TEREE
Transformer-based emotion recognition using EEG and Eye movement data

Esmi, Nima; Shahbahrami, Asadollah; Gaydadjiev, Georgi; de Jonge, Peter

DOI
10.1016/j.ibmed.2025.100305

Publication date
2025

Document Version
Final published version

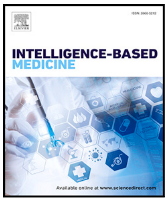
Published in
Intelligence-Based Medicine

Intelligence-Based Medicine 12 (2025) 100305

Contents lists available at ScienceDirect

Intelligence-Based Medicine

journal homepage: www.sciencedirect.com/journal/intelligence-based-medicine



TEREE: Transformer-based emotion recognition using EEG and Eye movement data

Nima Esmi [a,b,*], Asadollah Shahbahrami [c,b], Georgi Gaydadjiev [d], Peter de Jonge [e]

[a] Bernoulli Institute, University of Groningen, Groningen, The Netherlands
[b] Intelligent Systems Research Center, Khazar University, Baku, Azerbaijan
[c] Department of Computer Engineering, University of Guilan, Guilan, Iran
[d] Computer Engineering Laboratory, Delft University of Technology, Delft, The Netherlands
[e] Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, The Netherlands

ARTICLE INFO

Keywords:
Multimodal transformer
Emotion recognition
Bayesian spurious correlation minimization (BSCM)
Electroencephalogram (EEG)
Eye Movement (EM)

ABSTRACT

Background: Multimodal AI systems increasingly rely on biomedical signals such as EEG and eye movement data for emotion recognition. However, these models face challenges including limited training data, inter-subject variability, session-specific spurious correlations, and incomplete modality representation, all of which reduce generalization and reliability.
Method: We propose TEREE, a multimodal transformer-based model that integrates temporal, spatial, and spectral EEG features with eye movement data. To mitigate session-specific artifacts, Bayesian Spurious Correlation Minimization (BSCM) is applied. In addition, a holistic multimodal processing strategy enables robust handling of incomplete data. The model was trained and evaluated using the SEED and SEED-FRA benchmark datasets under one-to-one and multi-to-one transfer paradigms.
Results: TEREE achieved state-of-the-art performance, with average multi-to-one transfer accuracies of 97.7% on SEED and 98.8% on SEED-FRA. Ablation studies confirmed that fusing EEG with eye movement features consistently improved accuracy compared to unimodal baselines. Standard deviations across repeated experiments were below 5%, indicating stability.
Conclusion: By addressing inter-subject variability, spurious correlations, and incomplete modality issues, TEREE enhances the robustness and generalization of emotion recognition systems. These findings suggest that multimodal transformer-based models can substantially improve the reliability of affective computing applications such as human–computer interaction and mental health monitoring.

1. Introduction

Sentiment analysis has emerged as an essential tool across various domains, enabling applications such as mental health monitoring for early detection of conditions like depression, anxiety, or Alzheimer's disease; brain–computer interfaces that facilitate communication for individuals with disabilities; personalized learning that tailors educational content based on students' emotional states; human–computer interaction to enhance user experiences in gaming, virtual reality, and AI-driven systems; workplace productivity to monitor employee well-being; and market research to gauge consumers' emotional responses to advertisements and products [1–7].

Recent advancements in multimodal transformer-based models have significantly improved accuracy in detection and classification tasks, often surpassing traditional state-of-the-art methods [8,9]. These models excel at integrating multiple modalities, providing a comprehensive representation of complex data. In the context of human behavior and emotion analysis, such systems leverage diverse biomedical data, with electroencephalography (EEG) being a cornerstone modality due to its rich behavioral and emotional information [10–12]. Additionally, EEG-based models have shown promise in cognitive state recognition and neuropsychiatric disorder diagnosis, further expanding their utility [13,14].

Despite these advancements, the data-intensive nature of transformers, coupled with the high costs and constraints of collecting EEG datasets, presents significant challenges [15,16]. First, EEG signal features vary across subjects, leading to models that perform well for some individuals but poorly for others, thus hindering generalization [17–

[*] Corresponding author at: Bernoulli Institute, University of Groningen, Groningen, The Netherlands.
E-mail addresses: n.esmi.rudbardeh@rug.nl (N. Esmi), shahbahrami@guilan.ac.ir (A. Shahbahrami), g.n.gaydadjiev@tudelft.nl (G. Gaydadjiev), peter.de.jonge@rug.nl (P.d. Jonge).

https://doi.org/10.1016/j.ibmed.2025.100305
Received 9 June 2025; Received in revised form 19 September 2025; Accepted 13 October 2025
Available online 21 October 2025

19]. Second, external factors such as electrode placement, recording duration, and environmental conditions can introduce session-specific patterns, causing spurious correlations where models mistakenly associate recording conditions with class labels rather than intrinsic neural features [20,21]. Third, incomplete modality representation – caused by hardware malfunctions, poor electrode contact, or movement artifacts – results in missing or unusable signals, complicating classification [22,23].

To address these limitations, several multimodal approaches have been proposed. Weighted representation distribution alignment balances marginal and conditional distributions between source and target domains to mitigate individual differences in cross-subject emotion recognition, though significant domain disparities can lead to suboptimal alignment [24]. Domain-adversarial neural networks combine domain adaptation with deep feature learning to ensure discriminative and domain-invariant features, though large domain shifts may still impair generalization [25]. Dynamic domain adaptation algorithms adjust models in real time to address global and local domain divergences, enhancing performance in cross-subject and cross-session EEG emotion recognition; however, their reliance on domain-shift assumptions limits applicability in highly variable domains [26]. Additionally, advanced neural network architectures – such as spiking neural networks and graph neural networks – have been developed to improve EEG-based emotion recognition by capturing temporal and spatial dependencies in signals [12,27].

Motivated by the strengths and limitations of transformers in handling multimodal data, we propose a multimodal transformer-based model that integrates EEG with a supplementary modality, such as eye movement (EM). This approach, termed *Transformer-Based Emotion Recognition using EEG and Eye Movement Data (TEREE)*, aims to enhance classification performance by addressing the aforementioned challenges and enabling accurate identification of emotional states (positive, neutral, negative). The integration of EEG and EM leverages their complementary strengths, with EM providing direct insights into attentional and emotional states through gaze patterns and pupil dynamics, offering advantages over modalities like GSR or ECG [28,29]. For example, [24] demonstrated that the use of supplementary eye movement data can improve model accuracy by up to 10%. Additionally, [30] showed that EM can be used independently for emotion recognition with an accuracy exceeding 80%. Niaki et al. proposed a bipartite graph adversarial network that integrates bipartite graphs into a DANN framework to better handle cross-subject variability; their model achieved state-of-the-art or comparable performance, highlighting the effectiveness of graph-based domain adaptation for robust generalization [31]. Furthermore, incorporating paradigms from EEG-based disorder diagnosis and cognitive state recognition can enhance the robustness of our model across diverse applications [13,32].

EEG captures neural activity, while EM's precise behavioral data enhances robustness against incomplete data and individual variability. Validated on the SEED dataset, this combination outperforms alternative modality pairings in emotion recognition accuracy. Initially, to mitigate the effect of individual differences, we map EEG signals into a two-dimensional space, ensuring that different frequency bands and all channels are considered over time. Then, inspired by [33], we apply Bayesian Spurious Correlation Minimization (BSCM) to reduce session-specific artifacts. In our framework, BSCM models attention weights as Bayesian posterior distributions and applies KL-regularized variational inference, ensuring that the transformer relies less on session-dependent noise and more on causal neural–behavioral patterns. Finally, the entire session's data is fed into the model at each evaluation stage to minimize the impact of incomplete modality representations. This approach enables the model to leverage the global self-attention mechanism when processing the data, allowing it to capture correlations that are spatially and temporally distant despite the presence of noise.

The main contributions of our work are summarized as follows:

- To address individual differences, we employ a data-space transformation method combined with a Vision Transformer (ViT), ensuring that spatial, spectral, and temporal features are preserved and effectively analyzed for robust EEG-based emotion recognition.
- We mitigate spurious correlations from session-specific artifacts using Bayesian Spurious Correlation Minimization (BSCM). By modeling attention weights as a Bayesian posterior distribution with variational inference and KL-regularized loss, BSCM reduces reliance on non-causal patterns and enhances generalization across sessions and subjects, as validated by improved emotion recognition accuracy on the SEED and SEED-FRA datasets.
- To address incomplete modality representation, we feed the full session sample into the model at each analysis stage. Leveraging the self-attention mechanism, the model gains a global view of the data, allowing healthy segments to compensate for noisy or missing parts, thereby enhancing robustness in emotion recognition.

The remainder of this paper is structured as follows: Section 2 provides background on the challenges of transformers in emotion recognition using biomedical data. Section 3 discusses related work, highlighting existing methods and their limitations. Section 4 presents the design of our proposed transformer-based model, detailing its key components. Section 5 describes the implementation details and experimental setup. Section 6 evaluates our approach with benchmark datasets and compares it with state-of-the-art methods. Finally, Section 7 concludes the paper with a summary of findings and potential future research directions.

2. Background knowledge

In this section, we first discuss some of the challenges that models face when dealing with limited-scale datasets, including individual differences in data, spurious correlations between modalities, and incomplete modality representation. We then introduce key concepts related to attention mechanisms, including multi-head self-attention and multi-head cross-attention.

2.1. Challenges in models due to data characteristics

The performance of models is strongly influenced by the characteristics of the data used for training. Key challenges include individual differences in data, spurious correlations between modalities, and incomplete modality representations, which are discussed in the following sections.

Intrinsic differences in human physiology and brain function can significantly affect data analysis and the performance of machine learning models. EEG signals, for example, exhibit considerable variability across individuals, even when they perform the same task. These variations arise from factors such as brain structure, lifestyle, stress levels, and overall physical condition. For instance, in the SEED dataset, described in Section 5, EEG recordings are collected from 12 subjects of different genders and age groups while they are exposed to positive, neutral, or negative video stimuli using a 62-channel EEG device.

In EEG data analysis, one major challenge is the occurrence of spurious correlations, such as those caused by the limited spatial resolution of electrophysiological methods. Recorded EEG signals may result from the mixing of activity across multiple brain sources, which can lead to artificial correlations between regions. This issue may result in misinterpretations of functional brain connectivity. For example, Fig. 1 illustrates patterns observed in the left hemisphere of the brain (R: Real) alongside artificially induced patterns resulting from spurious correlations (A: Artificial), inspired by [34]. Such spurious correlations can obscure the true structure of brain networks.

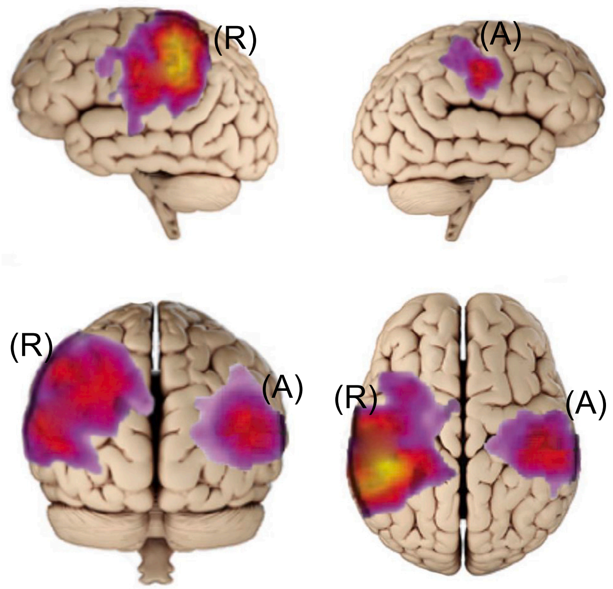


Fig. 1. Correlation patterns observed in the left hemisphere of the brain. (R: Real) shows true correlations, while (A: Artificial) depicts patterns arising from spurious correlations.
Source: Adapted from [34].

A critical challenge in multimodal learning is incomplete modality representation, where certain modalities may be partially or entirely missing. In real-world applications, missing data in one or more modalities is common. For example, in an emotion recognition system, some EEG channels may become disconnected from the subject's scalp, resulting in missing signals. In other cases, a modality is present but affected by high noise levels, distortion, or missing segments, making it difficult to extract meaningful information.

A real-world example from the SEED dataset illustrates this issue. Fig. 2 shows Short-Time Fourier Transform (STFT) representations from an experiment in which subjects watched a five-minute neutral video. Panel (a) represents the EEG signal from Channel 47, which remains unaffected, while Panel (b) represents Channel 48, where severe noise disrupts the entire frequency spectrum. As observed in panel (b), the adjacent channel remains unaffected, highlighting the variability of this problem.

The risk of incomplete modality representation increases in scenarios involving multiple modalities, as missing or corrupted data in one modality can reduce the overall performance of the model. Addressing this challenge requires robust techniques capable of handling missing or noisy data while maintaining model reliability across diverse conditions.

2.2. Data fusion using multi-head cross-attention

The presence of five distinct senses enables humans to perceive and interpret their surroundings effectively. For example, by simultaneously seeing and smelling a fruit, we can better assess its quality. Over time, the human brain has learned to associate visual appearance, aroma, and taste. While each sensory modality independently provides useful information, the likelihood of selecting a delicious strawberry increases when both vision and smell are combined.

A similar principle applies to machine learning models. One of the widely used mechanisms in deep learning is multimodal fusion, particularly in transformer-based architectures. In this paper, we focus on one type of fusion, while other common approaches are discussed in [35]. However, before addressing fusion in transformers, it is important to understand why transformers are advantageous in certain applications.

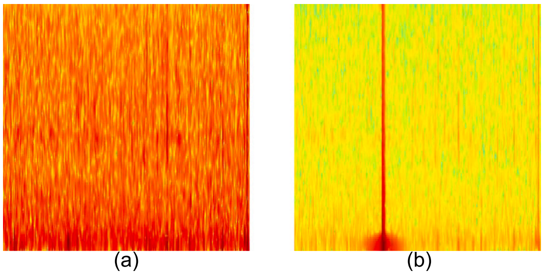


Fig. 2. Short-Time Fourier Transform representations of EEG signals during a five-minute neutral video. (a) Channel 47, unaffected. (b) Channel 48, where severe noise disrupts the entire frequency spectrum, illustrating the risk of incomplete modality representation in multimodal scenarios.

The fundamental operation of transformers is based on the attention mechanism. Fig. 3 provides a simplified comparison of how neurons interact in three deep learning models: CNNs, RNNs, and the attention mechanism.

In CNNs, the feedforward architecture ensures that only spatially adjacent input elements have direct connections. This design allows CNNs to effectively capture local relationships but limits their ability to identify global features in large-scale inputs. RNNs, in contrast, employ a sequential structure that captures order and positional dependencies of input elements. However, as the sequence length increases, the vanishing gradient problem reduces the influence of early inputs, restricting the model's ability to capture long-range dependencies.

Transformers address these limitations through the self-attention mechanism, which enables every input element to interact with all others. This property allows transformers to capture global dependencies effectively, albeit at a higher computational cost—a trade-off often justified in applications where accuracy is critical.

When modeling relationships across different modalities, multi-head cross-attention can be employed. As illustrated in Fig. 4, scaled dot-product attention serves as the core operation for both self-attention and cross-attention. It computes attention scores that determine how much one token (or feature) should attend to another. Multi-head self-attention focuses on dependencies within a single sequence, while multi-head cross-attention models interactions between different sequences. In applications such as EEG signal analysis, where related emotions may appear at distinct time intervals, multi-head attention is particularly beneficial for capturing temporal dependencies.

Multi-head cross-attention focuses on interactions between two different sequences, allowing one sequence to attend to and learn from the other. For example, EEG signal images and textual information related to eye movements, recorded simultaneously, are fed into two modalities $M1$ and $M2$. These are then processed by the multi-head cross-attention mechanism, producing $M1'$ and $M2'$, which represent the normalized modalities enriched by the influence of the other.

Using (Q_{M1}, K_{M1}, V_{M1}) and (Q_{M2}, K_{M2}, V_{M2}), the attention mechanism computes relevance scores that determine how much focus should be assigned to different parts of the input data. These scores guide the model in identifying the most relevant features for the current task.

3. Related work

This section reviews recent advancements in affective computing using multimodal biomedical data, as well as challenges specific to EEG-based emotion recognition. We organize the literature into two subsections: multimodal frameworks integrating EEG with other physiological signals (e.g., GSR, eye movement) and approaches addressing

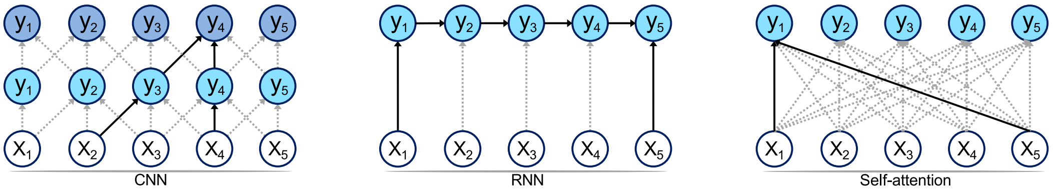


Fig. 3. Simplified comparison of neuron interactions in CNNs, RNNs, and the attention mechanism.

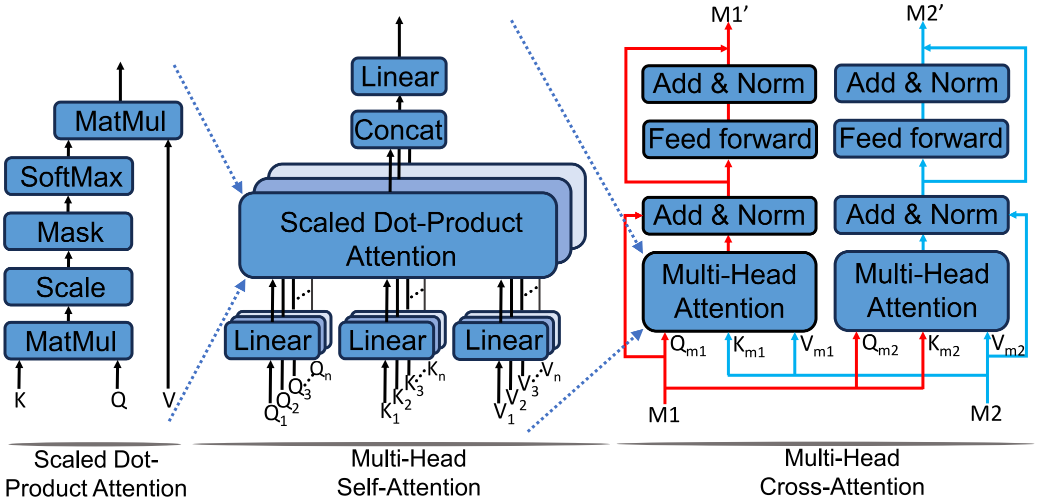


Fig. 4. Illustrating the relationship between scaled dot-product attention, multi-head self-attention, and multi-head cross-attention. Scaled dot-product attention forms the basic operation, while self-attention captures intra-sequence dependencies and cross-attention models interactions across different modalities.

EEG-specific challenges, including individual differences, spurious correlations, and incomplete modality representations. We critically evaluate these works, highlighting gaps that our proposed Transformer-Based Emotion Recognition (TEREE) model aims to address by simultaneously tackling all three challenges.

3.1. *Affective computing using multimodal biomedical data*

Multimodal frameworks that combine EEG with complementary signals such as galvanic skin response (GSR) or eye movement (EM) have shown significant potential for emotion recognition, yet they face persistent challenges. For instance, graph convolutional networks equipped with attention mechanisms have been applied to detect depression-related neural patterns in EEG data, yielding promising results but still struggling with inter-subject variability that limits generalization across diverse populations [36]. Similar efforts integrating EEG and GSR for stress detection in virtual reality environments have reported performance degradation when modalities are misaligned due to ambiguities in cross-modal correlations [37]. The value of eye-tracking for studying attentional biases in depression has also been highlighted, emphasizing EM's role as a complementary modality; however, this approach does not address incomplete EEG data, which frequently occurs in real-world applications due to hardware malfunctions or noise [38].

Transformer-based multimodal architectures have recently advanced the field by integrating diverse biomedical signals. One proposed solution introduces a BiProjection mechanism that unifies EEG and EM into a shared representation space, thereby improving emotion recognition accuracy [39]. Another approach employs self-distillation within transformers to reduce EEG signal noise, consequently enhancing cross-subject performance [40]. Nonetheless, these methods often fail to account for substantial inter-subject variability, which hinders generalization when training data is limited [41]. A unified transformer framework has also been introduced for joint emotion and intent recognition; however, its dependency on balanced multimodal data overlooks the prevalent challenge of incomplete modality representation, such as missing EEG channels [42,43]. Moreover, transformer-based models tailored for wearable emotion recognition exhibit limited robustness due to session-specific artifacts that produce spurious correlations [44]. Overall, the existing literature often addresses these challenges independently, lacking a holistic approach that concurrently mitigates individual differences, spurious correlations, and incomplete data—all of which are critical for the practical deployment of EEG-based emotion recognition systems.

3.2. *EEG challenges in emotion recognition*

EEG-based emotion recognition is promising due to its ability to directly capture neural correlates of emotional states. However, challenges such as individual differences, spurious correlations, and incomplete modality representations continue to hinder model accuracy and generalization.

To address individual differences, a model-agnostic meta-learning framework has been developed to rapidly adapt to individual variations via one-shot learning, leading to improved cross-subject generalization [45]. Research has also shown that variations in individual theta frequency are correlated with inhibitory control efficiency, with stronger connections observed in the superior temporal and inferior frontal gyri, reflecting neural dynamics that drive behavioral variability [46]. To enhance cross-subject performance, a spatio-temporal feature-fused convolutional graph attention network with multi-head attention has been introduced [47]. A domain-adaptive approach based on a cross-attention dilated causal convolutional neural network integrated with a domain discriminator has also been proposed to reduce both inter- and intra-subject variability [48]. In addition, contrastive learning has been employed to enable unsupervised multi-source domain adaptation by aligning conditional distributions across domains [49]. A multi-task self-supervised learning framework incorporating channel and frequency masking has likewise proven effective in mitigating individual and modality-related variations [50]. Nevertheless, large domain shifts in EEG data distributions remain a major obstacle, potentially limiting the generalizability of these approaches [25].

Spurious correlations – often introduced by session-specific artifacts or environmental noise – have also been extensively investigated. For example, constrained generalized Gaussian filters have been used to suppress such correlations in EEG signals [51]. A completeness-induced adaptive broad learning model has been proposed to generate comprehensive EEG and EM representations, thereby reducing inter-modality spurious correlations [24]. To minimize artefactual dependencies between pre-stimulus oscillations and behavioral responses, an end-point corrected Hilbert transform has been applied [52]. Filtering spurious EEG channel connections has also been addressed through an Adjacency-Explainable Graph Neural Network (AEG), which maximizes mutual information with true emotional states [53]. Techniques involving cross-scenario and cross-subject adaptation using adversarial learning and multi-kernel maximum mean discrepancy have further improved robustness [54], while an attentive simple graph convolutional network has been designed to mitigate sex-specific correlations [55]. Despite their effectiveness, these approaches tend to address narrow categories of spurious correlations and often overlook broader session-level artifacts.

Another critical challenge is incomplete modality representation, particularly in the presence of missing or corrupted EEG channels. This has been addressed through the use of an LSTM-enhanced multiview dynamical emotion graph that adaptively updates EEG graph structures [56]. Multimodal physiological signal fusion using self-attention and cross-attention transformers has also been explored to obtain more reliable EEG representations [57]. A graph convolutional network based on contrastive learning has been proposed to capture emotional features shared across modalities, thereby enhancing resilience to data loss [58]. More integrative frameworks have also emerged. For example, contrastive learning has been applied to extract invariant EEG features across multiple domains, effectively addressing challenges related to individual variability, spurious correlations, and data incompleteness [59]. Wang et al. [60] proposed a hierarchical spatial transformer that captures long-range dependencies from electrode to brain-region level. It outperforms CNNs and RNNs on DEAP and MAHNOB-HCI by emphasizing key brain regions. In [61], a spatio-temporal feature fusion network combining CNN-based spatial maps and temporal features with Bi-LSTM fusion is introduced. Improvements in signal representation have also been achieved by combining spatial graph-BERT and temporal LSTM in a spatio-temporal graph BERT model [62]. Furthermore, contrastive reinforced transfer learning, which uses reinforcement learning to dynamically select transferable EEG features, has shown promising results [63]. Finally, a multi-class transfer learning framework incorporating source label adaptive correction and nuclear norm maximization has been proposed to enhance model robustness [64].

Despite these advancements, the literature rarely offers unified solutions that simultaneously address individual differences, spurious correlations, and incomplete modality representation—especially in multimodal contexts. Many existing methods are tailored to unimodal EEG data, which limits their effectiveness in comprehensive frameworks such as TEREE that integrate EEG and EM for robust emotion recognition across varied scenarios.

4. Transformer-based emotion recognition using biomedical data

Fig. 5 illustrates the architecture of the proposed model, which consists of four main components: (i) data projection, (ii) multi-head cross-attention, (iii) a stack of self-attention blocks, and (iv) the classification head. Each component is described in detail in the following subsections.

4.1. Data projection

In the data projection stage, EEG and EM data are tokenized into two separate streams and projected linearly. Studies have shown that the spatial, spectral, and temporal aspects of EEG data each provide valuable information about individuals' emotional states. By analyzing all three aspects together, it is possible to minimize the effect of individual differences in data. To achieve this, the dimensionality of 1D EEG data is expanded into 2D to preserve all three aspects. This approach enhances the true correlations between samples within the same class and reduces the impact of inter-subject variability [65–67].

As shown in Fig. 6, the vertical axis (top to bottom) represents the channels (spatial) from 1 to 62. The frequency bands are separated using a bandpass filter, including Delta (0.5–4 Hz), Theta (4–8 Hz), Alpha (8–13 Hz), Beta (13–30 Hz), and Gamma (above 30 Hz). For all bands in each channel, 200 samples per second are recorded and converted into values from 0 to 255 to represent grayscale pixels (spectral), which are then arranged consecutively from left to right to encode the temporal dimension. For image tokenization and position embedding, we follow the Vision Transformer (ViT) approach [41]: the image is resized to 224 × 224 pixels, divided into 16 × 16 patches, linearly projected, and finally position embeddings are added to the tokens.

To mitigate potential quantization effects during the conversion of EEG signals into 2D images, we normalize the data prior to pixel mapping, which helps preserve the dynamic range. Furthermore, because the ViT-based attention mechanism emphasizes relative spatial–temporal–spectral patterns rather than absolute signal magnitudes, the model is less sensitive to scaling artifacts. For the EM stream, we explicitly use five feature categories: gaze X and Y coordinates (visual attention path), pupil diameter (arousal indicator), fixation duration, and saccade start–end times. These features are sampled at 200 Hz and normalized per subject to reduce inter-individual variability. Fixations and saccades are segmented using the dispersion-threshold algorithm provided with the SEED dataset. After preprocessing, EM features are linearly projected, followed by dropout and positional encoding, before being integrated via cross-attention with EEG representations. Dropout is applied before the position embedding operation to randomly remove some neurons during training, thereby reducing overfitting and mitigating spurious correlations between modalities [68].

4.2. Self-attention block

Although the self-attention block is the third stage of the model, we explain it first since multi-head cross-attention is essentially a combination of multiple self-attention mechanisms.

Self-attention is crucial for modeling the relationships between EEG and EM signals, as it captures long-range dependencies within fused multimodal representations. In our model, self-attention operates on tokenized sequences derived from the fusion of spatio–temporal–spectral EEG features and EM signals. The attention scores are computed as:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V. \tag{1}$$

where Q, K, and V are linear projections of the input, and d_k is the dimensionality of the key vectors. This mechanism allows each token to attend to all others, enabling effective integration of multimodal information.

To further refine the fused representations, multiple self-attention layers are stacked. Residual connections and layer normalization improve gradient stability and facilitate training deeper architectures:

$$X^{l+1} = \mathrm{MLP}\left(\mathrm{Norm}\left(\mathrm{Attention}(Q, K, V) + X^{l}\right)\right) \tag{2}$$

where X^l denotes the input at layer l, and the multi-layer perceptron (MLP) introduces additional feature transformations. By leveraging self-attention, the model captures both intra-modal dependencies within EEG and EM signals and inter-modal relationships, ensuring robust multimodal feature fusion.

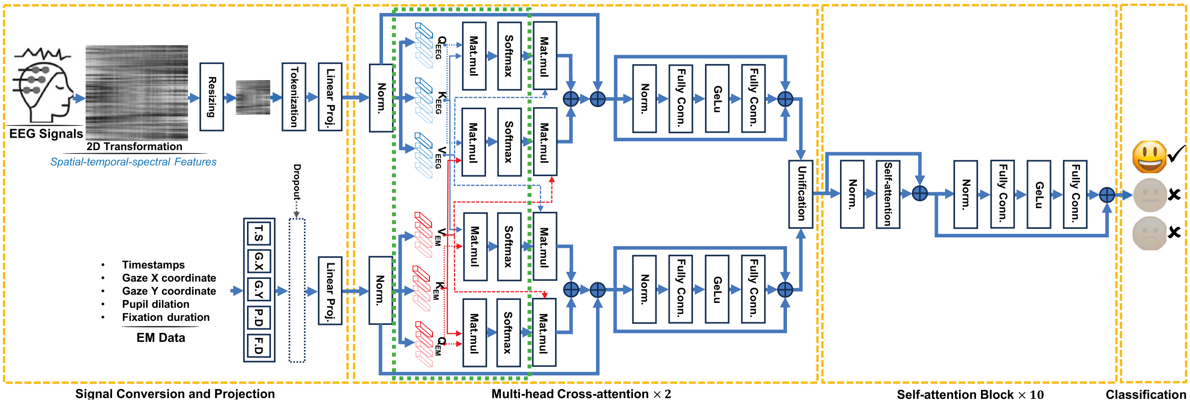


Fig. 5. Overall workflow of the proposed model for emotion recognition across negative, neutral, and positive states. The model comprises four main stages: (i) signal conversion and projection (EEG 1D→ 2D spatial–temporal–spectral mapping and EM feature projection); (ii) multi-head cross-attention (× 2) for multimodal fusion; (iii) a stack of self-attention blocks (× 10) for global representation learning; and (iv) classification. The green dotted line marks the Bayesian Spurious Correlation Minimization stage, where attention weights are modeled in a Bayesian manner and regularized to reduce session-specific spurious correlations.

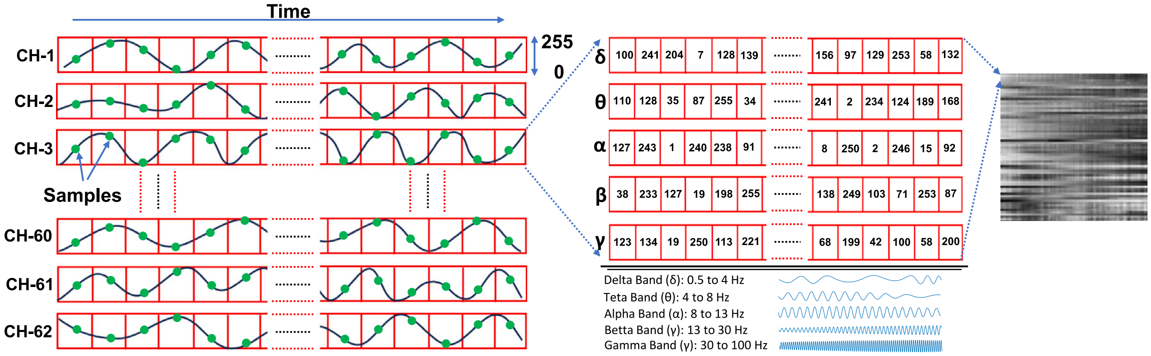

Time
CH-1
CH-2
CH-3
Samples
CH-60
CH-61
CH-62
255
0
δ 100 241 204 7 128 139 156 97 129 253 58 132
θ 110 128 35 87 255 34 241 2 234 124 189 168
α 127 243 1 240 238 91 8 250 2 246 15 92
β 38 233 127 19 198 255 138 249 103 71 253 87
γ 123 134 19 250 113 221 68 199 42 100 58 200
Delta Band (δ): 0.5 to 4 Hz
Teta Band (θ): 4 to 8 Hz
Alpha Band (α): 8 to 13 Hz
Betta Band (γ): 13 to 30 Hz
Gamma Band (γ): 30 to 100 Hz

Fig. 6. Illustration of EEG signal conversion into 2D images, preserving spatial information (channels), spectral information (frequency-band intensity), and temporal information (time).

4.3. Multi-head cross-attention

Multi-head cross-attention plays a central role in our model by integrating EEG and EM signals, enabling the network to learn cross-modal relationships. It aligns features between EEG and EM, ensuring effective multimodal fusion.

Given a query (Q) from one modality (e.g., EEG) and key (K) and value (V) from another (e.g., EM), the attention mechanism computes relevance scores as:

$$\text{Attention}(Q_{\text{EEG}}, K_{\text{EM}}, V_{\text{EM}}) = \text{softmax}\left(\frac{Q_{\text{EEG}} K_{\text{EM}}^{\text{T}}}{\sqrt{d_k}}\right) V_{\text{EM}} \tag{3}$$

To improve integration, we employ a bidirectional cross-attention mechanism, where EEG and EM mutually influence each other. This ensures that the learned representations capture both EEG-informed gaze patterns and gaze-informed neural responses. The attention outputs from cross-modal interactions are then combined with intra-modal self-attention, refining the feature space and reducing spurious correlations.

By leveraging multi-head cross-attention, the model effectively aligns spatial, temporal, and spectral features across modalities, enabling robust emotion recognition even in the presence of missing or noisy data.

4.3.1. Integrating BSCM into bidirectional multimodal attention

Algorithm 1 in TEREE integrates cross-attention transformers with variational autoencoder (VAE) losses to address spurious correlations and incomplete modality issues in EEG and EM data. The cross-attention mechanism aligns EEG's neural patterns with EM's behavioral cues (e.g., gaze coordinates, fixation durations), capturing inter-modal dependencies while filtering session-specific noise. VAE losses enforce a probabilistic latent space, enabling robust reconstruction of missing modalities by learning shared representations. Experiments on the SEED dataset confirm that TEREE achieves superior accuracy under missing EM data compared to unimodal models, highlighting the synergy of cross-attention and VAE in handling incomplete data and reducing overfitting.

The Bayesian Spurious Correlation Minimization (BSCM) aims to minimize the influence of spurious correlations by treating the model's decision function as a Bayesian posterior distribution rather than a fixed deterministic function. In this context, bidirectional multimodal attention learns posterior distributions over attention weights instead of fixed values. Formally, standard Invariant Risk Minimization seeks a representation Φ that remains consistent across environments $e \in \mathcal{E}$, with the optimal classifier w satisfying:

$$w \in \arg\min_{w} \sum_{e \in \mathcal{E}} R^{e}(w \circ \Phi), \tag{4}$$

where R^{e} represents the empirical risk in environment e. However, a deep model may still capture spurious correlations due to overfitting. To mitigate this, BSCM introduces Bayesian learning, which models the classifier w as a distribution rather than a fixed parameter:

$$p(w|D) \propto p(D|w)p(w), \tag{5}$$

where: $p(w|D)$ is the posterior distribution of the classifier, $p(D|w)$ is the likelihood of data given the model, $p(w)$ is the prior distribution over model parameters. As shown in Algorithm 1 BSCM incorporates this into bidirectional multimodal attention by sampling weights from a learned posterior rather than using deterministic attention scores.

Algorithm 1 Bayesian Spurious Correlation Minimization (BSCM) for Bidirectional Multimodal Attention in TEREE

```
Inputs:
    $X_{\mathrm{EEG}}$: EEG data
    $X_{\mathrm{EM}}$: Eye movement data
    $\Theta$: prior parameters $(\mu, \sigma^2)$ for attention weights
    $\alpha$: learning rate
    $N$: number of training epochs
    $\lambda$: KL regularization coefficient
Outputs:
    $M$: trained multimodal transformer with Bayesian attention
Initialization:
    Initialize variational parameters $(\mu, \rho)$ for attention weights $W$ with prior $\mathcal{N}(\mu, \sigma^2 I)$.
    Construct model architecture with bidirectional cross-attention and self-attention blocks.
for epoch $n = 1$ to $N$ do
    for each minibatch $(x_{\mathrm{EEG}}, x_{\mathrm{EM}}, y)$ do
        Bayesian Weight Sampling:
            Compute $\sigma = \log(1 + \exp(\rho))$.
            Sample $\epsilon \sim \mathcal{N}(0, I)$ and set $W^* = \mu + \sigma \odot \epsilon$.
        Forward Pass:
            Encode EEG and EM into embeddings.
            Apply bidirectional cross-attention with sampled weights:
                $A^*_{\mathrm{EEG}\to\mathrm{EM}} = \mathrm{softmax}\left(\frac{Q_{\mathrm{EEG}} W^* K_{\mathrm{EM}}^\top}{\sqrt{d}}\right) V_{\mathrm{EM}}$
                $A^*_{\mathrm{EM}\to\mathrm{EEG}} = \mathrm{softmax}\left(\frac{Q_{\mathrm{EM}} W^* K_{\mathrm{EEG}}^\top}{\sqrt{d}}\right) V_{\mathrm{EEG}}$
            Fuse attention outputs and pass through stacked self-attention + MLP layers.
        Loss Computation:
            Prediction loss: $\mathcal{L}_{\mathrm{pred}} = -\log p(y|h)$
            KL term: $\mathcal{L}_{\mathrm{KL}} = D_{\mathrm{KL}}(q(W|D) \,\|\, p(W))$
            Total loss: $\mathcal{L} = \mathcal{L}_{\mathrm{pred}} + \lambda \mathcal{L}_{\mathrm{KL}}$
        Backward + Update:
            Backpropagate $\nabla \mathcal{L}$.
            Update $(\mu, \rho)$ and model parameters via Adam with lr $\alpha$.
            Apply Bayesian dropout to attention layers for regularization.
    end for
    Evaluate model on validation set, record metrics.
end for
Return: trained model $M$.
```

4.4. Classification

The input to the classification head is the latent array obtained after processing through multiple bidirectional multimodal attention blocks and self-attention layers. This array contains the integrated and refined features extracted from both EEG and EM modalities. The classification head transforms this latent representation into logits, which are the raw, unnormalized scores for each class. These logits are then converted into class probabilities using a softmax activation function. During training, categorical cross-entropy loss is computed between the predicted logits and the ground-truth labels. Because a complete set of EEG data collected during a five-minute session, together with the corresponding EM signals, is processed at each stage, the self-attention block delivers a comprehensive multimodal representation to the classification stage.

5. Dataset description and experimental setup

Two well-known multimodal emotional datasets, SEED [69] and SEED-FRA [28], were used in our experiments, both involving EEG and eye movement signals. As shown in Table 1, although the SEED dataset originally contains 15 participants (referred to as subjects), simultaneous EEG and EM recordings are available for only 12; therefore, our experiments were conducted on this subset to ensure consistency across modalities.

In SEED, each subject viewed a total of 15 video clips during each session, categorized into three emotional valences: five positive, five neutral, and five negative. After an interval of approximately one week, the subjects participated in a second session with a new set of 15 video clips, followed by a third session, resulting in three experimental sessions for each subject. In SEED-FRA, eight subjects each watched 21 videos across three sessions.

Participants viewed the clips in a controlled environment while their EEG and eye-tracking data were recorded. Each trial lasted five minutes, sampled at 200 Hz, and was labeled as negative, neutral, or positive. EEG signals were recorded using a 62-channel ESI NeuroScan system, and the raw EEG data files were used. In the SEED dataset, eye-movement data were stored in an Excel file containing timestamps, gaze X and Y coordinates (e.g., X: 512, Y: 384), pupil dilation (e.g., 3.5 mm), fixation duration (e.g., 200 ms), and saccade start/end times (e.g., 10:01:23.456 to 10:01:23.789). These annotations provide behavioral insights into emotional responses, complementing EEG data and enhancing the accuracy of the proposed TEREE model.

The proposed TEREE model was implemented in PyTorch. Each transformer block consists of a multi-head self-attention or cross-attention layer followed by a feed-forward MLP. The MLP is composed of two fully connected layers with a hidden dimension of 1024, separated by a GeLU activation and followed by dropout (rate = 0.1). Layer normalization is applied before each attention and MLP sub-layer.

For optimization, we employed the Adam optimizer with weight decay set to 1×10^{-4}. A grid search was performed over hyperparameters to select the optimal configuration. The best performance was achieved with a learning rate of 1×10^{-4}, a batch size of 32, and 40 training epochs. To stabilize training, learning rate warm-up was applied during the first 10% of epochs, followed by cosine annealing decay.

For the BSCM integration, the posterior parameters were initialized with $\mu = 0$ and $\rho = \log(\exp(0.1) - 1)$, corresponding to a small initial variance, following standard practice in variational Bayesian neural networks. The KL regularization coefficient λ was selected via grid search on the validation set, and a linear annealing schedule was applied during the first 10 epochs to stabilize training. These settings ensure both stable optimization and reproducibility of our approach.

Dropout (0.1) and Bayesian dropout within attention weights were used as regularizers. For classification, categorical cross-entropy loss was employed, while for BSCM integration, an additional KL-divergence regularization term was included.

For our experimental environment, we utilized a workstation equipped with an Intel Core i7 CPU, 48 GB of RAM, and an NVIDIA GeForce GTX 1080 GPU.

To assess our model's performance, we employed two evaluation paradigms: one-to-one and multi-to-one. As described in Algorithm 2, in the one-to-one paradigm, the EEG and EM signals (labeled data) from a single subject (e.g., Subject 1) are used as the source domain, while the EEG data from each remaining subject (e.g., Subject 2) serves as a separate target domain. In simpler terms, the model is trained on data from Subject 1 and evaluated on data from Subjects 2 through 12. As shown in Algorithm 3, in the multi-to-one paradigm, data from several subjects are combined to improve learning and performance on a single target subject. The goal is to leverage the diversity and quantity of data from multiple sources to build a more robust model that can generalize effectively to the target subject.

Algorithm 2 One-to-One Transfer Paradigm

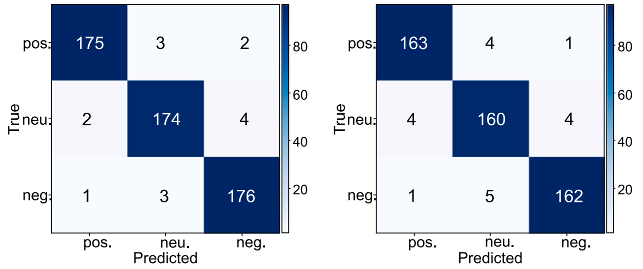
```
Require: Dataset $D$ with $N$ subjects, Model $M$
for each subject $i \in \{1, 2, \ldots, N\}$ do
    Assign subject $i$'s data as the test set $D_{\mathrm{test}}$
    Randomly select data from another subject $j \neq i$ as the training set $D_{\mathrm{train}}$
    Train the model $M$ on $D_{\mathrm{train}}$
    Evaluate $M$ on $D_{\mathrm{test}}$ and record the performance
end for
Compute and return the average performance metrics across all $N$ subjects
```

Algorithm 3 Multi-To-One Transfer Paradigm

```
Require: Dataset $D$ with $N$ subjects, Model $M$
for each subject $i \in \{1, 2, \ldots, N\}$ do
    Assign subject $i$'s data as the test set $D_{\mathrm{test}}$
    Use data from all other $N - 1$ subjects as the training set $D_{\mathrm{train}}$
    Train the model $M$ on $D_{\mathrm{train}}$
    Evaluate $M$ on $D_{\mathrm{test}}$ and record the performance
end for
Compute and return the average performance metrics across all $N$ subjects
```

Table 1
Summary of the multimodal datasets employed in this study, including SEED and SEED-FRA, with details on participants, sessions, recording modalities, and experimental settings.

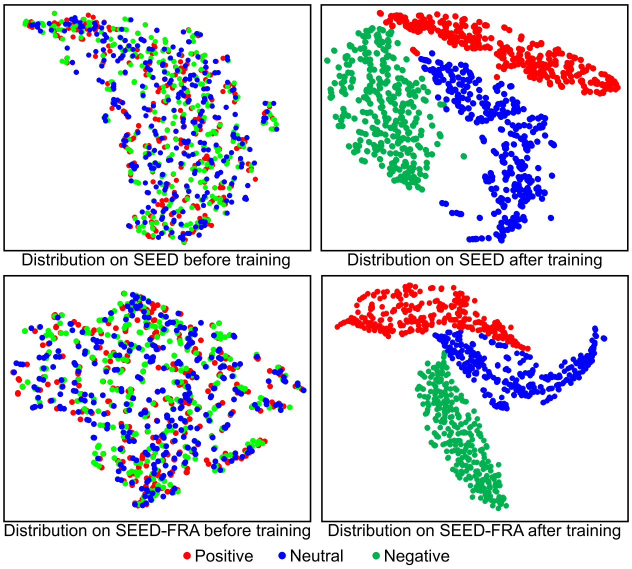
Dataset	SEED	SEED-FRA
Number of Participants	15 (12 both in EEG and EM)	8
Sessions	3 (15 clips per session)	3 (21 clips per session)
Emotion Categories	Positive, Neutral, Negative	Positive, Neutral, Negative
EEG System	62-channel ESI NeuroScan	62-channel ESI NeuroScan
Eye Movement Data	Timestamps, gaze X and Y, pupil dilation, fixation duration, saccades	
Eye Movement Data	5 min per video clip	
Sampling Rate	200 samples per second	



SEED-session1-3

SEED-FRA-session1-3

Fig. 7. Overall confusion matrices for the SEED and SEED-FRA datasets across three sessions in the multi-to-one transfer paradigm, illustrating classification performance across negative, neutral, and positive emotional states.


Distribution on SEED before training

Distribution on SEED after training

Distribution on SEED-FRA before training

Distribution on SEED-FRA after training

Fig. 8. t-SNE visualization of EEG feature distributions for Subject 1 under the multi-to-one setting. Pre-training distributions show substantial overlap among emotional classes, while post-training with the proposed TEREE model yields compact intra-class clusters and clearer inter-class separation across SEED and SEED-FRA datasets.

6. Experimental results

The confusion matrix in Fig. 7 shows both the correct classifications and misclassifications made by the model. The proposed model performs best in emotion recognition during the first session of the SEED dataset and the second session of the SEED-FRA dataset. However, the differences are not very significant. The model is better at identifying positive and negative emotions than neutral emotions, indicating that it more easily detects brain patterns associated with emotional states. This may be because the brain remains influenced by the previous emotional state and does not fully return to neutral before the next emotion is induced.

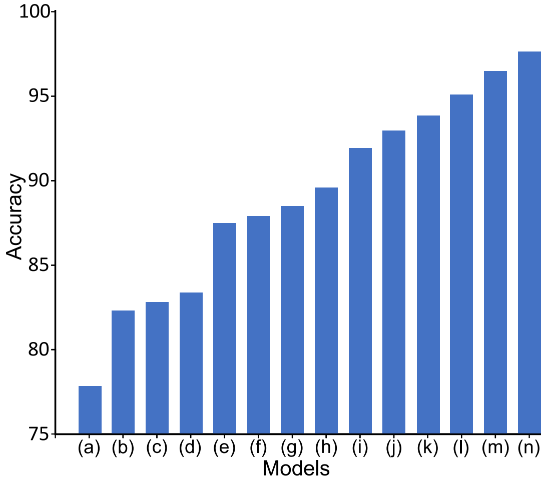


Fig. 9. Comparison of the average accuracy of different models across three sessions for all subjects, on the SEED dataset sorted in ascending order. (a):[58], (b):[50], (c):[45], (d):[25], (e):[64], (f):[70], (g):[59], (h):[62], (i):[56], (j):[26], (k):[63], (l):[47], (m):[24], (n): Ours.

To further evaluate our model's effectiveness, we use Subject 1 as the target subject and apply t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize the data distributions before and after training.

Fig. 8 shows the t-SNE visualization of EEG feature distributions. Before training, samples from different emotional states exhibit substantial overlap, making class boundaries indistinguishable. After training with the proposed TEREE model, the feature space becomes more structured: intra-class samples form compact clusters, while inter-class separation is clearly enhanced. This improvement is particularly evident in the distinct boundaries between positive, neutral, and negative emotions. These results demonstrate that the model, together with BSCM regularization, effectively learns discriminative and generalizable representations, thereby improving emotion classification performance. In both SEED and SEED-FRA datasets, emotion samples after training show tighter intra-class clustering and clearer inter-class separation, indicating that the model successfully extracts the most informative features from the source domain and adapts them to individual differences in the target domain.

Tables 2 and 3 report the performance of our model compared with the approaches in [24,25], and [26] under the one-to-one and multi-to-one transfer paradigms on the SEED dataset. Similarly, Tables 4 and 5 present the corresponding results on the SEED-FRA dataset.

These comparisons evaluate the ability of models to recognize emotions across different subjects. Accuracy was used as the evaluation metric, averaged over tasks from each session. As illustrated in Fig. 9, our proposed TEREE model consistently outperforms baseline methods, particularly in terms of average recognition accuracy across sessions.

On the SEED dataset, for example, our model achieved average emotion recognition accuracies of 93.5%, 96.2%, and 91.9% under the one-to-one transfer paradigm, and 97.7%, 97.2%, and 97.1% under the

Table 2
The emotion recognition accuracy (%) of the cross-subject experiment under the one-to-one transfer paradigm in three sessions on the SEED dataset. Subject 1 was used as a training subject. Other subjects were used one by one for testing. The Δ value represents the difference between the highest and lowest predicted accuracy values across different subjects.

Session1	Min and Max	Avg.(%)	Δ
[25]	Min: Sub2 (60.4%), Max: Sub12 (93.1%)	81.2	38.6
[24]	Min: Sub10 (75.1%), Max: Sub11 (99.0%)	87.1	23.8
[26]	Min: Sub10 (75.6%), Max: Sub9 (91.1%)	85.8	15.5
[31]	Min: Sub10 (70.5%), Max: Sub8 (92.7%)	83.7	22.5
Ours	Min: Sub10 (86.6%), Max: Sub11 (100.0%)	93.5	13.3
Session2	Min and Max	Avg.(%)	Δ
[25]	Min: Sub10 (72.4%), Max: Sub3 (92.6%)	82.4	25.5
[24]	Min: Sub10 (86.2%), Max: Sub5 (100.0%)	94.0	13.7
[26]	Min: Sub4 (63.5%), Max: Sub7 (100.0%)	88.1	36.4
[31]	Min: Sub10 (70.5%), Max: Sub8 (92.7%)	83.7	22.5
Ours	Min: Sub10 (93.3%), Max: Sub2 (100.0%)	96.2	6.6
Session3	Min and Max	Avg.(%)	Δ
[25]	Min: Sub10 (72.5%), Max: Sub3 (89.9%)	81.9	27.7
[24]	Min: Sub11 (73.9%), Max: Sub4 (96.4%)	86.9	22.4
[26]	Min: Sub10 (72.9%), Max: Sub7 (93.1%)	82.7	20.1
[31]	Min: Sub10 (70.5%), Max: Sub8 (92.7%)	83.7	22.5
Ours	Min: Sub11 (80.0%), Max: Sub8 (93.3%)	91.9	13.3

Table 3
The emotion recognition accuracy (%) of the cross-subject experiment under the multi-to-one transfer paradigm in three sessions on the SEED dataset. For each subject, the model is trained on other subjects and tested on that subject. The Δ value represents the difference between the highest and lowest predicted accuracy values across different subjects.

Session1	Min and Max	Avg.(%)	Δ
[25]	Min: Sub2 (72.3%), Max: Sub12 (93.2%)	82.9	20.9
[24]	Min: Sub1 (90.5%), Max: Sub8 (100.0%)	96.9	9.5
[26]	Min: Sub10 (81.4%), Max: Sub12 (100.0%)	92.0	18.5
Ours	Min: Sub1 (93.3%), Max: Sub8 (100.0%)	97.7	6.6
Session2	Min and Max	Avg.(%)	Δ
[25]	Min: Sub4 (64.9%), Max: Sub6 (100.0%)	81.55	35.0
[24]	Min: Sub11 (88.3%), Max: Sub7 (100.0%)	95.3	11.6
[26]	Min: Sub10 (80.6%), Max: Sub12 (100.0%)	92.4	19.3
Ours	Min: Sub9 (93.3%), Max: Sub5 (100.0%)	97.2	6.6
Session3	Min and Max	Avg.(%)	Δ
[25]	Min: Sub2 (72.2%), Max: Sub3 (98.4%)	85.2	26.2
[24]	Min: Sub1 (81.2%), Max: Sub6 (100.0%)	96.4	18.7
[26]	Min: Sub2 (81.3%), Max: Sub9 (100.0%)	93.7	18.6
Ours	Min: Sub10 (86.6%), Max: Sub3 (100.0%)	97.1	13.3

Table 4
The emotion recognition accuracy (%) of the cross-subject experiment under the one-to-one transfer paradigm in three sessions on the SEED-FRA dataset. Subject 1 was used as a training subject. Other subjects were used one by one for testing. The Δ value represents the difference between the highest and lowest predicted accuracy values across different subjects.

Session1	Min and Max	Avg.(%)	Δ
[25]	Min: Sub2 (60.4%), Max: Sub8 (87.5%)	83.9	27.1
[24]	Min: Sub8 (78.3%), Max: Sub5 (93.1%)	93.7	15.2
[26]	Min: Sub5 (67.3%), Max: Sub2 (85.8%)	76.8	18.5
[31]	Min: Sub2 (84.1%), Max: Sub6 (86.8%)	90.7	12.7
Ours	Min: Sub8 (87.5%), Max: Sub5 (100.0%)	94.6	12.5
Session2	Min and Max	Avg.(%)	Δ
[25]	Min: Sub4 (67.1%), Max: Sub3 (92.6%)	79.8	25.5
[24]	Min: Sub4 (87.4%), Max: Sub2 (100.0%)	97.6	12.6
[26]	Min: Sub3 (56.5%), Max: Sub4 (94.6%)	78.2	38.1
[31]	Min: Sub2 (84.1%), Max: Sub2 (86.8%)	90.7	12.7
Ours	Min: Sub7 (87.5%), Max: Sub3 (100.0%)	98.1	12.5
Session3	Min and Max	Avg.(%)	Δ
[25]	Min: Sub2 (65.2%), Max: Sub7 (92.9%)	76.9	27.7
[24]	Min: Sub3 (77.5%), Max: Sub4 (96.4%)	86.1	18.9
[26]	Min: Sub3 (72.0%), Max: Sub4 (85.7%)	79.8	13.7
[31]	Min: Sub2 (84.1%), Max: Sub6 (86.8%)	90.7	12.7
Ours	Min: Sub3 (87.5%), Max: Sub8 (100.0%)	90.1	12.5

Table 5
The emotion recognition accuracy (%) of the cross-subject experiment under the multi-to-one transfer paradigm in three sessions on the SEED-FRA dataset. For each subject, the model is trained on other subjects and tested on that subject. The Δ value represents the difference between the highest and lowest predicted accuracy values across different subjects.

Session1	Min and Max	Avg.(%)	Δ
[25]	Min: Sub8 (60.6%), Max: Sub6 (82.8%)	74.2	22.2
[24]	Min: Sub2 (76.5%), Max: Sub4 (94.7%)	91.9	18.2
[26]	Min: Sub2 (75.2%), Max: Sub6 (91.4%)	89.1	16.2
Ours	Min: Sub8 (87.5%), Max: Sub6 (100.0%)	95.1	12.5
Session2	Min and Max	Avg.(%)	Δ
[25]	Min: Sub2 (65.7%), Max: Sub5 (77.5%)	71.2	11.8
[24]	Min: Sub6 (82.7%), Max: Sub4 (97.7%)	92.2	15.0
[26]	Min: Sub6 (70.7%), Max: Sub1 (88.7%)	85.5	18.0
Ours	Min: Sub5 (87.5%), Max: Sub8 (100.0%)	98.8	12.5
Session3	Min and Max	Avg.(%)	Δ
[25]	Min: Sub1 (67.0%), Max: Sub5 (85.2%)	75.0	18.2
[24]	Min: Sub8 (82.3%), Max: Sub4 (95.0%)	86.4	12.7
[26]	Min: Sub2 (77.2%), Max: Sub4 (97.6%)	85.0	20.4
Ours	Min: Sub5 (87.5%), Max: Sub4 (100.0%)	94.9	12.5

multi-to-one paradigm. Similar trends were observed on the SEED-FRA dataset. Notably, the model consistently performed better in multi-to-one experiments compared to one-to-one experiments across both datasets.

The Δ value reported in each table represents the difference between the highest and lowest accuracy values across different subjects. For our model, this value is the smallest across all sessions, indicating that the choice of training subject has less impact on accuracy. Compared to prior methods [24–26], this suggests that individual differences exert a smaller influence on the performance of our model.

To investigate how EEG and EM signals complement each other, we conducted ablation studies (Table 6). The model was evaluated under four conditions: using only EEG, only EM, and the fusion of both modalities, in both one-to-one and multi-to-one settings. All reported results represent mean accuracy over five independent runs, with standard deviations consistently below 5%, confirming the stability of the model.

The results demonstrate that multimodal fusion significantly improves performance. For example, in Session 1 of the SEED dataset (one-to-one setting), the model achieved 93.5% accuracy, compared to 91.8% with EEG alone and 73.4% with EM alone. Across all experiments, EEG-only models generally performed better than EM-only models. However, in every session, combining EEG and EM yielded the highest accuracy, confirming that the modalities were effectively integrated and that spurious correlations did not impair performance. Moreover, by feeding the entire transformed EEG image into the transformer at once, the issue of incomplete modality representations was alleviated. As summarized in Table 6, the model's accuracy using EEG alone remained above 85% across all sessions.

Fig. 10 illustrates the training and test accuracy of the model with and without the BSCM mechanism over 50 epochs. As observed in the figure, although the model with BSCM initially learns slightly slower, the final accuracy remains largely similar to the model without BSCM. Additionally, the gap between test and training accuracy is reduced in the presence of BSCM, indicating a decrease in overfitting.

We analyze the channel attention by visualizing the weights learned by the proposed model. Fig. 11 shows the distribution of channel attention across the scalp for positive, neutral, and negative samples from the SEED dataset.

The red areas on the scalp indicate channels with higher attention weights, whereas the blue areas correspond to channels with lower weights. The relatively small extent of red regions compared to blue

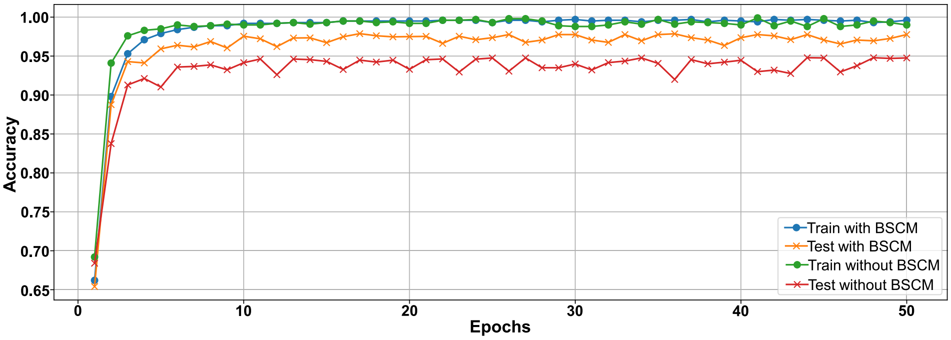


Fig. 10. Training and test accuracy in the multi-to-one configuration using the SEED dataset.

Table 6
Comparison of average accuracy (%) for the proposed TEREE model using EEG-only, EM-only, and fused EEG+EM modalities across one-to-one and multi-to-one transfer settings. Results highlight the performance gains achieved through multimodal fusion.

Paradigms		One-to-One			Multi-to-One		
Modality		EEG	EM	Fusion	EEG	EM	Fusion
SEED	Ses.1	91.8	73.4	93.5	86.1	84.5	**97.7**
	Ses.2	89.8	82.4	96.2	86.2	84.8	**97.2**
	Ses.3	85.7	80.6	91.9	85.9	85.0	**97.1**
Modality		EEG	EM	Fusion	EEG	EM	Fusion
SEED-FRA	Ses.1	87.2	77.8	94.6	91.6	85.7	**95.1**
	Ses.2	92.2	86.9	98.1	92.8	79.1	**98.8**
	Ses.3	87.8	76.1	90.1	90.0	78.9	**94.9**

suggests that only a limited number of EEG channels are strongly correlated with emotional states, while many others contribute less. Notably, the temporal lobes exhibit greater activation when subjects view both positive and negative videos. In addition, a stronger hemispheric imbalance is observed during negative video viewing, whereas the distribution of active regions appears more balanced when viewing neutral videos.

7. Conclusions

This study introduced TEREE, a transformer-based model for emotion recognition using multimodal biomedical signals, specifically EEG and EM data. TEREE addresses three key challenges in emotion recognition—individual differences, spurious correlations, and incomplete modality representation. The model transforms EEG signals into spatio–temporal–spectral 2D representations and integrates EM features through a modified multi-head cross-attention mechanism, further enhanced by Bayesian spurious correlation minimization. This design enables the model to capture nuanced emotional patterns while reducing the impact of subject variability.

Experimental results demonstrate that TEREE not only improves classification accuracy in cross-subject settings but also minimizes performance disparities among individuals, ensuring robustness and generalizability. The fusion of EEG and EM data consistently enhances accuracy across all sessions, confirming effective modality integration while mitigating the influence of spurious correlations. Moreover, by incorporating complete transformed EEG data into the transformer, TEREE effectively alleviates the challenge of incomplete modality representation. Notably, the model achieves accuracies of 97.7% and 98.8% in multi-to-one transfer paradigms, underscoring its potential for advancing multimodal emotion recognition. By improving classification accuracy, reducing inter-subject variability, and overcoming modality-specific challenges, TEREE contributes to more reliable and practical applications in human–computer interaction and affective computing.

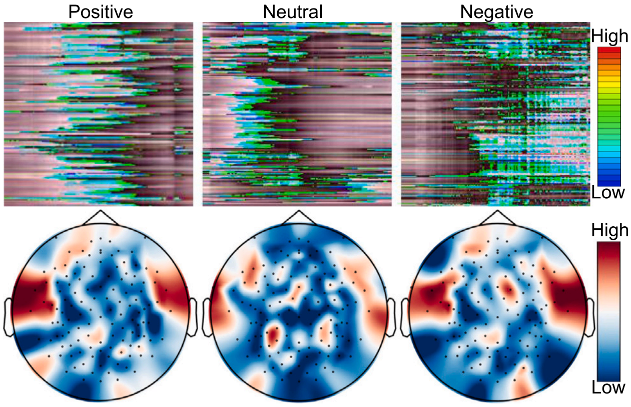


Fig. 11. (Top) Grad-CAM heatmaps of brain signal mappings for three video samples (Positive, Neutral, Negative) from Subject 1, trained under the multi-to-one configuration using the TEREE model on the SEED dataset. (Bottom) Topomap showing emotional neural patterns derived from aggregated frequency band estimates. Color intensity reflects importance, with "Low" indicating lower relevance and "High" indicating higher relevance. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Despite these promising results, several limitations remain. First, converting EEG signals into 2D images may still introduce minor quantization-related information loss, even though normalization and the ViT-based attention mechanism mitigate much of its effect. Second, hyperparameter choices, such as the KL regularization coefficient, require careful tuning, which may limit straightforward reproducibility across datasets. Third, while robustness to incomplete modalities was partially validated through ablation studies, further experiments with systematically varied levels of missing EEG channels or EM data are needed to provide stronger empirical evidence. Finally, although strong performance was demonstrated on SEED and SEED-FRA, broader evaluation on more diverse datasets and real-world scenarios is necessary to confirm generalizability. Future research could explore alternative signal representations that preserve the full dynamic range of EEG data, investigate adaptive strategies for hyperparameter scheduling, and extend the framework to additional modalities or larger-scale longitudinal studies.

CRediT authorship contribution statement

Nima Esmi: Writing – original draft, Visualization, Validation, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Asadollah Shahbahrami:** Writing – review & editing, Supervision, Resources, Conceptualization. **Georgi Gaydadjiev:** Writing – review & editing. **Peter de Jonge:** Writing – review & editing.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

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

Data availability

No new data were created or analyzed in this study.

References

[1] Mostafaei Sahar Hassanzadeh, Tanha Jafar, Sharafkhaneh Amir. A novel deep learning model based on transformer and cross modality attention for classification of sleep stages. J Biomed Inf 2024;157:104689.

[2] Helmy AbdelMoniem, Nassar Radwa, Ramdan Nagy. Depression detection for twitter users using sentiment analysis in English and Arabic tweets. Artif Intell Med 2024;147:102716. http://dx.doi.org/10.1016/j.artmed.2023.102716.

[3] Pfeffer Maximilian Achim, Ling Steve Sai Ho, Wong Johnny Kwok Wai. Exploring the frontier: Transformer-based models in EEG signal analysis for brain-computer interfaces. Comput Biol Med 2024;108705. http://dx.doi.org/10.1016/j.compbiomed.2024.108705.

[4] Balcioğlu Yavuz Selim, Çelik Ahmet Alkan, Altindağ Erkut. Sentiment analysis of reddit reviews on mobile gaming: Insights from the gaming community. Int J Hum-Comput Interact 2025;1–13. http://dx.doi.org/10.1080/10447318.2025.2464897.

[5] Yuan Aijia, Garcia Colato Edlin, Pescosolido Bernice, Song Hyunju, Samtani Sagar. Improving workplace well-being in modern organizations: A review of large language model-based mental health chatbots. ACM Trans Manag Inf Syst 2025;16(1):1–26. http://dx.doi.org/10.1145/3701041.

[6] Xu Haiyu, Guo Zhiwei, Saad Aldosary, Tolba Amr, Al-Dulaimi Anwer, Yu Keping, Rodrigues Joel JPC. Consumer QoE-aware cognitive semantic sentiment analysis via hybrid large models. IEEE Consum Electron Mag 2025;14(2):59–68. http://dx.doi.org/10.1109/MCE.2024.3437550.

[7] Acharya Madhav, Deo Ravinesh C, Barua Prabal Datta, Devi Aruna, Tao Xiaohui. EEGConvNeXt: A novel convolutional neural network model for automated detection of Alzheimer's disease and frontotemporal dementia using EEG signals. Comput Methods Programs Biomed 2025;108652. http://dx.doi.org/10.1016/j.cmpb.2025.108652.

[8] Jin Xiaofang, Xiao Jieyu, Jin Libiao, Zhang Xinruo. Residual multimodal transformer for expression-EEG fusion continuous emotion recognition. CAAI Trans Intell Technol 2024;9(5):1290–304. http://dx.doi.org/10.1049/cit2.12346.

[9] Li Jing, Chen Ning, Zhu Hongqing, Li Guangqiang, Xu Zhangyong, Chen Dingxin. Incongruity-aware multimodal physiology signals fusion for emotion recognition. Inf Fusion 2024;105:102220. http://dx.doi.org/10.1016/j.inffus.2023.102220.

[10] Pei Guanxiong, Shang Qian, Hua Shizhen, Li Taihao, Jin Jia. EEG-based affective computing in virtual reality with a balancing of the computational efficiency and recognition accuracy. Comput Hum Behav 2024;152:108085. http://dx.doi.org/10.1016/j.chb.2023.108085.

[11] Gao Ziheng, Huang Jiajin, Chen Jianhui, Zhou Haiyan. FAformer: parallel Fourier-attention architectures benefits EEG-based affective computing with enhanced spatial information. Neural Comput Appl 2024;36(8):3903–19. http://dx.doi.org/10.1007/s00521-023-09289-z.

[12] Xu FeiFan, Pan Deng, Zheng Haohao, Ouyang Yu, Jia Zhe, Zeng Hong. EESCN: A novel spiking neural network method for EEG-based emotion recognition. Comput Methods Programs Biomed 2024;243:107927. http://dx.doi.org/10.1016/j.cmpb.2023.107927.

[13] Zhao Yue, Zeng Hong, Zheng Haohao, Wu Jing, Kong Wanzeng, Dai Guojun. A bidirectional interaction-based hybrid network architecture for eeg cognitive recognition. Comput Methods Programs Biomed 2023;238:107593. http://dx.doi.org/10.1016/j.cmpb.2023.107593.

[14] Parsa Mohsen, Rad Habib Yousefi, Vaezi Hadi, Hossein-Zadeh Gholam-Ali, Setarehdan Seyed Kamaledin, Rostami Reza, Rostami Hana, Vahabie Abdol-Hossein. EEG-based classification of individuals with neuropsychiatric disorders using deep neural networks: A systematic review of current status and future directions. Comput Methods Programs Biomed 2023;240:107683. http://dx.doi.org/10.1016/j.cmpb.2023.107683.

[15] Han Zihao, De Wilde Philippe. OCT data is all you need: How vision transformers with and without pre-training benefit imaging. 2025, http://dx.doi.org/10.48550/arXiv.2502.12379, arXiv preprint arXiv:2502.12379.

[16] Peruzzo Elia, Sangineto Enver, Liu Yahui, De Nadai Marco, Bi Wei, Lepri Bruno, Sebe Nicu. Spatial entropy as an inductive bias for vision transformers. Mach Learn 2024;113(9):6945–75. http://dx.doi.org/10.1007/s10994-024-06570-7.

[17] Chen Bianna, Chen CL Philip, Zhang Tong. GDDN: Graph domain disentanglement network for generalizable EEG emotion recognition. IEEE Trans Affect Comput 2024;15(3):1739–53. http://dx.doi.org/10.1109/TAFFC.2024.3371540.

[18] Chang Jiang, Zhang Zhixin, Qian Yuhua, Lin Pan. Multi-scale hyperbolic contrastive learning for cross-subject EEG emotion recognition. IEEE Trans Affect Comput 2025;1–16. http://dx.doi.org/10.1109/TAFFC.2025.3535542.

[19] Yang Jianli, Zhang Zhen, Fu Zhiyu, Li Bing, Xiong Peng, Liu Xiuling. Cross-subject classification of depression by using multiparadigm EEG feature fusion. Comput Methods Programs Biomed 2023;233:107360. http://dx.doi.org/10.1016/j.cmpb.2023.107360.

[20] Jeganathan Jayson, Koussis Nikitas C, Paton Bryan, Phogat Richa, Pang James, Mansour Sina L, Zalesky Andrew, Breakspear Michael. Spurious correlations in surface-based functional brain imaging. Imaging Neurosci 2025;3(1):1–15. http://dx.doi.org/10.1162/imag_a_00478.

[21] Charlebois-Poirier Audrey-Rose, Davoudi Saeideh, Lalancette Ève, Knoth Inga Sophia, Lippé Sarah. The level of cognitive functioning in school-aged children is predicted by resting EEG directed phase lag index. Sci Rep 2025;15(1):1–13. http://dx.doi.org/10.1038/s41598-025-85635-6.

[22] Lin Changkai, Cheng Hongju, Rao Qiang, Yang Yang. M^{3}SA: Multimodal sentiment analysis based on multi-scale feature extraction and multi-task learning. IEEE/ACM Trans Audio Speech Lang Process 2024;32(1):1416–29. http://dx.doi.org/10.1109/TASLP.2024.3361374.

[23] Rukhsar Salim, Tiwari Anil Kumar. Lightweight convolution transformer for cross-patient seizure detection in multi-channel EEG signals. Comput Methods Programs Biomed 2023;242:107856. http://dx.doi.org/10.1016/j.cmpb.2023.107856.

[24] Gong Xinrong, Chen CL Philip, Hu Bin, Zhang Tong. CiABL: Completeness-induced adaptative broad learning for cross-subject emotion recognition with EEG and eye movement signals. IEEE Trans Affect Comput 2024;15(4):1970–84. http://dx.doi.org/10.1109/TAFFC.2024.3392791.

[25] Ganin Yaroslav, Ustinova Evgeniya, Ajakan Hana, Germain Pascal, Larochelle Hugo, Laviolette François, March Mario, Lempitsky Victor. Domain-adversarial training of neural networks. J Mach Learn Res 2016;17(59):1–35. http://dx.doi.org/10.48550/arXiv.1505.07818.

[26] Li Zhunan, Zhu Enwei, Jin Ming, Fan Cunhang, He Huiguang, Cai Ting, Li Jinpeng. Dynamic domain adaptation for class-aware cross-subject and cross-session EEG emotion recognition. IEEE J Biomed Heal Inf 2022;26(12):5964–73. http://dx.doi.org/10.1109/JBHI.2022.3210158.

[27] Lin Xuefen, Chen Jielin, Ma Weifeng, Tang Wei, Wang Yuchen. EEG emotion recognition using improved graph neural network with channel selection. Comput Methods Programs Biomed 2023;231:107380. http://dx.doi.org/10.1016/j.cmpb.2023.107380.

[28] Liu Wei, Zheng Wei-Long, Li Ziyi, Wu Si-Yuan, Gan Lu, Lu Bao-Liang. Identifying similarities and differences in emotion recognition with EEG and eye movements among Chinese, German, and French people. J Neural Eng 2022;19(2):1–20. http://dx.doi.org/10.1088/1741-2552/ac5c8d.

[29] Feng Naishi, Zhou Bin, Zhang Qianqian, Hua Chengcheng, Yuan Yue. A comprehensive exploration of motion sickness process analysis from EEG signal and virtual reality. Comput Methods Programs Biomed 2025;264:108714. http://dx.doi.org/10.1016/j.cmpb.2025.108714.

[30] Tarnowski Paweł, Kołodziej Marcin, Majkowski Andrzej, Rak Remigiusz Jan. Eye-tracking analysis for emotion recognition. Comput Intell Neurosci 2020;2020(1):2909267. http://dx.doi.org/10.1155/2020/2909267.

[31] Niaki Marzieh, Dharia Shyamal Y, Chen Yangjun, Valderrama Camilo E. Bipartite graph adversarial network for subject-independent emotion recognition. IEEE J Biomed Heal Informat. 2025;1–14. http://dx.doi.org/10.1109/JBHI.2025.3570187.

[32] Xu Yongjie, Yu Zengjie, Li Yisheng, Liu Yuehan, Li Ye, Wang Yishan. Autism spectrum disorder diagnosis with EEG signals using time series maps of brain functional connectivity and a combined CNN–LSTM model. Comput Methods Programs Biomed 2024;250:108196. http://dx.doi.org/10.1016/j.cmpb.2024.108196.

[33] Lin Yong, Dong Hanze, Wang Hao, Zhang Tong. Bayesian invariant risk minimization. In: Conference on computer vision and pattern recognition. 2022, p. 16000–9. http://dx.doi.org/10.1109/CVPR52688.2022.01555.

[34] Hipp Joerg F, Hawellek David J, Corbetta Maurizio, Siegel Markus, Engel Andreas K. Large-scale cortical correlation structure of spontaneous oscillatory activity. Nat Neurosci 2012;15(6):884–90. http://dx.doi.org/10.1038/nn.3101.

[35] Xu Peng, Zhu Xiatian, Clifton David A. Multimodal learning with transformers: A survey. IEEE Trans Pattern Anal Mach Intell 2023;45(10):12113–32.

[36] Zhang Zhongyi, Meng Qinghao, Jin LiCheng, Wang Hanguang, Hou Huirang. A novel EEG-based graph convolution network for depression detection: incorporating secondary subject partitioning and attention mechanism. Expert Syst Appl 2024;239(1):1–13. http://dx.doi.org/10.1016/j.eswa.2023.122356.

[37] Kim Hun-gyeom, Song Solwoong, Cho Baek Hwan, Jang Dong Pyo. Deep learning-based stress detection for daily life use using single-channel EEG and GSR in a virtual reality interview paradigm. PLoS One 2024;19(7):1–13. http://dx.doi.org/10.1371/journal.pone.0305864.

[38] Imbert Laetitia, Neige Cécilia, Moirand Rémi, Piva Giulia, Bediou Benoit, Vallet William, Brunelin Jerome. Eye-tracking evidence of a relationship between attentional bias for emotional faces and depression severity in patients with treatment-resistant depression. Sci Rep 2024;14(1):1–6. http://dx.doi.org/10.1038/s41598-024-62251-4.

[39] Moreno-Galván Diego Aarón, López-Santillán Roberto, González-Gurrola Luis Carlos, Montes-Y-Gómez Manuel, Sánchez-Vega Fernando, López-Monroy Adrián Pastor. Automatic movie genre classification and emotion recognition via a BiProjection multimodal transformer. Inf Fusion 2025;113(1):1–15. http://dx.doi.org/10.1016/j.inffus.2024.102641.

[40] Ma Hui, Wang Jian, Lin Hongfei, Zhang Bo, Zhang Yijia, Xu Bo. A transformer-based model with self-distillation for multimodal emotion recognition in conversations. IEEE Trans Multimed 2024;26(1):776–88. http://dx.doi.org/10.1109/TMM.2023.3271019.

[41] Dosovitskiy Alexey, Beyer Lucas, Kolesnikov Alexander, Weissenborn Dirk, Zhai Xiaohua, Unterthiner Thomas, Dehghani Mostafa, Minderer Matthias, Heigold Georg, Gelly Sylvain, Uszkoreit Jakob, Houlsby Neil. An image is worth 16x16 words: Transformers for image recognition at scale. Int Conf Learn Represent 2021;1–21.

[42] Singh Gopendra Vikram, Firdaus Mauajama, Ekbal Asif, Bhattacharyya Pushpak. EmoInt-trans: A multimodal transformer for identifying emotions and intents in social conversations. IEEE/ACM Trans Audio Speech Lang Process 2023;31(1):290–300. http://dx.doi.org/10.1109/TASLP.2022.3224287.

[43] Huang Jian, Tao Jianhua, Liu Bin, Lian Zheng, Niu Mingyue. Multimodal transformer fusion for continuous emotion recognition. Int Conf Acoust Speech Signal Process 2020;3507–11. http://dx.doi.org/10.1109/ICASSP40776.2020.9053762.

[44] Wu Yujin, Daoudi Mohamed, Amad Ali. Transformer-based self-supervised multimodal representation learning for wearable emotion recognition. IEEE Trans Affect Comput 2023;15(1):157–72. http://dx.doi.org/10.48550/arXiv.2303.17611.

[45] Chen Cheng, Fang Hao, Yang Yuxiao, Zhou Yi. Model-agnostic meta-learning for EEG-based inter-subject emotion recognition. J Neural Eng 2025;22(1):016008. http://dx.doi.org/10.1088/1741-2552/ad9956.

[46] Gómez-Lombardi Andre, Costa Begoña Góngora, Gutiérrez Pavel Prado, Carvajal Pablo Muñoz, Rivera Lucía Z, El-Deredy Wael. The cognitive triad network-oscillation-behaviour links individual differences in EEG theta frequency with task performance and effective connectivity. Sci Rep 2024;14(1):21482. http://dx.doi.org/10.1038/s41598-024-72229-x.

[47] Li Zhongjie, Zhang Gaoyan, Wang Longbiao, Wei Jianguo, Dang Jianwu. Emotion recognition using spatial–temporal EEG features through convolutional graph attention network. J Neural Eng 2023;20(1):016046. http://dx.doi.org/10.1088/1741-2552/acb79e.

[48] Li Chao, Bian Ning, Zhao Ziping, Wang Haishuai, Schuller Björn W. Multi-view domain-adaptive representation learning for EEG-based emotion recognition. Inf Fusion 2024;104:102156. http://dx.doi.org/10.1016/j.inffus.2023.102156.

[49] Xu Chengjian, Song Yonghao, Zheng Qingqing, Wang Qiong, Heng Pheng-Ann. Unsupervised multi-source domain adaptation via contrastive learning for eeg classification. Expert Syst Appl 2025;261:125452. http://dx.doi.org/10.1016/j.eswa.2024.125452.

[50] Li Guangqiang, Chen Ning, Niu Yixiang, Xu Zhangyong, Dong Yuxuan, Jin Jing, Zhu Hongqin. MSLTE: multiple self-supervised learning tasks for enhancing EEG emotion recognition. J Neural Eng 2024;21(2):024003. http://dx.doi.org/10.1088/1741-2552/ad3c28.

[51] Ludwig Siegfried, Bakas Stylianos, Adamos Dimitrios A, Laskaris Nikolaos, Panagakis Yannis, Zafeiriou Stefanos. EEGminer: discovering interpretable features of brain activity with learnable filters. J Neural Eng 2024;21(3):036010. http://dx.doi.org/10.1088/1741-2552/ad44d7.

[52] Vinao-Carl Matteo, Gal-Shohet Yuval, Rhodes Edward, Li J, Hampshire Adam, Sharp D, Grossman Nir. Just a phase? Causal probing reveals spurious phasic dependence of sustained attention. NeuroImage 2024;285:120477. http://dx.doi.org/10.1016/j.neuroimage.2023.120477.

[53] Yang Hua, Chen CL Philip, Chen Bianna, Zhang Tong. Improving the interpretability through maximizing mutual information for EEG emotion recognition. IEEE Trans Affect Comput 2024;1–14. http://dx.doi.org/10.1109/TAFFC.2024.3463469.

[54] Luo Yun, Liu Wei, Li Hanqi, Lu Yong, Lu Bao-Liang. A cross-scenario and cross-subject domain adaptation method for driving fatigue detection. J Neural Eng 2024;21(4):046004. http://dx.doi.org/10.1088/1741-2552/ad546d.

[55] Peng Dan, Zheng Wei-Long, Liu Luyu, Jiang Wei-Bang, Li Ziyi, Lu Yong, Lu Bao-Liang. Identifying sex differences in EEG-based emotion recognition using graph convolutional network with attention mechanism. J Neural Eng 2023;20(6):066010. http://dx.doi.org/10.1088/1741-2552/ad085a.

[56] Xu Guixun, Guo Wenhui, Wang Yanjiang. LSTM-enhanced multi-view dynamical emotion graph representation for EEG signal recognition. J Neural Eng 2023;20(3):036038. http://dx.doi.org/10.1088/1741-2552/ace07d.

[57] Zhang Yuzhe, Liu Huan, Wang Di, Zhang Dalin, Lou Tianyu, Zheng Qinghua, Quek Chai. Cross-modal credibility modelling for EEG-based multimodal emotion recognition. J Neural Eng 2024;21(2):026040. http://dx.doi.org/10.1088/1741-2552/ad3987.

[58] Zhang Yiling, Liao Yuan, Chen Wei, Zhang Xiruo, Huang Liya. Emotion recognition of EEG signals based on contrastive learning graph convolutional model. J Neural Eng 2024;21(4):046060. http://dx.doi.org/10.1088/1741-2552/ad7060.

[59] Hu Mengting, Xu Dan, He Kangjian, Zhao Kunyuan, Zhang Hao. Cross-subject emotion recognition with contrastive learning based on EEG signal correlations. Biomed Signal Process Control 2025;104:107511. http://dx.doi.org/10.1016/j.bspc.2025.107511.

[60] Wang Zhe, Wang Yongxiong, Hu Chuanfei, Yin Zhong, Song Yu. Transformers for EEG-based emotion recognition: A hierarchical spatial information learning model. IEEE Sens J 2022;22(5):4359–68. http://dx.doi.org/10.1109/JSEN.2022.3144317.

[61] Wang Zhe, Wang Yongxiong, Zhang Jiapeng, Hu Chuanfei, Yin Zhong, Song Yu. Spatial-temporal feature fusion neural network for EEG-based emotion recognition. IEEE Trans Instrum Meas 2022;71(1):1–12. http://dx.doi.org/10.1109/TIM.2022.3165280.

[62] Yan Jingjie, Du Chengkun, Li Na, Zhou Xiaoyang, Liu Ying, Wei Jinsheng, Yang Yuan. Spatio-temporal graph bert network for EEG emotion recognition. Biomed Signal Process Control 2025;104:107576. http://dx.doi.org/10.1016/j.bspc.2025.107576.

[63] Zang Zhibang, Yu Xiangkun, Fu Baole, Liu Yinhua, Ge Shuzhi Sam. Contrastive reinforced transfer learning for EEG-based emotion recognition with consideration of individual differences. Biomed Signal Process Control 2025;106:107622. http://dx.doi.org/10.1016/j.bspc.2025.107622.

[64] Zhu Lei, Xu Mengxuan, Huang Aiai, Zhang Jianhai, Tan Xufei. Multiple class transfer learning framework with source label adaptive correction for EEG emotion recognition. Biomed Signal Process Control 2025;104:107536. http://dx.doi.org/10.1016/j.bspc.2025.107536.

[65] Zhou Hong-Yu, Yu Yizhou, Wang Chengdi, Zhang Shu, Gao Yuanxu, Pan Jia, Shao Jun, Lu Guangming, Zhang Kang, Li Weimin. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. Nat Biomed Eng 2023;7(6):743–55. http://dx.doi.org/10.1038/s41551-023-01045-x.

[66] Li Rui, Ren Chao, Zhang Sipo, Yang Yikun, Zhao Qiqi, Hou Kechen, Yuan Wenjie, Zhang Xiaowei, Hu Bin. STSNet: a novel spatio-temporal-spectral network for subject-independent EEG-based emotion recognition. Heal Inf Sci Syst 2023;11(1):1–25. http://dx.doi.org/10.1007/s13755-023-00226-x.

[67] Luo Jie, Cui Weigang, Xu Song, Wang Lina, Li Xiao, Liao Xiaofeng, Li Yang. A dual-branch spatio-temporal-spectral transformer feature fusion network for EEG-based visual recognition. IEEE Trans Ind Inf 2024;20(2):1721–31. http://dx.doi.org/10.1109/TII.2023.3280560.

[68] Lengerich Benjamin J, Xing Eric, Caruana Rich. Dropout as a regularizer of interaction effects. Int Conf Artif Intell Stat 2022;7550–64.

[69] Zheng Wei-Long, Lu Bao-Liang. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. IEEE Trans Auton Ment Dev 2015;7(3):162–75. http://dx.doi.org/10.1109/TAMD.2015.2431497.

[70] Zhang Haokai, Li Pengrui, Chang Hongli, Liu Shihong, Qin Yun, Xie Jiaxin, Wang Manqing, Gao Dongrui, Wu Dingming. A coupling of common–private topological patterns learning approach for cross-subject emotion recognition. Biomed Signal Process Control 2025;105:107550. http://dx.doi.org/10.1016/j.bspc.2025.107550.