offers enhanced performance by integrating multiple data sources. This paper introduces TMNet, an innovative multimodal emotion recognition framework that leverages both speech and Electroencephalography (EEG) signals to deliver superior accuracy. This framework utilizes cutting-edge technology, employing a Transformer model to effectively fuse the CNN-BiLSTM and BiGRU architectures, creating a unified multimodal representation for enhanced emotion recognition performance. By utilizing a diverse set of datasets RAVDESS, SAVEE, TESS, and CREMA-D for speech, along with EEG signals captured via the Muse headband. The multimodal model achieves impressive accuracies of 98.89% for speech and EEG signal processing.

1. Introduction

Understanding human emotions is key to effective human-machine interaction [1]. Emotion recognition has therefore become a research focus in psychology, neuroscience, and computer science. Emotion-aware systems can adapt responses to match users' emotional states, enhancing user experience. For instance, a framework could respond to frustration with reassurance or to delight with enthusiasm. Emotion recognition also allows users to receive feedback on their emotions, promoting self-awareness and personal growth. Earlier research has employed text, facial expressions, acoustics, and physiological signals for emotion detection. Each method has limitations, such as text analysis's challenges with context and sarcasm, facial analysis's issues with cultural and individual differences, and EEG's sensitivity to subtle brainwave variations and ethical concerns [2]. Speech analysis, which is more user-friendly, also requires addressing emotional nuance, language, and cultural diversity.

Speech emotion recognition (SER) has gained traction for scalable emotion detection, yet speech alone may not fully capture emotions. EEG has shown promise for identifying underlying neural signals of emotions, but combining EEG and speech can yield more robust results [3]. Multimodal frameworks that integrate EEG with speech outperform unimodal systems by providing richer emotional insights [4]. Despite advancements, EEG-speech integration still poses challenges in data synchronization, preprocessing, and balancing modality representation. EEG and audio require distinct approaches to handle issues like artifact sensitivity in EEG and relevant feature extraction in audio. Optimal algorithm selection is essential to leverage both modalities without redundancy for improved framework performance [5].

This work introduces a multimodal emotion recognition framework that combines EEG and speech signals to classify emotions accurately. Our approach aims to improve the accuracy and robustness of emotion recognition systems utilizing advanced deep learning technologies like bidirectional GRUs, LSTMs, and transformers, with potential applications in healthcare, education, and entertainment. The contributions of this work can be summarized as follows:

- This work introduces a multimodal emotion recognition framework using a CNN-BiLSTM model for speech and a BiGRU model for EEG signals. The CNN-BiLSTM identifies emotions from audio cues, while the BiGRU detects emotions from EEG patterns, creating a robust tool for enhanced emotion recognition in human–computer interactions.
- We design a Transformer-driven multimodal framework that fuses outputs from the speech and EEG-based emotion recognition models. The Transformer captures complex interrelationships between modalities, enhancing emotion recognition accuracy and resilience. This Transformer-fused approach outperforms

* Corresponding author.
E-mail address: taesoo.jun@kumoh.ac.kr (T. Jun).

https://doi.org/10.1016/j.icte.2025.04.007
Received 9 December 2024; Received in revised form 2 April 2025; Accepted 14 April 2025
Available online 20 May 2025

unimodal methods by effectively integrating rich features from both EEG and speech signals.
- Extensive evaluations on publicly available datasets (RAVDESS, SAVEE, TESS, and CREMA-D) demonstrate the effectiveness of our proposed framework, validating its superior performance in emotion recognition across diverse data sources.

The rest of the paper is organized as Section 2 related works, Section 3 delves into the framework design, Section 4 is for evaluation and discussion. Section 5 discusses the limitations and future work of the paper, and Section 6 provides a didactic conclusion.

2. Related works

Various methods, including facial expression analysis, voice, physiological signals, and behavioral observations, have been explored for emotion recognition. Speech and EEG signals stand out for their complementary features: speech conveys emotional cues through tone and pitch, while EEG offers direct insight into brain activity related to emotions [1]. In [2], a speech-emotion model aimed at helping children with Autism Spectrum Disorder (ASD) was developed, using the RAVDESS dataset. However, overfitting due to excessive variation remained a challenge. Another model, MLT-SER, used dilated CNN to capture emotional elements in speech but required more energy in real-time applications. In [6], a CNN-LSTM model trained on the RECOLA dataset demonstrated context-aware emotion recognition, though it suffered from dataset limitations and a simplistic CNN. In [1], a ResNet50-based mobile app for ASD displayed promising accuracy, though the training was lengthy due to model complexity. EEG is also effective for emotion recognition, as it captures unique patterns representing different emotional states [3]. EEG signals, particularly beneficial for ASD studies, reveal sharper wave discharges compared to neurotypical individuals [7]. EEG-based methods such as those in [8], which used genetic algorithms for feature extraction, and [9], which leveraged deep learning on raw EEG data with restricted Boltzmann machines, have shown improved accuracy. In [10], physiological data collected from wearable devices was analyzed with SVM to predict emotions in children with ASD, based on physiological markers like Galvanic Skin Response (GSR) and Heart Rate Variability (HRV). In advancing multimodal emotion recognition, combining EEG and speech has shown the potential to capture nuanced emotional states by integrating linguistic and physiological data [11]. This approach, however, demands careful synchronization and computational resources, and the ethical implications of EEG use require consideration. Despite challenges, bimodal recognition frameworks offer a promising pathway to deeper insights into emotions. Our work addresses gaps in prior studies by leveraging multiple well-established datasets and integrating speech and EEG signals through deep neural networks and transformers for enhanced emotion recognition. The proposed multimodal framework, TMNet, combines advanced AI models and extensive datasets to improve robustness and accuracy, providing a scalable solution that surpasses limitations in current unimodal or bimodal models.

3. Proposed framework overview

The proposed multimodal emotion recognition framework is designed to classify emotions accurately by processing EEG and speech signals simultaneously. The technical workflow, as shown in Fig. 1, integrates complementary emotional cues from both modalities through systematic data acquisition, preprocessing, feature extraction, and classification. EEG signals are recorded using headgear, and speech signals are acquired via a microphone. To overcome equipment constraints, we utilized open-source EEG datasets [12] alongside publicly available speech datasets. Synchronizing EEG and speech data presents significant challenges due to inherent differences in sampling rates, modality-specific temporal distortions, and variations in response timing leading to discrepancies in temporal resolution. Additionally, cognitive and vocal responses to stimuli may have different latencies, further complicating direct alignment. To address these challenges, a two-stage synchronization approach is implemented. First, timestamps from both modalities are aligned based on stimulus presentation times. Next, Dynamic Time Warping (DTW) [13] is applied to refine temporal alignment by compensating for non-linear variations in signal timing. This ensures precise synchronization before modality fusion. We choose DTW over other potential methods for temporal alignments, such as Connectionist Temporal Classification (CTC) and temporal convolutions, because CTC is often better suited for tasks where a direct, end-to-end mapping is needed, and is typically applied when sequence lengths may vary significantly without requiring the explicit synchronization of two different modalities and Temporal Convolutions can handle alignment more adaptively, it requires substantial computational resources and complex model training. The EEG signal processing pipeline involved several stages: acquisition, data preprocessing, feature extraction, and classification. For noise filtering a bandpass filter is applied to retain frequencies between −600 to 800 Hz. This process eliminated low-frequency drift and high-frequency noise:

$$\mathrm{EEG}_{\mathrm{filtered}}(t) = \mathrm{EEG}_{\mathrm{raw}}(t) * h(t) \tag{1}$$

where $h(t)$ is the impulse response of the bandpass filter. Independent Component Analysis (ICA) [14] is used to separate mixed EEG signals into independent components, removing artifacts such as muscle movements and eye blinks:

$$\mathrm{EEG}_{\mathrm{clean}}(t) = \mathrm{EEG}_{\mathrm{filtered}}(t) - \mathrm{Artifacts}(t) \tag{2}$$

Fig. 2 shows the artifact removal from the EEG signal of positive emotion. The cleaned signals are normalized to zero mean and unit variance:

$$\mathrm{EEG}_{\mathrm{norm}}(t) = \frac{\mathrm{EEG}_{\mathrm{clean}}(t) - \mu}{\sigma} \tag{3}$$

where μ and σ are the mean and standard deviation. The normalized signals are segmented into overlapping 1-second windows with a 50% overlap to capture temporal dynamics:

$$\mathrm{EEG}_{\mathrm{seg}}(t) = \mathrm{Segment}(\mathrm{EEG}_{\mathrm{norm}}(t), \\ \text{window size} = 1s, \text{overlap} = 50\%) \tag{4}$$

Frequency domain features are extracted, focusing on emotion-specific frequency bands: Neutral emotions: 5 Hz to 25 Hz, Negative emotions: −600 Hz to 600 Hz, Positive emotions: Above 600 Hz. Power Spectral Density (PSD) and numerical encoding are applied to prepare the features for input into a neural network. These features are then classified into Neutral, Positive, or Negative categories using a BiGRU model.

The speech signal processing pipeline followed a structured approach involving preprocessing, feature extraction, and classification. Noise is removed using Voice Activity Detection (VAD) to exclude irrelevant segments, and spectral subtraction is used for background noise reduction. The recordings are resampled to 22,050 Hz for uniformity. Feature Extraction: Mel-Frequency Cepstral Coefficients (MFCCs) [4] are extracted from 7,664 audio recordings, each labeled with emotion. Using Librosa's parameters with a frame length of 512 and a hop length of 512, 40 MFCCs are extracted per recording:

$$\mathrm{MFCCs} = \mathrm{Librosa}(F, \text{frame length} = 512, \\ \text{hop length} = 512 \tag{5}$$

The resulting feature matrices are standardized to a fixed length of 500 to ensure consistency. The standardized features are processed using a CNN-BiLSTM model, which classifies the emotional content into predefined categories based on the extracted acoustic features. DTW is employed to synchronize emotional cues across the modalities to minimize temporal misalignment. The alignment is computed as:

$$\mathrm{DTW\ Alignment}(S_{\mathrm{speech}}, S_{\mathrm{EEG}}) =$$

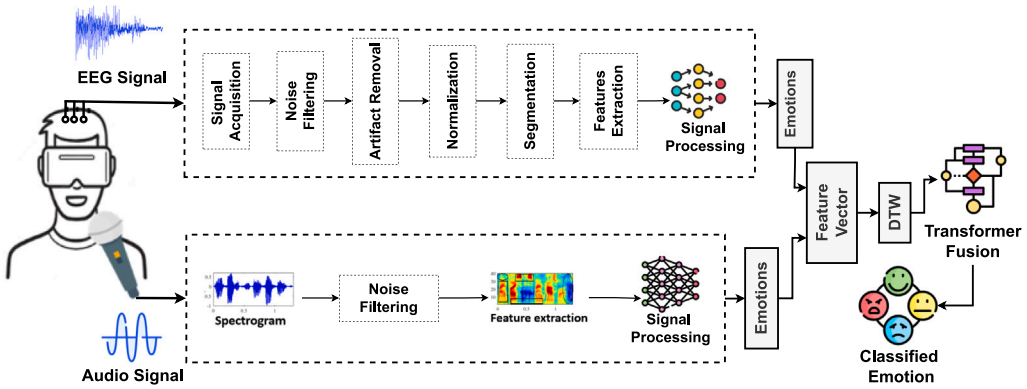

Fig. 1. Proposed workflow of the multimodal emotion recognition framework.

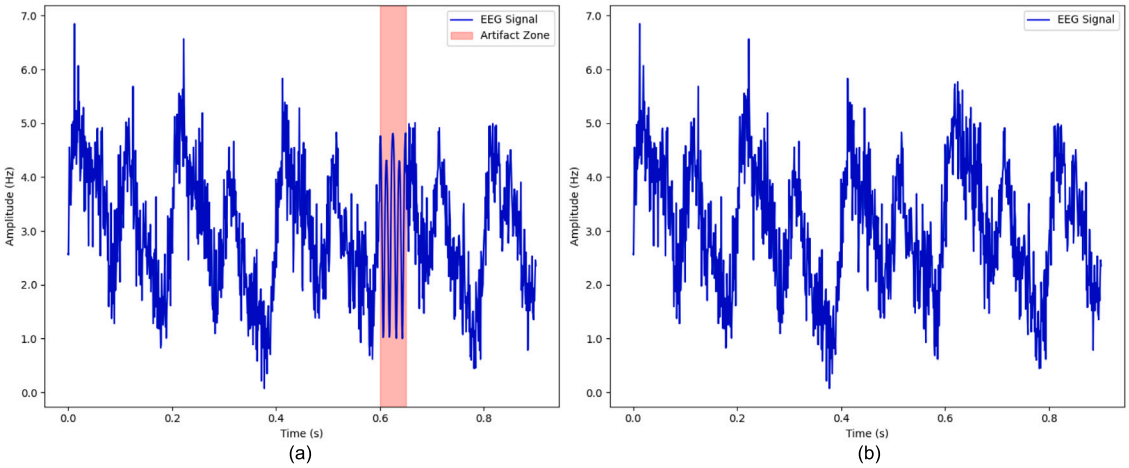

Fig. 2. Artifact removal from EEG signal of positive emotion (a) Original EEG signal with artifacts, (b) Artifact-free signal after ICA processing.

$$\arg\min_{\text{path}} \sum_{i=1}^{n} d(S_{\text{speech}}[i], S_{\text{EEG}}[i]) \tag{6}$$

where S_{speech} and S_{EEG} represent the feature sequences from the speech and EEG modalities, respectively, and $d(\cdot,\cdot)$ denotes the distance metric as Euclidean distance. DTW ensures that the emotional features from both modalities align in the temporal domain, even if the signals exhibit non-linear temporal differences. The output is integrated using a Transformer-based module, which acted as a meta-learner to combine insights from both modalities it performs well on unseen data due to the Transformer's self-attention mechanism, context-aware feature learning, and pre-training capabilities. These properties allow the model to capture long-range dependencies, dynamically weigh important features, and adapt to new variations in input distributions, enhancing its generalization to unseen data.

3.1. Dataset description

The datasets used include both speech-emotional and EEG data. For speech-emotion analysis, we selected four comprehensive datasets: *RAVDESS* [15], *TESS* [16], *SAVEE* [17], and *CREMA-D* [18]. These datasets cover a range of universal emotions in wav format. From these, we focused on five primary emotions: *happiness*, *anger*, *calmness*, *sadness*, and *neutral*. Table 1 summarizes the sample sizes and sources for each emotion type. Most emotions have comparable sample sizes across datasets, with the exception of *calmness* (present only in *RAVDESS*) and *neutral* (which varies significantly across datasets, except for *TESS*). This overview supports the diversity needed for robust emotion recognition model evaluation. The EEG signal dataset was collected from two individuals, a male, and a female, with recordings of the three emotional states—positive, neutral, and negative using dry electrodes and an EEG headgear the brainwave activity data were recorded [19] . These datasets were chosen for their high-quality annotations, emotional diversity, and prior benchmarking relevance, facilitating reliable evaluation and comparability with existing models. These datasets are publicly available and collected under ethical guidelines with informed consent. The data is anonymized, ensuring participant privacy, and no personally identifiable information (PII) is included.

Table 1
Dataset characterization.

Emotion	CREMA-D	SAVEE	RAVDESS	TESS
Calmness	0	0	192	0
Happiness	1271	60	192	400
Neutral	1087	120	96	400
Sadness	1271	60	192	400
Anger	1271	60	192	400

To ensure generalization to unseen data, the datasets were split into 70% training, 15% validation, and 15% testing, with proportional distribution across emotions for speech data and a balanced representation of emotional states for EEG. A dedicated test set, unseen during training and validation, was used to reliably evaluate model performance.

3.2. Speech and EEG models

Existing approaches in multimodal EEG and speech processing primarily focus on sequential feature extraction, where temporal dependencies are modeled in a unidirectional manner. While these methods

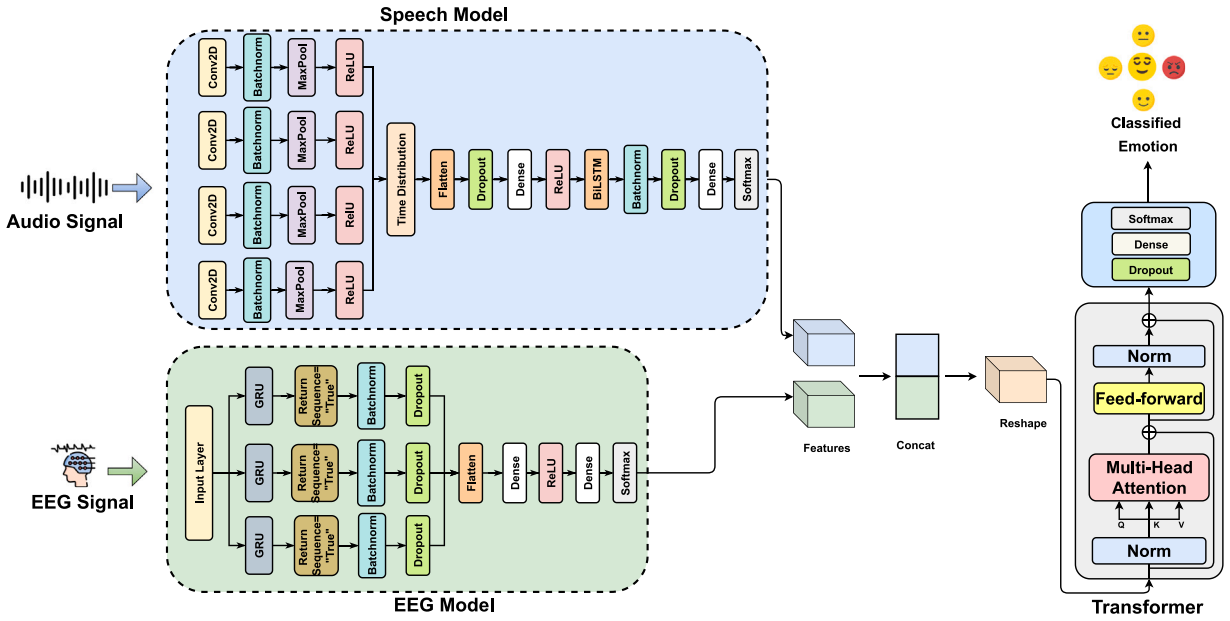

Fig. 3. Proposed TMNet multimodal architecture highlights the convergence of the Speech and EEG signals into the Transformer for complex model feature extraction and prediction.

capture time-series patterns effectively, they often fail to retain long-range dependencies and contextual relationships between EEG and speech modalities. In contrast, TMNet introduces a bidirectional feature extraction strategy allowing the model to process information from both past and future contexts simultaneously. This bidirectional analysis enhances temporal modeling and improves cross-modal alignment, leading to a more comprehensive representation of multimodal affective signals.

For the speech modality, we utilized the CNN-BiLSTM architecture because it effectively leverages the complementary strengths of CNNs and RNNs. The CNN layers are adept at extracting robust spatial features from the MFCCs, capturing the local patterns and nuances inherent in speech signals. Meanwhile, the BiLSTM layers are capable of modeling the temporal dependencies from both past and future contexts, which is crucial for understanding the dynamic nature of speech. This combination not only improves the representation of the audio signal but also addresses the limitations of using purely convolutional or unidirectional recurrent architectures. Existing literature supports that integrating CNN with BiLSTM enhances performance in speech emotion recognition tasks by capturing both local and long-range dependencies, thereby providing a more holistic understanding of the input data. Using Keras, the CNN-BiLSTM model processes 3D tensors of dimensions (max_len, num_mfcc, 1), where max_len represents the audio signal length and num_mfcc denotes the number of MFCC features. The model is built with Conv2D layers for feature extraction, followed by MaxPooling layers for dimension reduction and a TimeDistributed Flatten layer. A BiLSTM layer with 128 units captures temporal dependencies, while the final Dense layer with softmax activation classifies the emotions into five categories. Fig. 3 shows the detailed architecture of the multimodal model.

The convolutional layer's output, represented by $\mathbb{C}_{\text{output}}$, is given as $\mathbb{C}_{\text{output}} = R(W_c \cdot X + b_c) \equiv \propto$, where R is the ReLU activation function, W_c is the weight matrix, and b_c is the bias term. The convolutional layer is chosen for its ability to efficiently extract spatial features from the input data, which is essential in capturing local patterns in the speech signals represented by MFCCs. These local patterns are key to distinguishing emotions based on speech characteristics such as tone and pitch. The MaxPooling layer reduces spatial dimensions, expressed as $\mathbb{P}_{\text{output}} = \text{Max}\mathbb{P}_{\propto} \equiv \delta$. MaxPooling is used to down-sample the output from the convolutional layer, reducing its dimensionality while retaining the most important features, thus improving computational efficiency and preventing overfitting. Next, the TimeDistributed layer performs a linear transformation on the pooled feature maps, defined as $\mathbb{T}\text{Dist}_{\text{output}} = W_{\text{TDist}} \cdot \delta + b_{\text{TDist}} \equiv \mu$, where W_{TDist} represents learnable weights and b_{TDist} the bias term. This layer is essential because it applies the same transformation to each time step of the feature map, enabling the model to handle sequential data without losing the temporal order of the features. A dropout layer is included to prevent overfitting, denoted by $\mathbb{D}_{\text{output}} = \mathbb{D}(\mu) \equiv \gamma$. Dropout is a regularization technique that randomly drops units during training to reduce model overfitting and improve generalization to new data. The BiLSTM layer, which captures temporal dependencies in both forward and backward directions, is represented as $\mathbb{L}_{\text{output}} = \text{BiLSTM}(\mathbb{L}_{\text{units}}) \cdot \gamma \equiv \beta$, where $\mathbb{L}_{\text{units}}$ refers to the number of LSTM units. BiLSTM was selected for its ability to model long-range temporal dependencies in speech signals. The bidirectional nature of the LSTM allows the model to consider both past and future context, which is crucial for accurately interpreting emotions in speech, where context plays a key role in emotional meaning. Finally, the output layer applies softmax for classification, computed as $\Omega_{\text{output}} = \text{Softmax}(W_{\text{output}} \cdot \beta + b_{\text{output}})$, where W_{output} and b_{output} are the weight matrix and bias term, respectively. Softmax is used to convert the output of the LSTM into a probability distribution over the emotion categories, making the model capable of classifying the speech signal into predefined emotional categories.

We selected the BiGRU model for EEG modality due to its efficiency and effectiveness in handling temporal sequences. EEG signals, which inherently contain sequential data with complex temporal dependencies, benefit significantly from gated recurrent units. The bidirectional

Table 2
Comparing the CNN-BiLSTM with existing methods.

Model	Datasets	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
CNN-MLP [20]	R	86.40	86.50	86.30	86.32
CNN [21]	R	71.61	71.80	71.50	71.65
CNN [22]	R+S+T+C	92.60	92.70	92.50	92.60
CNN [23]	R+S+T+C	90.00	90.10	89.90	89.91
CNN-BiLSTM	R+S+T+C	**95.33**	**95.40**	**95.30**	**95.35**

*R: RAVDESS, *S: SAVEE, *T: TESS, *C: CREMA-D

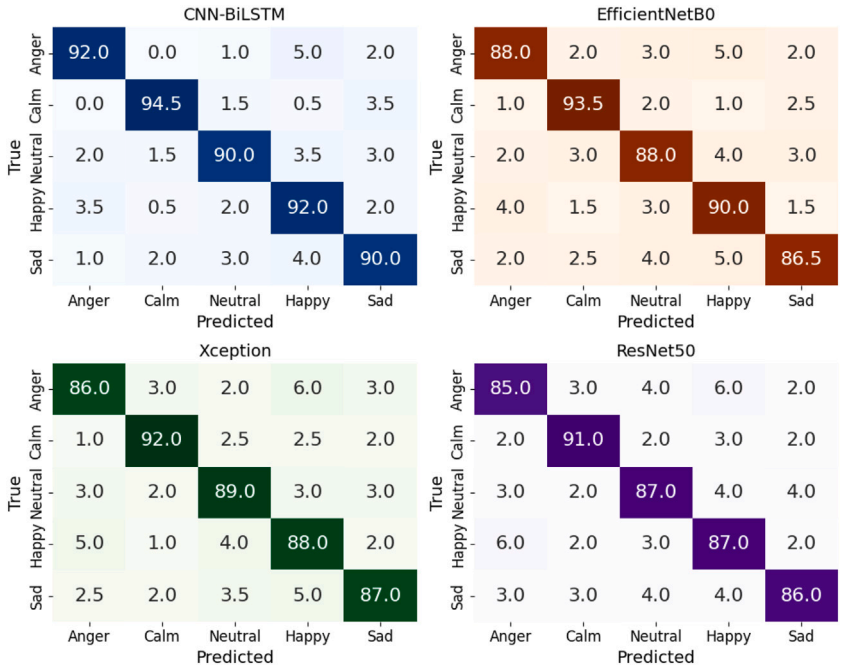

Fig. 4. Comparison of the proposed CNN-BiLSTM with state-of-the-art model.

nature of BiGRU further enhances the model's ability to capture information from past and future time steps, ensuring a more comprehensive representation of the neural activity related to emotional states. Compared to other architectures, such as standard RNNs or even LSTMs, BiGRUs offer a more streamlined structure with fewer parameters, reducing computational complexity while maintaining high performance. This balance between efficiency and accuracy makes BiGRU particularly well-suited for real-time EEG-based emotion recognition tasks, as supported by recent studies in the field. The model processes 3D tensors of shape $(\text{max\_len}, \text{num\_features}, 1)$, where max_len is the sequence length, and num_features is the number of EEG features. Each input sequence X passes through three stacked BiGRU layers, chosen for their ability to efficiently capture temporal dependencies in sequential EEG data. The output of the lth layer is defined as $\mathbb{G}^{(l)}_{\text{output}} = \text{BiGRU}(W^{(l)}_{\text{GRU}} \cdot X + b^{(l)}_{\text{GRU}})$, where $W^{(l)}_{\text{GRU}}$ and $b^{(l)}_{\text{GRU}}$ are the weights and biases. A Flatten layer transforms the final BiGRU output into a 1D vector, $\mathbb{F}_{\text{output}} = \text{Flatten}(\mathbb{G}^{(3)}_{\text{output}})$. The Dense layer with softmax activation then produces the emotion classification across three categories, computed as $\Omega_{\text{output}} = \text{Softmax}(W_{\text{Dense}} \cdot \mathbb{F}_{\text{output}} + b_{\text{Dense}})$, with W_{Dense} and b_{Dense} as the weight and bias. This architecture is chosen for its effectiveness in capturing EEG signal dynamics and accurately recognizing emotional states from the data.

3.3. Multimodal fusion with transformer

The proposed multimodal framework utilizes a Transformer encoder layer to fuse the outputs from the CNN-BiLSTM speech model and the BiGRU EEG model, enhancing emotion recognition by capturing dependencies between these two modalities. First, each model's output is concatenated into a single vector, $\mathbf{H} = [\mathbf{H}_{\text{speech}}, \mathbf{H}_{\text{EEG}}]$, representing the joint feature space. This concatenated vector is then processed by the Transformer encoder, which applies multi-head self-attention and feed-forward transformations to learn inter-modal interactions. In the multi-head attention mechanism, the input $\mathbf{H}$ is transformed into query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) matrices through learned weights. The attention score for each head is calculated as $\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$, where d_k is the key dimension, stabilizing gradients. Each head captures different aspects of the input, and their outputs are concatenated and linearly transformed, yielding a unified attention vector. The multi-head output, $\mathbf{Z}$, then passes through a feed-forward network (FFN) with a ReLU activation function, $\text{FFN}(\mathbf{Z}) = \text{ReLU}(\mathbf{Z}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2$, introducing non-linearity for more complex feature representation. Layer normalization and dropout are applied to stabilize learning and prevent overfitting. Finally, a dense layer with softmax activation is used for classification, where the probability distribution for each emotion class is obtained as $\text{Output} = \text{softmax}(\mathbf{T}\mathbf{W}_{\text{output}} + \mathbf{b}_{\text{output}})$.

This Transformer-driven fusion enables the model to capture intricate interdependencies between EEG and speech data, effectively leveraging complementary information from both modalities. By selectively focusing on the most relevant features, the Transformer enhances classification accuracy, resulting in a more robust and nuanced emotion recognition framework.

Table 3
Comparison of BiGRU with existing methods.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
DeepMaxout [24]	90.84	91.00	90.50	90.75
SVM [10]	90.00	90.20	89.80	89.88
GRU [25]	96.95	97.00	96.90	96.51
BiGRU	**97.25**	**97.40**	**97.20**	**97.21**

4. Result discussion and evaluation

4.1. Speech model evaluation

The proposed CNN-BiLSTM model outperforms state-of-the-art architectures such as EfficientNetB0, Xception, and ResNet50 in emotion classification, as shown in Fig. 4. The CNN-BiLSTM achieves consistently higher accuracy across all emotion categories, particularly excelling in distinguishing Neutral (94.5%), Happy (92.0%), and Sad (90.0%) emotions. This superior performance is attributed to the hybrid architecture, where CNN layers effectively extract spatial features from speech signals, while the BiLSTM layers capture temporal dependencies, enhancing sequential pattern recognition. Unlike standard CNN-based models that primarily focus on spatial representations, the

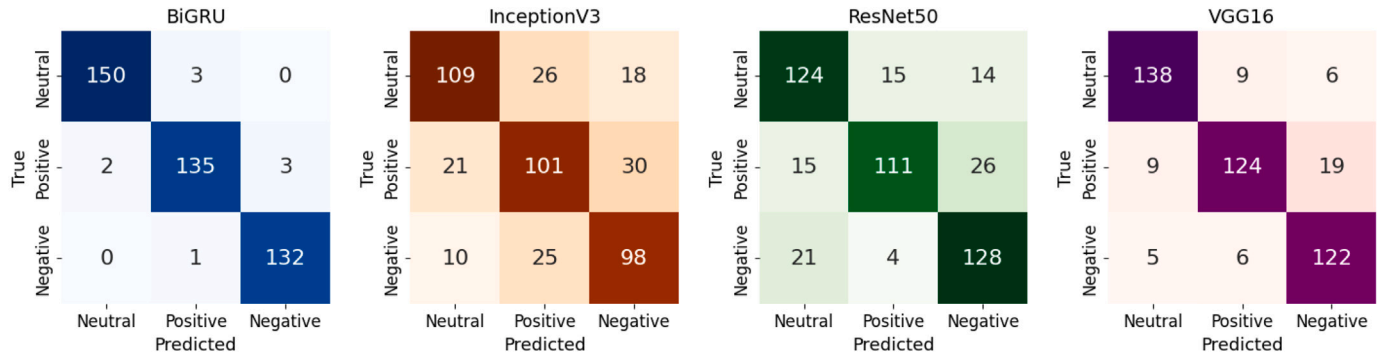

Fig. 5. Comparison of the proposed BiGRU with state-of-the-art model.

Table 4
Comparison of TMNet with existing approaches.

Model	Method	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Multimodal Transformer [5]	Transformer, SVM Classifier	91.79	92.00	91.50	91.70
Late Fusion [26]	CNN, GRU, Encoder	95.39	95.50	95.30	95.38
Feature Level Fusion [27]	SVM	87.44	87.60	87.30	87.45
Decision Level Fusion [4]	kNN, SVM	96.24	96.12	95.91	95.87
Hierarchical LSTM Fusion with Self-Attention [28]	LSTM, Attention	94.86	93.96	94.89	94.88
Multi CNN Fusion [29]	DeepVANet, CNN	95.61	95.11	95.08	95.01
TMNet	CNN-BiLSTM, BiGRU, Transformer	**98.89**	**99.00**	**98.80**	**98.88**

BiLSTM component in the proposed approach allows the model to better understand contextual variations in speech, leading to more precise predictions. The results demonstrate that CNN-BiLSTM's ability to learn both local and long-range dependencies significantly improves emotion recognition accuracy compared to conventional deep learning models.

Similarly, Table 2 shows the comparison of the proposed CNN-BiLSTM with existing methods where similar datasets are used. The results demonstrate that CNN-BiLSTM's ability to learn both local and long-range dependencies significantly improves emotion recognition accuracy compared to existing methods [20–23] that uses conventional CNN based approach.

4.2. EEG signal model evaluation

The BiGRU model achieves state-of-the-art performance in emotion classification, outperforming based models like InceptionV3, ResNet50, and VGG16, as demonstrated in Fig. 5. The confusion matrices show that BiGRU effectively distinguishes Neutral (150), Positive (135), and Negative (132) emotions with minimal misclassification, highlighting its ability to model long-range dependencies in sequential data. BiGRU leverages temporal relationships, making it more suitable for emotion recognition.

In Table 3, BiGRU achieves the highest accuracy (97.25%), precision (97.40%), recall (97.20%), and F1 score (97.21%), surpassing DeepMaxout (90.84%), SVM (90.00%), and GRU (96.95%). The improvement over GRU demonstrates the effectiveness of bidirectional processing, allowing BiGRU to capture contextual dependencies in both forward and backward directions, leading to a more robust emotion classification model. These results validate the effectiveness of BiGRU in achieving higher recognition accuracy compared to conventional models.

4.3. Multimodal emotion recognition: TMNet performance evaluation

To assess TMNet's effectiveness in multimodal emotion recognition, we compare it against several state-of-the-art models employing Transformer, CNN, LSTM, SVM, kNN, and various fusion strategies, as shown in Table 4. The Multimodal Transformer model [5] performs well in capturing long-range dependencies but struggles with complex feature representations due to SVM constraints, which limit its ability to handle diverse feature types effectively, leading to an accuracy of 91.79% and an F1-score of 91.70%. TMNet solves this issue by employing a Transformer-based fusion that allows for more dynamic and adaptive integration of features from both speech and EEG modalities. By using multi-head self-attention, TMNet effectively captures the complex interdependencies between these modalities, allowing it to outperform the Multimodal Transformer model by 7.1% in accuracy, and 3.5% in F1-score. The Late Fusion model [26], which processes EEG and speech signals separately before merging them, achieves 95.39% accuracy and an F1-score of 95.38%. However, its late fusion strategy limits effective cross-modal interactions, preventing it from leveraging the rich, complementary information present in both modalities. TMNet addresses this limitation by fusing EEG and speech data at an earlier stage in the model, allowing for simultaneous processing of both modalities. This early fusion strategy enables TMNet to model bidirectional dependencies and contextual relationships between the EEG and speech signals, which improves its ability to recognize emotions across both modalities. As a result, TMNet outperforms the Late Fusion model by 3.5% in accuracy and 11.45% in F1-score. The Feature-Level Fusion model [27], relying on manually extracted EEG features and lacking sequential modeling, has the lowest accuracy (87.44%) and precision (87.60%), highlighting its poor generalizability. In contrast, TMNet leverages end-to-end deep learning to automatically learn and extract sequential features from both EEG and speech signals. The use of BiGRU for EEG and CNN-BiLSTM for speech allows TMNet to capture the temporal dynamics and long-range dependencies inherent in both signal types. This end-to-end learning, combined with the Transformer for fusion, enables TMNet to outperform Feature-Level Fusion by 11.45% in accuracy and achieve higher precision and generalizability.

Other fusion strategies also exhibit limitations. The Decision-Level Fusion model [4], despite achieving a strong 96.24% accuracy, struggles with intra-modal dependency modeling, which affects its recall (95.91%) and overall feature alignment. This limitation arises from its approach of processing each modality separately before merging, which fails to fully exploit the rich, interdependent information between EEG and speech signals. TMNet addresses this challenge by employing a Transformer-based fusion, which integrates EEG and speech data at an earlier stage, allowing for dynamic and adaptive integration of features from both modalities. The use of multi-head self-attention in

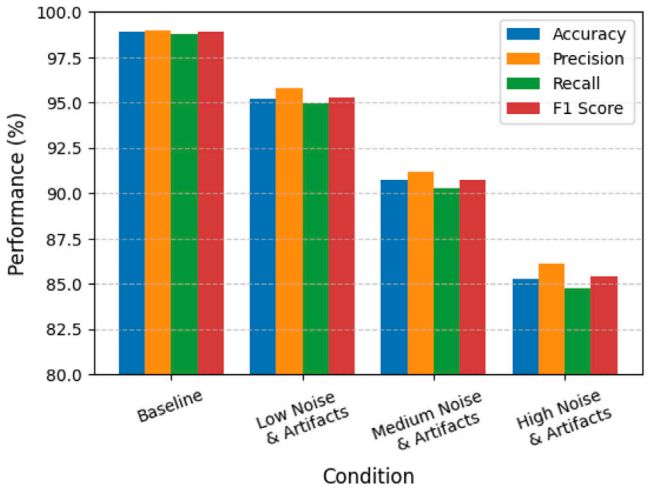

Fig. 6. Performance evaluation of TMNet under different noise and artifact conditions.

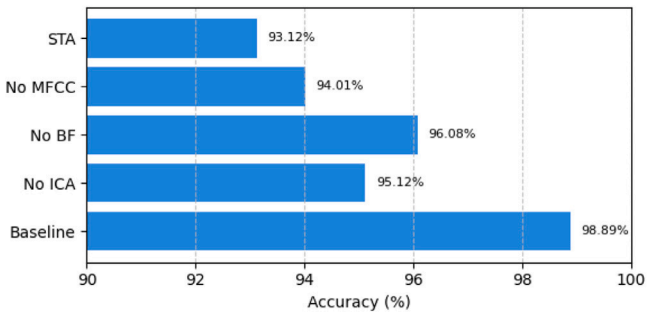

Fig. 7. Impact of each component on TMNet's performance.

Table 5
Performance comparison across different architectures of the proposed model.

Speech Model	EEG Model	MAE	Parameters	Inference Time (s)
BiGRU	**CNN-BiLSTM**	**0.05**	**5.6M**	**0.12**
GRU	LSTM	0.12	4.8M	0.18
CNN	CNN-LSTM	0.15	3.9M	0.15
RNN	MLP	0.35	1.4M	0.08
BiGRU	DCNN	0.18	5.5M	0.20
CNN-GRU	CNN-BiLSTM	0.10	7.3M	0.25
LSTM	CNN-BiLSTM	0.13	6.4M	0.22
DCNN-BiGRU	DCNN-BiLSTM	0.9	9.5M	0.30

the Transformer enables TMNet to capture the intricate interdependencies between EEG and speech, leading to improved alignment and more accurate emotion recognition, thereby surpassing the Decision-Level Fusion model by 7.1% in accuracy and 3.5% in F1-score. The Hierarchical LSTM Fusion model [28], while incorporating LSTM and self-attention, shows good recall (94.89%) but suffers from unidirectional modeling constraints, limiting its ability to capture long-term dependencies in sequential data. This approach fails to fully account for bidirectional dependencies, which are crucial for understanding the temporal dynamics of emotions. In contrast, TMNet leverages Bi-GRU and CNN-BiLSTM layers, which effectively model bidirectional temporal dependencies in both EEG and speech signals. This enables TMNet to capture both past and future contextual information, offering a richer, more comprehensive representation of the sequential affective patterns present in the data. As a result, TMNet outperforms the Hierarchical LSTM Fusion model by 3.5% in accuracy and 11.45% in F1-score. The Multi-CNN Fusion model [29], which excels in spatial feature extraction, attains high precision (95.11%) but struggles to capture the temporal dependencies inherent in EEG and speech signals, limiting its overall performance with a capped accuracy of 95.61%. This model primarily focuses on spatial patterns and overlooks the sequential nature of the data, which is critical for emotion recognition. TMNet, however, addresses this limitation by incorporating BiGRU and CNN-BiLSTM, which provide end-to-end deep learning capabilities for both temporal and spatial feature extraction. This allows TMNet to capture both spatial and temporal dependencies in EEG and speech signals, resulting in superior performance. The fusion mechanism in TMNet further optimizes this integration, ensuring that both modalities contribute effectively to the final classification. Consequently, TMNet surpasses the Multi-CNN Fusion model by 3.28% in accuracy. These architectural advancements in TMNet, such as early fusion, bidirectional temporal modeling, and Transformer-driven integration, directly address the limitations of previous models that either rely on shallow feature extraction or separate modality processing. By dynamically capturing the complex interactions between EEG and speech, TMNet significantly improves the model's ability to recognize emotions with greater accuracy and robustness.

Fig. 6 illustrates the performance of TMNet under different noise and artifact conditions, simulating real-world scenarios where EEG data is often contaminated by various factors such as muscle activity, eye blinks, and electromagnetic interference. In the Baseline condition, TMNet achieves optimal performance, benefiting from our comprehensive data preprocessing pipeline, including ICA for artifact removal and noise filtering. Performance naturally declines as noise and artifacts are progressively injected during the data augmentation phase. In the Low Noise & Artifacts condition, where we introduced slight muscle movement and low-frequency noise, TMNet maintains strong performance with 95.20% accuracy. This demonstrates the model's resilience to minor distortions typically seen in real-world data. However, when we introduce Medium Noise & Artifacts, such as more significant muscle activity and eye blinks, the accuracy drops to 90.75%, and recall decreases to 90.30%. This highlights the impact of moderate signal corruption on the model's feature extraction process, emphasizing the challenges posed by more pronounced artifacts and noise in EEG signals. In the High Noise & Artifacts setting, where severe data degradation was simulated by introducing electromagnetic interference, strong muscle artifacts, and motion-induced noise, TMNet still retains reasonable classification performance with 85.30% accuracy, 86.10% precision, and an F1-score of 85.40%. Despite the significant noise, the model continues to classify emotions effectively, showcasing its robustness in handling real-world noisy data. These results validate TMNet's adaptability and its capacity to function reliably even in challenging conditions, where the data quality is compromised.

Fig. 7 shows the ablation study comparing the baseline model, TMNet that uses our EEG and Speech signal preprocessing pipeline, with variants where a single component is removed or replaced. ICA reduces accuracy to 95.12%, highlighting the importance of artifact removal in EEG signals. Omitting the Bandpass Filter (BF) drops accuracy to 96.08%, indicating that retaining relevant frequency bands is crucial for robust feature extraction. Excluding MFCCs, our primary speech features cause accuracy to fall to 94.01%. Finally, replacing DTW with Simple Timestamp Alignment (STA) yields the lowest accuracy (93.12%), underscoring the significance of precise synchronization for multimodal fusion. Collectively, these results demonstrate that each component, along with the parameters we have used in our framework, meaningfully contributes to TMNet's overall performance.

We experimented with several EEG and Speech model architectures, and the proposed BiGRU and CNN-BiLSTM architecture emerges as the best choice for multimodal emotion recognition, as demonstrated in Table 5. A higher number of parameters enhances the model's capacity to capture complex, non-linear relationships within the data, thereby reducing the MAE. However, this comes with the caveat that excessively large models can lead to overfitting and increased computational demands. This combination achieves the lowest MAE of 0.05,

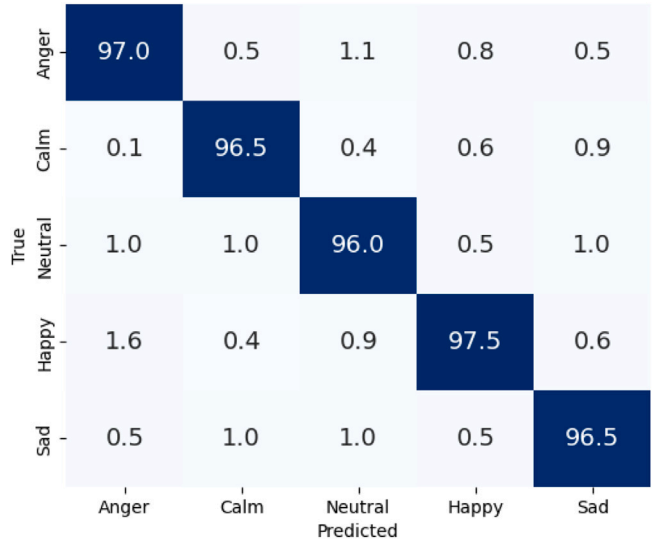

Fig. 8. Evaluation of the proposed TMNet for multilinguality and cultural diversity on the EmoDB dataset.

indicating superior accuracy in emotion classification with 5.6M parameters, striking a balanced trade-off between model complexity and performance. Compared to other model architectural combinations, such as GRU, LSTM (MAE = 0.12), and CNN, CNN-LSTM (MAE = 0.15), the BiGRU, CNN-BiLSTM model offers a significant improvement in accuracy while maintaining a lower computational burden. Other combinations, like lightweight RNN, MLP, and DCNN, BiGRU, show substantially higher MAE values (0.35 and 0.18, respectively), we observed that architectures with fewer parameters like RNN-MLP with 1.4M parameters struggled to capture the intricate temporal and spatial dynamics inherent in EEG and speech signals, resulting in higher MAE values up to 0.35. The combination of LSTM and CNN-BiLSTM shows that our choice of BiGRU for the speech model is optimal, as it increases the number of parameters but performs lower with MAE 0.13 and higher inference time.

To evaluate the linguistic and cultural diversity of the proposed TMNet, we conducted an additional evaluation using the EmoDB [30] dataset, as shown in Fig. 8, which provides annotated speech data in the German language. For consistency, we employed the same emotional categories as in the other datasets used in this study, ensuring that the dataset distribution and experimental setup were comparable. This approach allowed for a balanced and fair performance assessment across different linguistic contexts. The evaluation results demonstrate that TMNet performs effectively across all emotional categories, with Anger and Happy emotions achieving the highest recognition accuracies of 97.0% and 97.5%, respectively. Although there is a slight misclassification between Neutral and Sad emotions, the overall accuracy remains high, further validating the model's robustness. These results confirm that TMNet is capable of handling linguistic and cultural diversity, showing its generalization ability across different languages and emotional expressions.

5. Limitations and future works

TMNet demonstrates high accuracy, yet it remains prone to overfitting due to its complex architecture and relatively homogeneous training datasets. Additionally, the cultural and linguistic scope of the current corpora is limited, raising uncertainties about the model's ability to generalize to real-world scenarios with diverse languages, accents, and environmental noise. TMNet's resource-intensive design further constrains its deployment on edge devices and other low-power platforms.

To overcome these challenges, future efforts will focus on developing more diverse, realistic datasets that incorporate different cultural backgrounds, languages, and noisier EEG recordings. We will also explore model compression, pruning, and quantization techniques to reduce computational demands and facilitate real-time processing on constrained hardware. These enhancements aim to improve TMNet's robustness, transferability, and practical feasibility in a wide range of deployment environments.

6. Conclusion

This paper proposes a novel multimodal framework that leverages both speech and EEG signals for enhanced emotion recognition. The framework begins with an efficient data preprocessing pipeline that effectively handles speech and EEG signals by removing noise and artifacts from the data. It utilizes MFCCs from speech signals to serve as input features for a CNN-BiLSTM model, which effectively captures both spatial and temporal characteristics of the data. Simultaneously, a BiGRU-based neural network is employed to classify emotions from EEG signals. The chosen CNN-BiLSTM & BiGRU model outperforms other architectures like GRU & LSTM, CNN & CNN-LSTM. The outputs of both models are then fused through DTW, which refines temporal alignment and feeds the feature vector into the Transformer, which captures inter-modal dependencies to achieve an integrated and robust emotional representation. This Transformer-driven fusion approach handles the limitations of the previous single-modal and fusion-based approaches, thus achieving higher emotion recognition performance, achieving an accuracy of 98.89%. The proposed framework demonstrates promising performance under stress testing conditions with noise and artifacts, showcasing its robustness in real-world scenarios. The combination of the speech and EEG models in the proposed architecture stands out due to its lower inference time (0.12 sec) and reduced number of trainable parameters (5.6M), making it an efficient and scalable solution for emotion recognition applications.

CRediT authorship contribution statement

Md Mahinur Alam: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Mohamed A. Dini:** Writing – original draft, Visualization, Software, Resources, Methodology, Investigation, Formal analysis, Data curation. **Dong-Seong Kim:** Writing – review & editing, Supervision. **Taesoo Jun:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that there is no conflict of interest in this paper

Acknowledgments

This research received 34% funding from the Innovative Human Resource Development for Local Intellectualization Program, South Korea (IITP-2025-RS-2020-II201612) through the Institute of Information & Communications Technology Planning & Evaluation (IITP) under the Ministry of Science and ICT (MSIT), the Basic Science Research Program, South Korea (2018R1A6A1A03024003) through the National Research Foundation of Korea (NRF), contributing 33%, and the Information Technology Research Center (ITRC) Program, South Korea (IITP-2025-RS-2024-00438430) funded by MSIT through IITP(Institute for Information Communications Technology Planning Evaluation), South Korea, contributing 33%.

References

[1] N. Chitre, N. Bhorade, P. Topale, J. Ramteke, C.R. Gajbhiye, Speech Emotion Recognition to assist Autistic Children, Institute of Electrical and Electronics Engineers Inc., ISBN: 9781665497107, 2022, pp. 983–990, http://dx.doi.org/10.1109/ICAAIC53929.2022.9792663.

[2] R. Matin, D. Valles, A Speech Emotion Recognition Solution-based on Support Vector Machine for Children with Autism Spectrum Disorder to Help Identify Human Emotions, Institute of Electrical and Electronics Engineers Inc., ISBN: 9781728142913, 2020, http://dx.doi.org/10.1109/IETC47856.2020.9249147.

[3] H. Liu, T. Lou, Y. Zhang, Y. Wu, Y. Xiao, C.S. Jensen, D. Zhang, EEG-based multimodal emotion recognition: a machine learning perspective, IEEE Trans. Instrum. Meas. (2024).

[4] Q. Wang, M. Wang, Y. Yang, X. Zhang, Multi-modal emotion recognition using EEG and speech signals, Comput. Biol. Med. 149 (2022) 105907.

[5] F. Zhu, J. Zhang, R. Dang, B. Hu, Q. Wang, MTNet: Multimodal transformer network for mild depression detection through fusion of EEG and eye tracking, Biomed. Signal Process. Control. 100 (2025) 106996.

[6] P. Tzirakis, J. Zhang, B.W. Schuller, End-to-End Speech Emotion Recognition Using Deep Neural Networks, vol. 2018-April, Institute of Electrical and Electronics Engineers Inc., (ISSN: 15206149) ISBN: 9781538646588, 2018, pp. 5089–5093, http://dx.doi.org/10.1109/ICASSP.2018.8462677.

[7] L. Zhang, Y. Zhou, P. Gong, D. Zhang, Speech imagery decoding using EEG signals and deep learning: A survey, IEEE Trans. Cogn. Dev. Syst. (2024).

[8] M. Tahir, A. Tubaishat, F. Al-Obeidat, B. Shah, Z. Halim, M. Waqas, A novel binary chaotic genetic algorithm for feature selection and its utility in affective computing and healthcare, Neural Comput. Appl. (2022) 1–22.

[9] T. Liu, M.Z.H. Shah, X. Yan, D. Yang, Unsupervised feature representation based on deep boltzmann machine for seizure detection, IEEE Trans. Neural Syst. Rehabil. Eng. 31 (2023) 1624–1634.

[10] N. Krupa, K. Anantharam, M. Sanker, S. Datta, J.V. Sagar, Recognition of emotions in autistic children using physiological signals, Heal. Technol. (ISSN: 21907196) 6 (2016) 137–147, http://dx.doi.org/10.1007/s12553-016-0129-3.

[11] S.O. Ajakwe, D.-S. Kim, J.M. Lee, CogNet: Cognitive super resolution network for persistent end-to-end mobility communication in MIMO systems, in: 2023 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication, IConSCEPT, 2023, pp. 1–4, http://dx.doi.org/10.1109/IConSCEPT57958.2023.10170238.

[12] J.J. Bird, D.R. Faria, L.J. Manso, A. Ekárt, C.D. Buckingham, A deep evolutionary approach to bioinspired classifier optimisation for brain-machine interaction, Complexity 2019 (1) (2019) 4316548.

[13] H. Lerogeron, R. Picot-Clémente, A. Rakotomamonjy, L. Heutte, Approximating dynamic time warping with a convolutional neural network on EEG data, Pattern Recognit. Lett. 171 (2023) 162–169.

[14] K. Gorur, E. Olmez, Z. Ozer, O. Cetin, EEG-driven biometric authentication for investigation of Fourier synchrosqueezed transform-ICA robust framework, Arab. J. Sci. Eng. 48 (8) (2023) 10901–10923.

[15] R.F. Livingstone SR, The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north American english, PLoS One 13 (5) (2018) http://dx.doi.org/10.1371/journal.pone.0196391.

[16] M.K. Pichora-Fuller, K. Dupuis, Toronto Emotional Speech Set (TESS), Borealis, 2020, http://dx.doi.org/10.5683/SP2/E8H2MF.

[17] S. Haq, P. Jackson, Speaker-dependent audio-visual emotion recognition, in: Proc. Int. Conf. on Auditory-Visual Speech Processing, AVSP'08, Norwich, UK, 2009.

[18] H. Cao, D.G. Cooper, M.K. Keutmann, R.C. Gur, A. Nenkova, R. Verma, CREMA-D: Crowd-sourced emotional multimodal actors dataset, IEEE Trans. Affect. Comput. 5 (4) (2014) 377–390, http://dx.doi.org/10.1109/TAFFC.2014.2336244.

[19] J.J. Bird, D.R. Faria, L.J. Manso, A. Ekárt, C.D. Buckingham, A deep evolutionary approach to bioinspired classifier optimisation for brain-machine interaction, Complexity (ISSN: 10990526) 2019 (2019) http://dx.doi.org/10.1155/2019/4316548.

[20] Y. Ammanavar, S. Patil, A.P. Bidargaddi, P. Deshanur, R. Rathod, S. Meena, Speech emotion recognition on RAVDESS dataset using deep learning, in: 2024 5th International Conference for Emerging Technology, INCET, IEEE, 2024, pp. 1–6.

[21] D. Issa, M.F. Demirci, A. Yazici, Speech emotion recognition with deep convolutional neural networks, Biomed. Signal Process. Control. 59 (2020) 101894.

[22] S. Ullah, Q.A. Sahib, S. Ullahh, I.U. Haq, I. Ullah, et al., Speech emotion recognition using deep neural networks, in: 2022 International Conference on IT and Industrial Technologies, ICIT, IEEE, 2022, pp. 1–6.

[23] S.S. Pujitha, V.A. Sai, J.B. Sree, S. Shareefunnisa, B. Suvarna, Enhanced speech emotion recognition through convolutional neural networks, in: 2024 15th International Conference on Computing Communication and Networking Technologies, ICCCNT, IEEE, 2024, pp. 1–6.

[24] G.K. Chakravarthy, M. Suchithra, S. Thatavarti, EEG-based multimodal emotion recognition with optimal trained hybrid classifier, Multimedia Tools Appl. 83 (17) (2024) 50133–50156.

[25] M.A. Dini, M.J.A. Shanto, G.-H. Kwon, D.-S. Kim, T. Jun, Emotion type recognition scheme using EEG based signals, 2023, pp. 489–490.

[26] A. Das, P. Soni, M.-C. Huang, F. Lin, W. Xu, Multimodal speech recognition using EEG and audio signals: A novel approach for enhancing ASR systems, Smart Heal. 32 (2024) 100477.

[27] Z. Ning, H. Hu, L. Yi, Z. Qie, A. Tolba, X. Wang, A depression detection auxiliary decision system based on multi-modal feature-level fusion of EEG and speech, IEEE Trans. Consum. Electron. (2024).

[28] D. Wu, J. Zhang, Q. Zhao, Multimodal fused emotion recognition about expression-EEG interaction and collaboration using deep learning, IEEE Access 8 (2020) 133180–133189, http://dx.doi.org/10.1109/ACCESS.2020.3010311.

[29] S. Wang, J. Qu, Y. Zhang, Y. Zhang, Multimodal emotion recognition from EEG signals and facial expressions, IEEE Access 11 (2023) 33061–33068.

[30] S. Akinpelu, S. Viriri, A. Adegun, Lightweight deep learning framework for speech emotion recognition, IEEE Access 11 (2023) 77086–77098.