ARTICLE
https://doi.org/10.26599/AIR.2026.9150001

CAAI Artificial Intelligence Research

GE-LSTM-Net: A Global-Enhanced LSTM Network with Attention for Spatiotemporal Feature Extraction in EEG-based Vigilance Estimation

Hao Lan[1], Meng Tang[1], Xiangyu Ju[1], Ming Li[1]✉, Dewen Hu[1]✉

ABSTRACT

Electroencephalograph (EEG)-based vigilance estimation methods have achieved significant progress. In vigilance-associated EEG signals, non-adjacent electrodes exhibit strong coupling, and distant time points also demonstrate significant dependence, not limited the adjacent ones. Therefore, fully extracting such rich global-local spatiotemporal characteristics is critical. In this paper, we propose the Global Enhanced LSTM Network (GE-LSTM-Net), in which synergizes the Transformer's attention mechanism with LSTM to enhance the extraction of spatiotemporal features. Firstly, a specialized sample partitioning strategy along with the designed feature fusion module is adopted to reorganize raw EEG signals into structured 3D differential entropy (DE) feature representations, effectively preserving spatiotemporal and frequency dependencies across electrode channels and time points. Secondly, the attention mechanism and LSTM are encapsulated into a novel module (GE-LSTM module), serving as the core of the proposed GE-LSTM-Net to simultaneously extract spatiotemporal features from 3D representations. In this module, the attention mechanism will extract global information and integrate it into each unit of the LSTM, enabling LSTM to focus on more critical electrode channels and time points and extract richer global-local features. Subsequently, the GE-LSTM-Net demonstrate competitive performance and achieved SOTA results compared to existing methods on two public vigilance datasets. The codes are available at: https://github.com/Lanhao23-nudt/GE-LSTM-Net.

KEYWORDS

electroencephalograph (EEG), vigilance estimation, self-attention, Transformer, LSTM

Vigilance refers to the ability to focus on a specific task for an extended period of time [1]. Accurate vigilance estimation, which can help to issue timely alerts when the workers is in low vigilance level, is essential in many task scenarios. Various physiological signals that are not influenced by the environment are widely used in vigilance estimation [2, 3, 4, 5]. Among them, the electroencephalograph (EEG) is considered a highly reliable signal for vigilance estimation, due to its high temporal resolution and characteristics that can directly reflect brain activity [6, 7, 8, 9].

In recent years, deep learning-based EEG-decoding method demonstrated significant progress. The proposed famous EEGNet by Lawhern et. al [10] marked the popularity of deep learning-based EEG decoding. Subsequently, models such as the proposed EEG-Conformer [11], TSception [12], and CSF-NET [13] continuously improved EEG decoding accuracy across various EEG-decoding areas. However, it is still challenging to efficiently extract spatiotemporal features which represents the vigilance. In spatial information extraction, even distant electrodes can exhibit significant coupling [14, 15], and in temporal extraction, strong temporal dependencies also exist across distant time points [16, 17] , not limited to adjacent ones. To extract the rich global-local dependencies, several Recurrent Neural Network (RNN)-based methods were proposed [18, 19, 20]. Cui et al. [21] developed the GRU-MCC model, which leverages Gate Recurrent Unit (GRU) to extract spatial information. Li et al. [22] introduced the BiHDM model, and Du et al. [23] proposed the ATDD-LSTM, utilizing RNNs and Long Short-Term Memory (LSTM) for spatial feature extraction, respectively. For temporal modeling, Gao et al. [24] proposed SFT-Net, and Singh et al. [25] designed the 2L-LSTM network, both employing LSTM networks to capture short-term and long-term temporal dependencies. However, although these RNN-based models are capable of capturing some short-term and long-term dependencies, the way of which remembering the long-term information based on the updates of hidden states, may dilute the dependence between distant positions and overlook critical units (such as critical electrodes and time points).

To fully extract the spatiotemporal information of EEG signals in EEG-based vigilance estimation, we present the GE-LSTM-Net (global enhanced long-short-term memory network) in which we introduce the multi-heads self-attention mechanism of the Transformer to enhance global dependency

1 The College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China
✉ Corresponding authors: Ming Li (liming78@nudt.edu.cn) and Dewen Hu (dwhu@nudt.edu.cn)

extraction capability of LSTM. Firstly, a specialized sample partitioning strategy along with the designed feature fusion module was adopted, which reorganized raw EEG signals into structured 3D differential entropy (DE) feature representations and effectively preserve spatiotemporal and frequency dependencies between electrode channels and time points. Then, the attention mechanism and LSTM were encapsulated into a novel module (GE-LSTM module), in which the attention mechanism extracted global information and integrated it into each unit of the LSTM, enabling LSTM to extract richer global-local features and focused on more critical electrodes and time points. In addition, the network demonstrated competitive performance on the SEED-VIG and MMV public vigilance estimation dataset and achieved SOTA results compared to existing methods. In summary, the main contributions of this paper can be summarized as follows:

(1) The adopted sample partitioning method combined with the designed DE feature fusion module in this paper enables the model to effectively extract frequency, spatial, and temporal information from EEG signals.

(2) The proposed GE-LSTM module, which serves as the core of the GE-LSTM-Net, can be used to extract both spatial and temporal information between different electrodes and time points, respectively, and can be easily adopted to other EEG decoding tasks.

(3) We proved that the DE features extracted from the finer filtering techniques contains richer vigilance associated information than traditional filtering technique by the visualization of saliency map and can improve the performance of proposed GE-LSTM-Net.

(4) The GE-LSTM-Net built using GE-LSTM module achieved SOTA performance on the public vigilance dataset SEED-VIG and MMV, outperforming existing advanced baseline methods.

1 Related work

1.1 Handcrafted features for vigilance estimation

Before the widespread use of deep learning, research on EEG-based vigilance estimation focused on the handcrafted features [26, 27, 28, 29]. Brahim et al. [30] extracted spectral-spatial features through two specially designed filters, and selected the most important one by means of Mutual Information before classification. Lin et al. [31] confirmed that logarithmic transformations of EEG spectral amplitudes exhibit strong linear correlations with wake-sleep transitions. Angari et al. [32] studied the correlation between sample entropy and changes in fatigue levels, which improved the accuracy of fatigue monitoring. Guo et al. [33] computed the power spectrum in multiple frequency bands and brain regions to investigate its relationship with behavioral performance at varying levels of vigilance. Shi et al. [34] proposed a novel feature called DE and demonstrated its efficiency by comparing it with four commonly used vigilance estimation features. Lu et al. [16] utilized traditional support vector regression (SVR) to perform regression fitting between DE features extracted from raw EEG signals and vigilance labels, quantitatively assessing vigilance levels. Notably, the DE feature remains one of the most popular features in this field today [16, 35].

1.2 End-to-end deep learning model for EEG decoding

With the popularity of deep learning, deep learning EEG decoding models, demonstrating excellent performance, gradually replaced traditional manual feature extraction methods. These models include a large number of end-to-end models [36, 37, 38], which break the boundaries between different brain decoding tasks. They can automatically learn task-related features from the raw data and directly output results. For example, Schirrmeister et al. [39] proposed shallow ConvNet and deep ConvNet that directly performed convolution and pooling operations on the original signal and directly output the categories. Lawhern et al. [10] designed EEGNet using separable convolution and separable convolution, achieving a lightweight model for accurate EEG decoding of EEG signals in multiple tasks. Gong et al. proposed an attention-driven spatial-temporal dual-stream fusion network that captures the spatiotemporal coupled and complementary characteristics of EEG signals [40]. Ding et al. [12] introduced prior knowledge of the asymmetry between the left and right hemispheres of the brain and designed the TSception network, which further improved the accuracy of EEG decoding. Song et al. [11] introduced the Transformer to extracted the temporal information from raw EEG signals. Besides, recent studies have proposed spiking neural networks for joint spatio-temporal feature learning in EEG-Based emotion recognition [41]. In summary, there has been significant progress in end-to-end EEG decoding models in recent years.

1.3 Deep learning models based on some handcrafted features

Directly inputting raw EEG signals into deep networks can significantly increase the computational burden of deep learning models due to the large volume of data. Therefore, some researchers attempt to extract handcrafted features first before feeding them into deep learning networks [42, 43, 44, 45]. For instance, Lin et al. [46] extracted the power spectral density (PSD) of raw EEG signals, transforming them into a series of power energy maps before importing them into their designed network for fatigue monitoring. Wang [47] extracted both DE features and symmetry quotient (SQ) before inputting to a CNN-LSTM neural network for classification task. Shen et al. [42] extracted DE features across different frequency bands from raw signals, then employed convolutional layers to capture spatial information and LSTM to model temporal dependencies. Building on this approach, Gao et al. adopted the same DE feature extraction strategy and developed more sophisticated convolutional networks, including CSF-GTNet [13] and SFT-Net [24], which have consistently achieved state-of-the-art results on the vigilance estimation dataset SEED-VIG. In summary, the approach of first extracting certain features and then inputting them into a deep network demonstrate significant performance.

2 Method

As shown in Fig.1, the structure of GE-LSTM-Net includes 4 parts: Sample Partitioning, DE Feature Fusion module, GE-LSTM module and Classification module. The GE-LSTM modules used to extract spatial and temporal

information are referred to as Spa-GE-LSTM and Tim-GE-LSTM, respectively. The details will be introduced in turn.

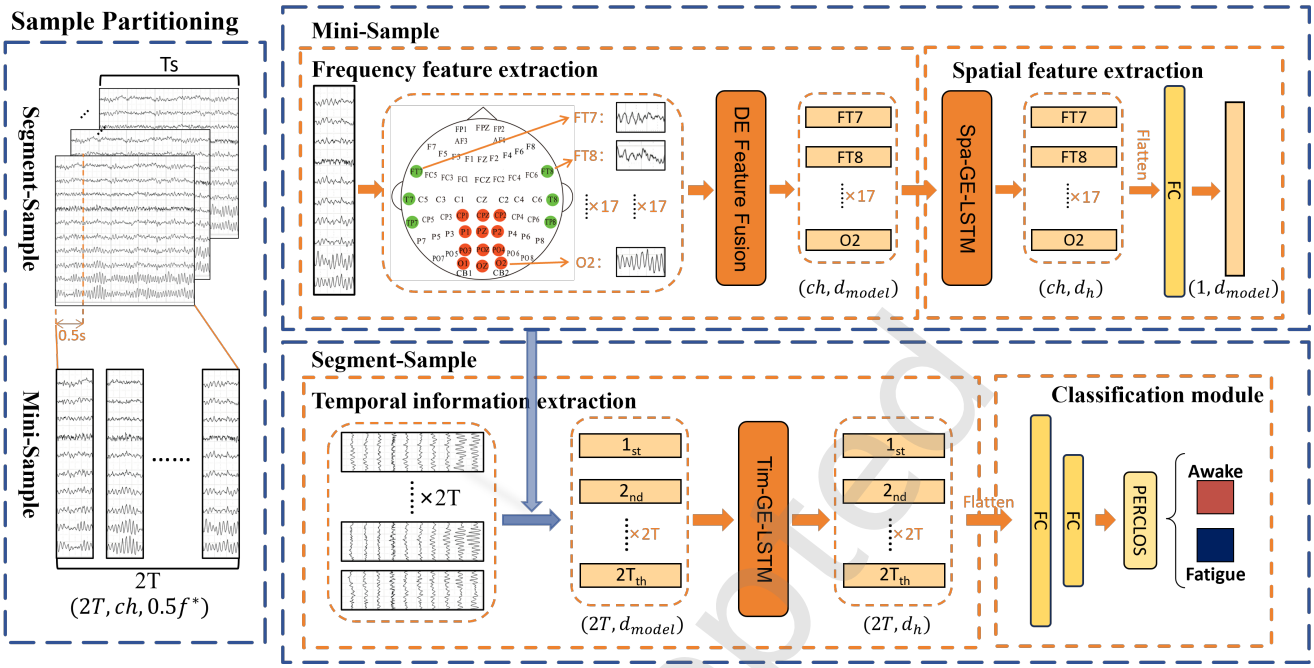

Fig. 1 The framework of GE-LSTM-Net, including 4 parts: Sample Partitioning, DE Feature Fusion module, GE-LSTM Module and Classification Module. Through the four parts of the network, it can progressively extract and integrate frequency, spatial, and temporal information from the EEG signals. The Spa-GE-LSTM and Tim-GE-LSTM in the figure refer to GE-LSTM module for extracting spatial and temporal information, respectively.

2.1 Sample Partitioning

As shown on the left side of Fig.1, the EEG signals from all electrodes are divided into samples using non-overlapping Ts time windows, where $T = 8$. These samples, which are called segment-samples, will be then divided into minor sample parts through 0.5s non-overlapping time windows, consistent with the previous setting in and ensuring a fair comparison of model performance under equivalent data volumes [24, 42]. And these minor samples are called mini-samples. The segment-sample and mini-sample of raw EEG data can be denoted as $S_R \in R^{2T \times ch \times f_s}$ and $M_R \in R^{ch \times f_s}$, respectively. Here, ch represents the number of EEG electrode channels, and f_s represents the sampling rate of EEG signals. Based on this sample partitioning method, the frequency and spatial information within every mini-sample are firstly extracted based on the proposed DE Feature Fusion and Spa-GE-LSTM modules. And the temporal information within $2T$ mini-samples of segment-sample are further extracted from the Tim-GE-LSTM module. By means of this sample partitioning method, the model can effectively extract frequency, spatial, and temporal information within EEG signals.

2.2 DE Feature Fusion Module

Inspired by the previous findings that finer filtering techniques (filtered the raw EEG signals according 2Hz windows) before extracting DE features can improve the performance of previous EEG decoding methods [16, 17]. To verify the efficiency of finer filtering techniques in the proposed GE-LSTM-Net, we adopted three distinct filtering techniques. As shown on the Fig.2, the first filtering technique filters the raw EEG signal according to five traditional frequency bands (Fre-Bands): delta (δ, 1-4 Hz), theta (θ, 4-8 Hz), alpha (α, 8-13 Hz), beta (β, 13-30 Hz), and gamma (γ, 30-51 Hz). The other two finer filtering techniques filter the raw EEG signal according to the windows of 5 Hz and 2 Hz, respectively. These different filtering methods will generate 5, 10, and 25 Fre-Bands, respectively. Subsequently, we utilize the popular DE feature, which has demonstrated effective performance in both emotion recognition and vigilance estimation [34, 42, 13]. Its calculation formula is as follows:

$$DE = \int_Z f(z) log(f(z)) dz \tag{1}$$

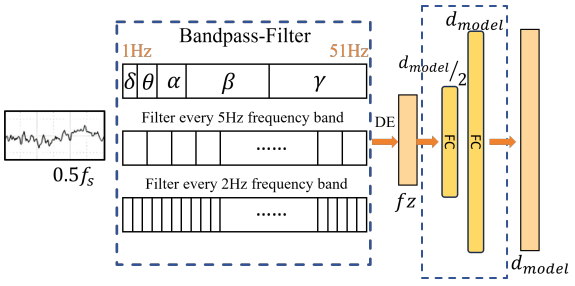

Fig. 2 DE Feature Fusion module. It includes three different filtering methods, as well as two linear layers and a rectified linear unit (ReLU) activation function between to fuse the linear and nonlinear interactions between different frequency bands (Fre-Bands) and expand the dimensionality of DE features.

In the formula, z represents the filtered raw EEG data, and it is generally assumed that z follows a Gaussian distribution with the density function $f(\cdot)$. As a result, the formula can be further simplified as:

$$DE = \int_Z \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}} log(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}) dz$$
$$= \frac{1}{2} log(2\pi e\sigma^2) \quad (2)$$

where e denotes the base of the natural logarithm and σ^2 represents the sample variance of filtered raw EEG data.

According to the formula, the DE feature is extracted from every Fre-band. Based on three different filtering techniques, the dimensionality of the DE feature extracted from each electrode will be $f_z = 5, 10, or 25$. As a result, the Mini-Sample can be further represented as $M_D \in R^{ch \times f_z}$ and the Segment-Sample can be represented as $S_D \in R^{2T \times ch \times f_z}$.

As shown on the right side of Fig.2, two fully connected (FC) layers, with $d_{model}/2$ and d_{model} neurons, associated with a Rectified Linear Unit (ReLU) activation function between are designed in the DE feature fusion module. This combination captures both linear dependencies through weight matrices and non-linear interactions via activation transformations. After passing through this module, the dimension of each mini-sample becomes $ch \times d_{model}$. The default value of d_{model} is set to 64.

2.3 GE-LSTM Module

In this paper, we propose a novel GE-LSTM module, designed to enhance global dependency information of LSTM network in EEG signal analysis. The proposed modules used to extract spatial and temporal dependencies are termed Spa-GE-LSTM and Tim-GE-LSTM, respectively, which include identical architectures. Each GE-LSTM module consists of multiple GE-LSTM units. As shown in Fig.3, each GE-LSTM unit is formed by cascading a Transformer Encoder layer (TranEnlayer) in front of the LSTM, leveraging the multi-head self-attention mechanism in TranEnlayer to extract global dependency information and integrate it into each unit of the LSTM for further capturing long-term and short-term dependencies. More specifically, The features of electrodes or time points after the TranEnlayer will integrated global information into feature of each electrode or point and strengthen importance of specific electrode and point. Feature of each electrode or time point will be then input into each unit of LSTM to extract long-short-term dependence. This GE-LSTM unit is actually a two-stage pipeline, a easy but efficient construction for performance enhancement.

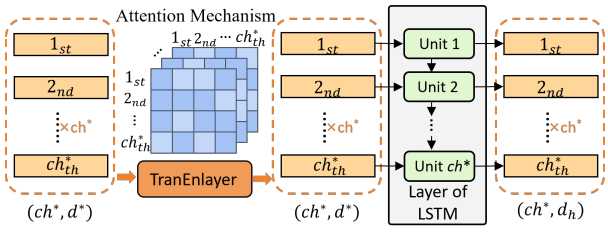

Fig. 3 GE-LSTM units of GE-LSTM module, which is formed by cascading a TranEnlayer before an LSTM layer. The attention mechanism in TranEnlayer will extract global dependency information and integrate it into each unit of the LSTM.

The following section will detail introduce the components of the GE-LSTM module: positional embedding (PE), which is typically needed before the TranEnlayer, TranEnlayer and LSTM.

Positional Embedding

Due to the parallel computing characteristics of the Transformer, the computation of vectors does not inherently capture positional relationships [48]. Therefore, position embedding (PE) is typically required and added to the original data to improve the model performance [49, 50].

Currently, in sequential information processing, the commonly used PE method is 1-dimensional (1D) sinusoidal PE [48, 50], which leverages sine and cosine functions of different frequencies. In the Tim-GE-LSTM module, the temporal information is sequential, and thus we adopt 1D sinusoidal PE. The specific calculation formula is as follows:

$$PE_{pos,2i} = \sin(pos/10000^{2i/d_{model}})$$
$$PE_{pos,2i+1} = \cos(pos/10000^{2i/d_{model}}) \quad (3)$$

Here, pos denotes the position index of each electrode, and $2i$, $2i+1$ represent the positions of the parameters in the PE vector. However, the relationship between electrodes is clearly not a simple sequential pattern. Therefore, we map the electrodes according to their positions into a 2-dimensional (2D) image, as shown in Fig.4, and explore several other PE methods used in Vision Transformers (VIT) [49], the learnable PE and the 2D sinusoidal PE. The former directly adds learnable parameters to the original data while the calculation formula for the latter is as follows:

$$PE_{pos,2i} = \sin(pos_x/10000^{4i/d_{model}})$$
$$PE_{pos,2i+1} = \cos(pos_x/10000^{4i/d_{model}})$$
$$PE_{pos,2j} = \sin(pos_y/10000^{4j/d_{model}})$$
$$PE_{pos,2j+1} = \cos(pos_y/10000^{4j/d_{model}}) \quad (4)$$

In this equation, pos represents the position index of each electrode, and $2i$, $2i+1$, $2j$, and $2j+1$ denote the positions of the specific value in the PE vector, where $i \in \{n \in Z | 0 \le n < d_{model}/4\}$ and $j \in \{n \in Z | d_{model}/4 \le n < d_{model}/2\}$. pos_x and pos_y refer to the vertical and horizontal coordinates of the electrode after being mapped to a 2D image, which is shown in Fig.4.

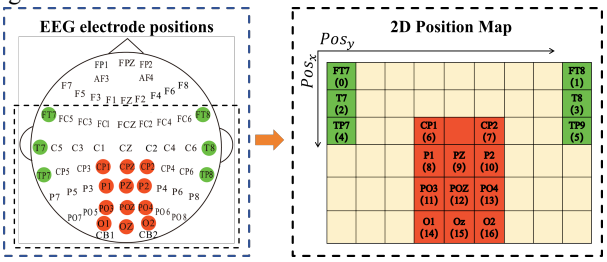

Fig. 4 The 2D position image mapped from the EEG electrode position. CPZ serves as the reference electrode and the index number in parentheses represents the pos of each electrode in 2D sinusoidal PE.

TranEnlayer

The overall architecture of TranEnlayer is shown in Fig.5, consistent with original Transformer architecture [48]. TranEnlayer consists of two core sub-layers, the multi-head self-attention mechanism and feed-forward network. A dropout with a rate of 0.1 and a layer normalization follows each sub-layer. In addition, residual connections are used around the two sub-layers, meaning that the final output of each sub-layer is given by $LayerNorm(x + Sublayer(x))$, where $Sublayer(x)$ represents the direct output of each sub-layer. The dimensions of the input and output data of the TranEnlayer module remain unchanged.

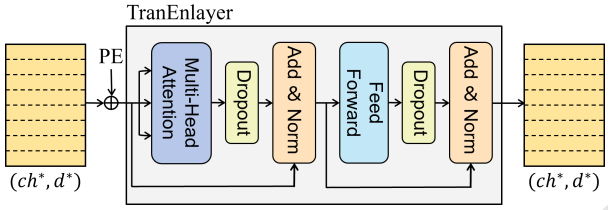

Fig. 5 Overall structure of TranEnlayer, which mainly consists of two core sub-layers, the multi-head self-attention mechanism and feed-forward network.

In the multi-head self-attention mechanism, the input data is first mapped to query(Q), keys(K), and values(V) through different linear layers. The dimensions of Q, K, and V are the same as the input data, which are expressed as $Q, K, V \in R^{ch^* \times d^*}$. Then, based on the number of heads, Q, K, and V are divided into Q_i, K_i, and V_i. The attention of different heads, capturing different dependencies between the input vectors, are further computed as follows :

$$\begin{aligned} heads_i &= Attention(Q_i, K_i, V_i) \\ &= \text{softmax}(\frac{Q_i K_i^T}{\sqrt{d_k}})V_i \end{aligned} \tag{5}$$

in which $i = 0, 1, ..., h_n - 1$, where h_n denotes the number of heads. $Q_i, K_i, V_i \in R^{ch^* \times d_k}$, where $d_k = \frac{d^*}{h_n}$. $\text{softmax}(\frac{Q_i K_i^T}{\sqrt{d_k}})$ represents the different learned association matrices between the input vectors, which is also called the attention score. $heads_i \in R^{ch^* \times \frac{d^*}{h_n}}$ represents the updated features after integrating the global association. As shown in Fig.6, the attention scores dynamically learn the association weights between input tokens, and through matrix multiplication, the output tokens naturally acquire global information from other tokens and place greater focus on those with higher association weights (for example, the token corresponding to the attention weight of 0.6 during matrix multiplication in the figure). Through the attention mechanism, token representations with richer local-global information are obtained.

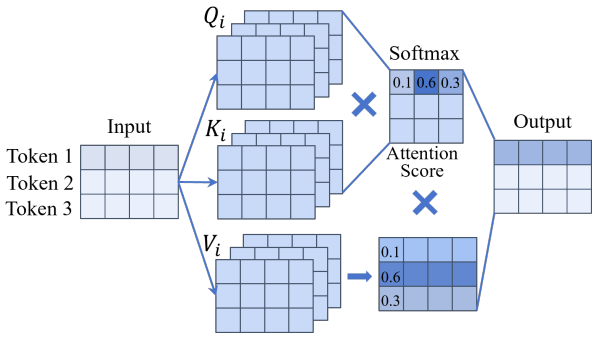

Fig. 6 The data flow of attention mechanism, the attention scores can dynamically integrate global information into every token (electrode or time point) and focusing on the token with high attention weights.

At the end of the attention mechanism, the outputs of all the attention heads are concatenated and passed through an FC layer to integrate the diverse association information learned from different heads. The input and output dimensionality of the FC layer is consistent.

The feed-forward network consists of two FC layers with a ReLU activation in between. The first layer expands the input dimensionality by an expansion factor of n, while the second layer reduces it back to the original size. n is set to 4 in this paper. The ReLU activation adds non-linearity, enabling the model to learn more intricate patterns.

LSTM

LSTM is a network with a chain-like structure of repeating units. Each repeating units contains three special gates: the forget gate, the input gate, and the output gate, along with a special cell state that captures the long-term dependencies between the input data. The vectors input into different repeating units are the data from different electrodes and time points in this paper. The calculation formulas for the three gates and the update formula for cell state are as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{6}$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{7}$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{8}$$
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{9}$$
$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{10}$$
$$h_t = o_t * \tanh(C_t) \tag{11}$$

Here, f_t and i_t denote the forget gate and the input gate, respectively, and C_t denotes the cell state. The forget gate and the input gate jointly complete the update of C_t, while C_{t-1} represents the previous cell state and $\tilde{C}_t$ represents the candidate value in current units that could be added to the cell state. h_{t-1} represents the output (also called the hidden state) of the previous unit. x_t represents the input of the current units. The symbol "[,]" means the concatenation operation. o_t denotes the output gate, which, together with the updated cell state, determines the current hidden state h_t. W and b denote the weight matrices and the bias term, respectively. The symbols σ and tanh represent the Sigmoid activation function and the hyperbolic tangent activation function.

2.4 Classification Module

The final output of Tim-GE-LSTM is represented as $T_n \in R^{ch \times d_h}$. After flattening all dimensions of T_n into a vector, a linear layer with 120 neurons is first applied for dimensionality reduction, followed by a second linear layer that reduces the dimensionality to 1. The final output $\hat{y}_n$ are compared with the target label PERCLOS y_n, which serves as an indicator of vigilance level, to perform the fitting task. A detailed introduction to PERCLOS is provided in Section 4.1. In addition, in scenarios such as fatigue driving detection, the primary operational need is determining whether to issue alerts (low-vigilance warnings). This critical 'warning decision' necessitates thresholding vigilance into two actionable states: 'attention required' or 'no alert needed'. Therefore, a binary classification task was further conducted based on the fitting results, where a value of 0.35 is adopted as the threshold to separate the states of awake and fatigue. The threshold of 0.35 has been widely adopted in previous studies [24, 13, 16], in which it is gradually accepted when PERCLOS is less than 0.35, the subjects are in an awake state, while the value exceeds 0.35, they begin to enter a fatigue state. In practical work, this means that appropriate alerts are needed. To ensure equitable comparison, we employed identical thresholds to those used in benchmark literature. The specific calculation process is as follows:

$$\text{Classification} = \begin{cases} \text{Awake} & (\hat{y}_n \geq 0.35) \\ \text{Fatigue} & (\hat{y}_n < 0.35) \end{cases} \tag{12}$$

3 Experiment

3.1 Dataset Description

The SEED-VIG dataset is a subset of the SEED dataset developed by the Brain-like Computing and Machine Intelligence (BCMI) laboratory at Shanghai Jiao Tong University [16]. The experiments are conducted on a virtual driving system simulating a monotonous four-lane straight road scenario. The simulated driving sessions are scheduled post-lunch and lasted for 2 hours, a timeframe particularly prone to fatigue induction. During the experiment, both EEG and EOG signals are recorded, though only EEG signals are utilized in this study. The dataset comprises a total of 23 subjects. Meanwhile, only a limited number of electrodes from the occipital and temporal regions, arranged according to the 10-20 electrode system, are utilized in this dataset. Excluding the reference electrodes, the dataset contains only 17 electrodes.

The MMV dataset, jointly constructed by the Chinese Academy of Sciences and Tianjin University, induces vigilance decline through prolonged visual tasks using either Rapid Serial Visual Presentation (RSVP) or Steady-State Visual Evoked Potential (SSVEP) paradigms [51]. While the dataset captures seven physiological signals, this study exclusively employs EEG signals for model evaluation. The EEG signals in MMV include 62 electrodes similarly arranged according to the international 10-20 system and are downsampled to 200 Hz. The dataset comprises 18 subjects, each participating in four experimental sessions (two RSVP and two SSVEP sessions), with each session lasting 2 hours.

In both datasets, eye movement signals are collected, including PERCLOS which serves as the most widely used indicator for measuring vigilance levels [52, 35]. PERCLOS, which denotes the percentage of eye blinks and eye closures within a certain duration, can be calculated as follows:

$$\text{PERCLOS} = \frac{\text{blink} + \text{CLOS}}{\text{blink} + \text{fixation} + \text{saccade} + \text{CLOS}} \tag{13}$$

where CLOS denotes the duration of eye closure. PERCLOS is computed every 8 seconds in the SEED-VIG dataset and 4 seconds in the MMV dataset. As indicated by the formula, PERCLOS ranges from 0 to 1, with higher values corresponding to lower vigilance levels.

3.2 Experiment Setting

The experiment was carried out using PyTorch 1.13.0 in a Python 3.8 environment, with a 12-core Intel Xeon Silver 4214R CPU, an RTX 3080 Ti GPU (12GB RAM). The learning rate was set to 3e-4, the batch size to 150, and the AdamW optimizer with a weight decay of 2e-2. The loss function was the Mean Squared Error (MSE) for the output values to PERCLOS. We performed mixed-subject five-fold cross-validation to evaluate the model's performance, a common experimental approach for evaluating the performance of EEG decoding models [18, 13, 24]. Under the mixed-subject paradigm, individual information from different subjects is considered as guidance for improving the model performance. However, this paradigm inherently requires including some training samples from the test subject before real-world deployment. Specifically, data from all subjects were mixed to form a unified dataset. For the SEED-VIG dataset, data from all 23 subjects were combined, while for the MMV dataset, data from the first session under the RSVP paradigm of 17 subjects were merged. In the five-fold cross-validation process, the dataset was randomly split into 5 folds, with one fold serving as the test set and the remaining four as the training set [16, 13]. In addition, five different initialization seeds were used to mitigate the effect of model initialization, with the average result considered the final performance metric.

All experimental results were evaluated using six classification metrics (Accuracy, Precision, Recall, F1-score, Kappa, Area Under the ROC Curve (AUC)) and two regression metrics (Root Mean Square Error (RMSE) and Correlation Coefficient (COR)). Accuracy measures the overall proportion of correct predictions, Precision and Recall focus on the accuracy and capture ability of positive class predictions, while F1-score combines both, making it suitable for class imbalance problems. Kappa measures the consistency of prediction, accounting for randomness. AUC provides a comprehensive evaluation of classifier performance across all possible decision thresholds, making it especially suitable for imbalanced classification tasks. RMSE reflects the deviation between predicted and actual values, and COR measures the linear correlation between predicted and actual values. Together, these metrics provide a comprehensive evaluation of both regression and classification quality.

The five classification metrics are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{14}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{15}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{16}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

$$\text{Kappa} = \frac{\text{Accuracy} - P_e}{1 - P_e} \tag{18}$$

$$\text{AUC} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} I(p_i^M > p_j^N)}{M \times N} \tag{19}$$

Where

$$P_e = \frac{(TP+FP)(TP+FN)+(TN+FP)(TN+FN)}{(M+N)^2} \tag{20}$$

$$I(p_i^M, p_j^N) = \begin{cases} 1 & \text{if } p_i^M > p_j^N \\ 0.5 & \text{if } p_i^M = p_j^N \\ 0 & \text{if } p_i^M < p_j^N \end{cases} \tag{21}$$

In these definitions, TP (True Positive) and FP (False Positive) refer to the number of samples correctly and incorrectly predicted as positive, respectively, while TN (True Negative) and FN (False Negative) refer to the number of samples correctly and incorrectly predicted as negative, respectively.M and N represents the number of fatigue and awake samples. p_i^M and p_j^N denote the predicted vigilance level of fatigue and awake samples.

The two regression metrics are defined as follows.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{22}$$

$$\text{COR} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{23}$$

Where y_i and $\hat{y}_i$ denote the labels and outputs, respectively, while $\bar{y}$ and $\bar{\hat{y}}$ denote their means.

4 Result and Discussion

This section presents a comprehensive analysis of the experimental results and discussions, focusing on multiple key factors that impact the performance of the GE-LSTM module and GE-LSTM-Net. First, we compared the performance of GE-LSTM-Net with several other competitive methods to evaluate its superiority. Next, we investigated the effect of finer Fre-Bands on model performance and introduced the saliency maps for visual analysis. Subsequently, we examined the effect of the number of units in the GE-LSTM module on its performance, followed by an analysis of the impact of the number of attention heads and PE on the module, respectively. Afterward, a series of ablation experiments were conducted to validate the contribution of each module to the final results. Finally, a real-time and deployability research was conducted to evaluate the practicality and efficiency of GE-LSTM-Net in real-world application.

4.1 Comparison with Other Models

The effectiveness of the proposed model in EEG-based vigilance estimation is evaluated on two public datasets: SEED-VIG and MMV. In the SEED-VIG dataset, the proposed model is compared against 10 existing methods, including ten open-source models—ShallowConvNet [39], DeepConvNet [39], EEGNet [10], EEG-Conformer [11], SFT-Net [24], TSception [12], LGG [53], ARNN [54], STGCN [55] and Deformer [38]—and four results replicated in the literature [24]. In the MMV dataset, the model is primarily compared with ten open-source models. All comparisons are made under identical experimental conditions, including the dataset, data division, optimizer, etc. Among these models: SFT-Net employs convolution to extract spatial signals followed by LSTM for temporal information extraction; EEG-Conformer and Deformer leverages the Transformer architecture to capture temporal dependencies; TSception, EEGNet, DeepConvNet, and ShallowConvNet are convolution-based models designed to extract spatiotemporal dependence; LGG and STGCN are graph-based models with attention mechanism; while ARNN employs the architecture combined both attention and recurrence.

The proposed GE-LSTM-Net demonstrated leading performance in EEG-based vigilance estimation tasks. In the SEED-VIG dataset, as shown in Table 1, the proposed model achieved leading results across seven evaluation metrics. Specifically, GE-LSTM-Net obtained a result of 0.943, outperforming the EEG-Conformer by 2.8%, with a p-value of $p < 0.001$, and the 95% confidence interval for its AUC is reported as [0.93608, 0.94991]. Furthermore, the results of paired t-test for Accuracy and Delong's test for AUC demonstrated that the proposed GE-LSTM-Net significantly outperformed all other models ($p < 0.001$). Besides, while the SFT model exhibited a higher Recall than GE-LSTM-Net based on public data, its Precision was significantly lower by 10.7%, revealing an imbalance in its results. Compared to convolutional-based spatiotemporal models (e.g., TSception, EEGNet), GE-LSTM-Net improved accuracy by 8–14%. The EEG-Conformer delivered relatively better performance, but there remained a substantial gap between its RMSE and COR values (RMSE: 0.124–0.100; COR: 0.888–0.929). The extreme case, ESTCNN, despite achieving the second-highest Recall (0.919), exhibited severely limited applicability due to excessively high false positives (Precision = 0.657). In the MMV dataset, as shown in Table 2, GE-LSTM-Net outperformed all six compared models across all seven evaluation metrics, while the results of the paired t-test for Accuracy and Delong's test for AUC also demonstrated its significant superiority over other baseline models ($p < 0.001$), demonstrating the superiority of the proposed model in multiple datasets.

4.2 Impact of Finer Fre-Bands and visualization

Inspired by the previous findings, finer filtering techniques (filtered the EEG signals according 2Hz windows, rather than traditional five bands filtering [56, 57]) before extracting DE features can improve the performance of previous EEG decoding methods [17, 16]. As a result, to verify the efficiency of finer filtering techniques in deep learning based model, We compared the the performance of the proposed GE-LSTM-Net in three different filtering techniques (traditional five bands filtering technique, 5Hz filtering techniques and 2Hz filtering technique) in SEED-VIG dataset.

As shown in Table 3, the results demonstrated that finer filtering technique can indeed improve the performance of

the proposed GE-LSTM-Net, and the finer filtering technique achieved the better results across seven evaluation metrics except Precision (2Hz filtering > 5Hz filtering > traditional five bands filtering). To further explore how finer frequency band divisions improve performance compared to the traditional five-frequency band segmentation, we introduced the saliency maps to visualize the contributions of different regions and Fre-bands in the states of fatigue and awake. Saliency map visualization is an interpretability technique that highlights the critical input features influencing the model's predictions, and has been widely used in emotion recognition tasks [12, 58].

As shown in Fig. 8, the visualization in the 10 Hz filtering technique shows the gamma range (30-50Hz) has great differences in lateral comparisons, while in tradition filtering technique it is considered as a single band and should have the same performance. This demonstrates that in vigilance estimation, treating the 30-50Hz range as a single Gamma frequency band fails to adequately represent the change of vigilance level while the finer Fre-bands contain richer information. Additionally, in other parts of the saliency map, the elevation of theta, alpha, and beta waves (1-30Hz) [15, 59] in the temporal lobe, as well as the theta (4-8Hz) and alpha (8-13Hz) waves in the occipital lobe, are enhanced [14, 60] when fatigue sets in, which is consistent with previous studies.

4.3 Visualization of Attention Score

To better understand the role of the attention mechanism in extracting spatial features within the proposed GE-LSTM module, we performed visualization on the saved model. Using the 3,000 most fatigued and most alert samples respectively, the attention scores were averaged to generate the results shown in Fig. 7. Each row of the data represents the association weights between the current token vector and all other global tokens. As observed in the association matrix for the awake state in Fig. 7 (a), the columns CP1 and CP2 exhibit relatively high attention weights, indicating that these two electrodes are generally more associated with others in the awake state and thus represent more pivotal electrodes. In contrast, the columns for T7 and T8 display relatively lower attention weights, suggesting that the information learned by the model from these electrodes exerts comparatively lesser influence during awake. Similarly, in the fatigue state attention score in Fig. 7 (b), columns CP2 and P2 show higher attention weights, indicating their relatively stronger relevance under fatigue state.This visualization demonstrates that in the proposed GE-LSTM module, the attention scores enhance global information extraction and focus more on the key electrodes.

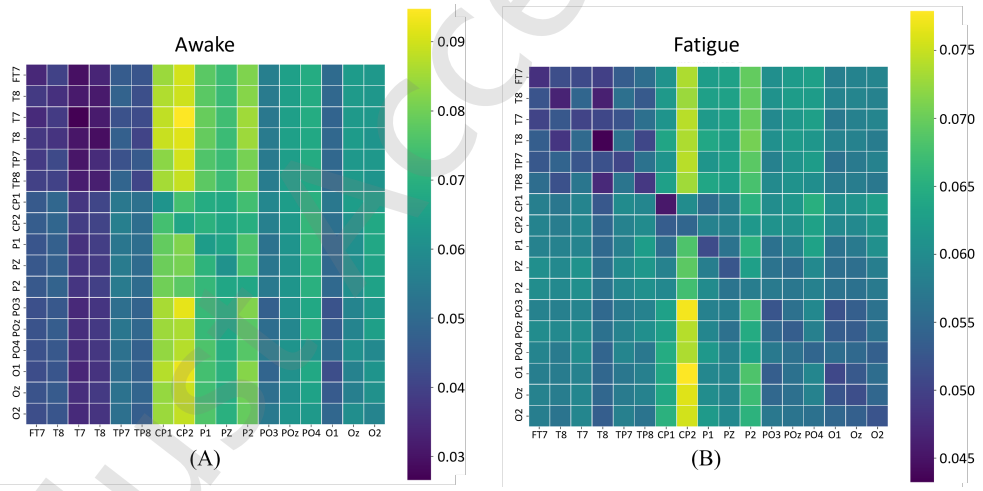

Fig. 7 Visualization of attention score for spatial feature extraction, where (a) represents the average attention score of awake status, and (b) represents the average attention score of fatigue status.

4.4 Effect of GE-LSTM Units on Module Performance

In this part, we explore the impact of the number of GE-LSTM units on the performance of the GE-LSTM module in the SEED-VIG dataset. As shown in Table. 4, with the increase in the number of GE-LSTM units, the performance of the GE-LSTM module first improves and then declines. The configuration with 2 GE-LSTM units performs the best across multiple metrics, except recall, outperforming other configurations. As the number of units continue to increase, the performance of model rapidly declines.In the configuration with 4 GE-LSTM units, the accuracy decreased to 0.873 (down from 0.886), Kappa dropped to 0.722 (from 0.751), the RMSE rose to 0.109 (from 0.100), the COR fell to 0.915 (from 0.929) and AUC dropped to 0.933 (from 0.943). Therefore, it is recommended to use two GE-LSTM units, and this follows common practices in LSTM networks that using two LSTM layers generally contributes to better performance [25, 24].

Table 1 The comparison with ten baseline models in SEED-VIG dataset

	Accuracy	Kappa	F1-Score	Recall	Precision	RMSE	COR	AUC
EEGNet	0.747	0.436	0.810	0.778	0.844	0.178	0.754	0.817
DeepConvNet	0.809	0.573	0.854	0.834	0.879	0.147	0.848	0.876
ShallowConvNet	0.805	0.571	0.851	0.830	0.873	0.145	0.844	0.874
AMS-CNN	0.802	0.554	0.844	0.874	0.820	-	-	-
EEG-Conv	0.795	0.536	0.848	0.896	0.805	-	-	-
EEG-Conv-R	0.768	0.478	0.827	0.872	0.787	-	-	-
ESTCNN	0.641	0.083	0.764	0.919	0.657	-	-	-
SFT-Net	0.871	0.717	0.901	**0.922**	0.820	-	-	-
EEG-Comformer	0.851	0.672	0.885	0.868	0.902	0.124	0.888	0.915
TSception	0.834	0.629	0.876	0.838	0.917	0.144	0.847	0.897
LGG	0.826	0.623	0.864	0.839	0.892	0.140	0.813	0.893
ARNN	0.863	0.705	0.890	0.873	0.914	0.118	0.906	0.922
STGCN	0.793	0.591	0.847	0.841	0.869	0.149	0.820	0.819
Deformer	0.845	0.663	0.876	0.856	0.887	0.126	0.867	0.903
GE-LSTM-Net	**0.886**	**0.751**	**0.912**	0.897	**0.927**	**0.100**	**0.929**	**0.943**

Table 2 The comparison with six baseline models in MMV dataset

	Accuracy	Kappa	F1-Score	Recall	Precision	RMSE	COR	AUC
EEGNet	0.889	0.726	0.803	0.802	0.804	0.149	0.855	0.941
DeepConvNet	0.876	0.682	0.765	0.823	0.717	0.164	0.823	0.922
ShallowConvNet	0.868	0.679	0.772	0.754	0.791	0.166	0.818	0.920
SFT-Net	0.910	0.774	0.836	0.856	0.817	0.125	0.900	0.953
EEG-Conformer	0.903	0.755	0.822	0.847	0.798	0.136	0.882	0.945
TSception	0.887	0.712	0.788	0.834	0.747	0.163	0.825	0.925
LGG	0.891	0.728	0.807	0.820	0.809	0.136	0.876	0.936
ARNN	0.907	0.763	0.830	0.851	0.806	0.131	0.890	0.949
STGCN	0.877	0.684	0.769	0.825	0.721	0.169	0.825	0.921
Deformer	0.886	0.702	0.786	0.821	0.763	0.156	0.860	0.924
GE-LSTM-Net	**0.923**	**0.806**	**0.859**	**0.880**	**0.840**	**0.114**	**0.918**	**0.963**

Table 3 The performance comparison of finer Fre-Bands for GE-LSTM-Net in SEED-VIG dataset

	Accuracy	Kappa	F1-Score	Recall	Precision	RMSE	COR	AUC
25 Fre-Bands	**0.889** ±0.003	**0.759** ±0.007	**0.914** ±0.003	**0.903** ±0.006	0.926 ±0.009	**0.098** ±0.002	**0.932** ±0.003	0.947 ±0.003
10 Fre-Bands	0.886 ±0.003	0.751 ±0.006	0.912 ±0.003	0.897 ±0.005	**0.927** ±0.009	0.100 ±0.002	0.929 ±0.002	0.943 ±0.002
5 Fre-Bands	0.882 ±0.003	0.746 ±0.006	0.908 ±0.002	0.894 ±0.005	0.923 ±0.007	0.103 ±0.002	0.924 ±0.003	0.939 ±0.002

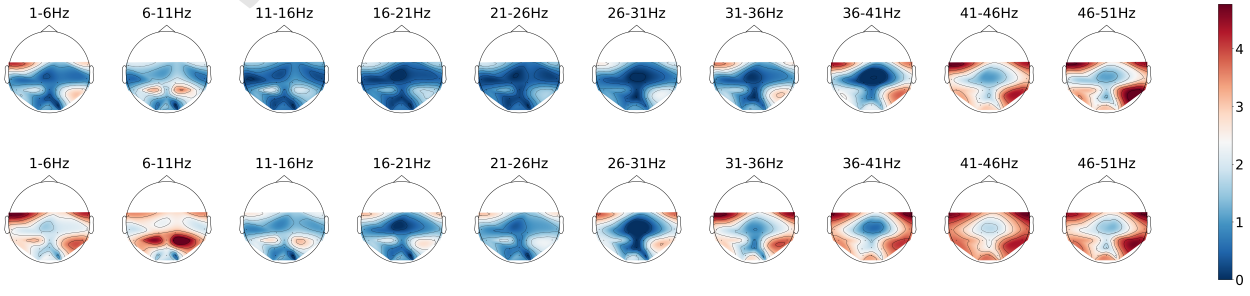

Fig. 8 The Saliency Map of 10 finer Fre-Bands in SEED-VIG dataset, which indicated the importance of different Fre-Bands and brain region for model prediction, with the first row representing the awake state and the second row representing the fatigue state.

Table 4 The impact of the number of GE-LSTM units on the performance of GE-LSTM module in SEED-VIG dataset

	Accuracy	Kappa	F1-Score	Recall	Precision	RMSE	COR	AUC
1 GE-LSTM Unit	0.882 ±0.003	0.742 ±0.006	0.908 ±0.003	0.896 ±0.005	0.921 ±0.008	0.101 ±0.002	0.927 ±0.002	0.939 ±0.002
2 GE-LSTM Units	**0.886** ±0.003	**0.751** ±0.006	**0.912** ±0.003	0.897 ±0.005	**0.927** ±0.011	**0.100** ±0.002	**0.929** ±0.002	**0.943** ±0.002
3 GE-LSTM Units	0.883 ±0.003	0.745 ±0.007	0.909 ±0.003	**0.899** ±0.007	0.920 ±0.010	0.103 ±0.002	0.925 ±0.003	0.940 ±0.003
4 GE-LSTM Units	0.873 ±0.005	0.722 ±0.015	0.902 ±0.004	0.885 ±0.012	0.921 ±0.011	0.109 ±0.003	0.915 ±0.004	0.933 ±0.003

4.5 Impact of the Number of Attention Heads

The number of heads is an important parameter in the Transformer-based model. As d_{model} and d_h are set to 64 and 32, we explored the results for h_n = 1, 2, 4, 8, and 16, to ensure the dimension is divisible by the number of heads. As shown in Fig.9, the experimental results in SEED-VIG dataset indicate that the accuracy of the model remains stable (with an average of 88.5% to 88.7%) when h_n is between 1 and 8. As confirmed by the t-paired test for Accuracy, there are no significant differences between them ($p > 0.05$). This shows that the performance of the model is not sensitive to changes in the number of heads, which is consistent with previous conclusions of other Transformer-based models in the field of EEG decoding [11, 50]. However, when h_n increases to 16, a significant difference is observed in the results compared to when the number of heads is 4, $p < 0.05 (p = 0.024)$. This result aligns with the findings that excessive number of heads can lead to dimensional fragmentation, thereby degrading model performance and there should be a balance between the number of heads and the dimensionality [61].

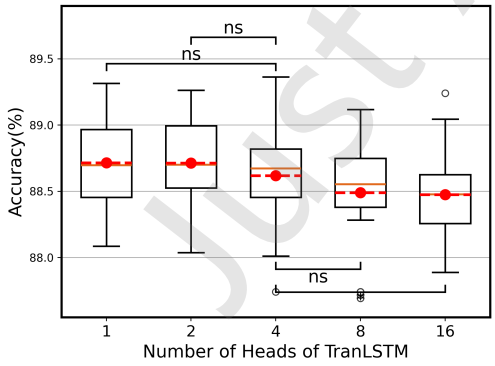

Fig. 9 The box and whisker plot of accuracy for different numbers of heads of the GE-LSTM module in SEED-VIG dataset. The red dashed line and dots indicate the mean value, while the orange line represents the median. The symbol "ns" ($p > 0.05$) and "$*$" ($p < 0.05$) denote significant differences obtained through t-paired test. "ns" means "not significance."

4.6 Effect of PE on the performance of model

PE is an important factor affecting performance in Transformer. It is generally believed that adding an appropriate PE can significantly improve the model's performance. For temporal information extraction, 1D sinusoidal PE is commonly used [48, 50]. For extracting 2D image information, in addition to 1D sinusoidal PE, 2D sinusoidal PE and learnable PE are also used [49]. Therefore, in the Tim-GE-LSTM section, we used 1D sinusoidal PE. After mapping the electrodes to 2D images, we applied the other two PE methods in the Spa-GE-LSTM section.

As shown in Fig.10, after t-paired test, the results show no significant difference between adding PE and not adding PE ($p > 0.05$) in the Tim-GE-LSTM module on the SEED-VIG dataset using five random seeds.In addition, as shown in Table 5, the results after the addition of three different types of PE in the Spa-GE-LSTM module are very similar to the results without any PE. In other words, the role played by PE is not as significant as previously reported in prior studies [48, 49, 50]. We speculate that this may be due to the following reasons. In spatial information extraction, the spatial associations between spatial electrodes are relatively complex and weak. As a result, the 1D/2D sinusoidal PE we employed may struggle to accurately represent this spatial information, and the learnable PE may not effectively capture the relevant spatial information, leading to no improvement and, in fact, causing some interference. However, although no PE is added, the attention mechanism of TranEnlayer can still extract position-independent information, such as the associations between electrodes, and may implicitly learn the electrode topology, thereby enhancing the model's performance. In temporal information extraction, the rationale for adding positional encoding to TranEnlayer stems from the fact that during the parallel matrix multiplication in the attention mechanism, the sequential ordering information across different tokens is inherently discarded. In our proposed module, however, the attention-enhanced outputs are fed into an LSTM for sequential computation. The inherent sequential processing nature of LSTM automatically preserves the order information. Consequently, ablating PE demonstrates negligible impact on performance in our experiments.

Table 5 Performance comparison of different kinds of PE in Spa-GE-LSTM module in SEED-VIG dataset

	Accuracy	Kappa	F1-Score	Recall	Precision	RMSE	COR	AUC
No PE	**0.886** ±0.003	**0.751** ±0.006	**0.912** ±0.003	0.897 ±0.005	**0.927** ±0.011	**0.100** ±0.002	**0.929** ±0.002	**0.943** ±0.002
1D sinusoidal PE	0.884 ±0.003	0.746 ±0.006	0.910 ±0.002	0.896 ±0.006	0.926 ±0.008	0.101 ±0.001	0.927 ±0.002	0.942 ±0.002
2D sinusoidal PE	0.884 ±0.002	0.747 ±0.004	0.910 ±0.002	**0.899** ±0.005	0.922 ±0.008	0.101 ±0.002	0.927 ±0.002	0.942 ±0.003
Learned PE	0.885 ±0.003	0.747 ±0.006	0.911 ±0.003	0.895 ±0.005	0.927 ±0.007	0.101 ±0.002	0.927 ±0.003	0.942 ±0.002

Table 6 Results of ablation experiment of GE-LSTM module in SEED-VIG dataset

Model Variant	Accuracy	Kappa	F1-Score	Recall	Precision	RMSE	COR	AUC
GE-LSTM	**0.886** ±0.003	**0.751** ±0.006	**0.912** ±0.003	**0.897** ±0.005	**0.927** ±0.011	**0.100** ±0.002	**0.929** ±0.002	**0.943** ±0.002
Ablation of TranEnlayer	0.863 ±0.005	0.700 ±0.010	0.894 ±0.004	0.879 ±0.006	0.910 ±0.006	0.114 ±0.001	0.907 ±0.002	0.921 ±0.003
Ablation of LSTM	0.872 ±0.004	0.719 ±0.008	0.902 ±0.003	0.881 ±0.005	0.925 ±0.007	0.107 ±0.001	0.919 ±0.002	0.932 ±0.002
Ablation of GE-LSTM	0.763 ±0.005	0.479 ±0.012	0.818 ±0.005	0.800 ±0.009	0.838 ±0.014	0.171 ±0.001	0.774 ±0.003	0.834 ±0.002

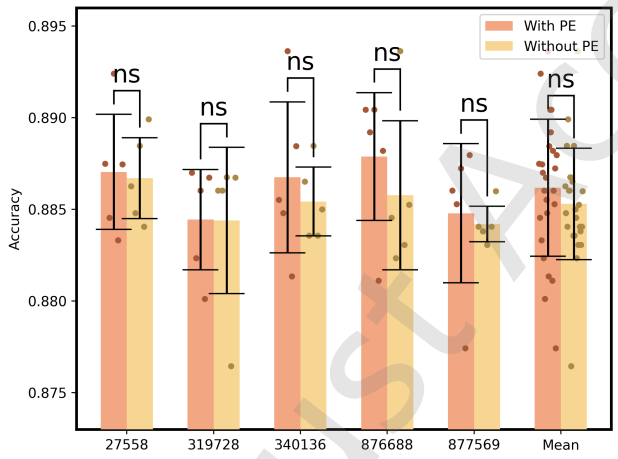

Fig. 10 Accuracy comparison chart of Tim-GE-LSTM with and without PE under 5 different random seeds and average case in SEED-VIG dataset. The numbers on the x-axis represent different random seeds. "ns" means "not significance."

4.7 Ablation Study

In this section, we conducted an ablation study on the validation set of SEED-VIG. Specifically, we first conducted ablation experiments on the GE-LSTM module to explore the contribution of each component to the spatiotemporal feature extraction performance. Then, we performed separate ablation experiments on the DE Feature Fusion module, Spa-GE-LSTM module, and Tim-GE-LSTM module to investigate the specific feature extraction contribution of each module.

In the ablation experiment of GE-LSTM module, we first removed the TranEnlayer and LSTM from the GE-LSTM module, retaining only the LSTM and TranEnlayer, respectively. Subsequently, we replaced the entire module with fully connected layers.As shown in Table 6, the GE-LSTM-Net with the complete structure achieves the best performance with an accuracy of 0.886, a Kappa of 0.751, an F1-score of 0.912, Precision of 0.927, Recall of 0.897, a minimum RMSE of 0.100, and a maximum COR of 0.929, demonstrating strong overall performance. After removing the TranEnlayer and LSTM component, model's accuracy drops to 0.863 and 0.872, Kappa drops to 0.700 and 0.719 and significant declines in both F1-score and COR. When the entire module was removed, the model's performance significantly deteriorates, with accuracy dropping to 0.763, Kappa falling to 0.479, and substantial decreases in F1-score and recall. These results highlights the importance of the TranEnlayer in enhancing the capability of global dependency extraction and the effectiveness of the hybrid architectures (GE-LSTM module) in spatiotemporal feature extraction from EEG signals.

In the ablation experiments for the DE Feature Fusion module, Spa-GE-LSTM module, and Tim-GE-LSTM module, the results are shown in Table 2. When the Spa-GE-LSTM module or Tim-GE-LSTM is ablated individually, the model's accuracy decreases by 1.1% and 2.6%, respectively. And ablating both GE-LSTM modules results in a 12.3% decrease in accuracy. When the DE Feature Fusion module is also ablated, the accuracy further decreases to 0.728 (by 15.8%). This conclusion further demonstrates the impact of each proposed module on performance and highlights the ability of the proposed GE-LSTM module in extracting spatial and temporal features separately.

Table 7 Ablation of DE Fusion module, Spa-GE-LSTM module and Tim-GE-LSTM module

DE Feature Fusion	Spa-GE-LSTM	Tim-GE-LSTM	Accuracy	Change(%)
-	-	-	0.728	-15.8%
√	-	-	0.763	-12.3%
√	√	-	0.840	-2.6%
√	-	√	0.875	-1.1%
√	√	√	0.886	-

4.8 Real-Time and Deployability Research

To explore the real-time performance and deployability of the proposed model, we compare it with six baseline models on three metrics: Floating Point Operations (FLOPs), P99 latency, and parameter count. FLOPs and P99 represent the total arithmetic operations required per inference and the worst-case response time for 99% of requests, respectively. As shown in Table 8, the proposed GE-LSTM-Net achieves a relatively low FLOPs of 218.64 and contains only 1.52M parameters among the compared models. This result demonstrates that the proposed model can be directly deployed on commonly used edge computing platforms, such as the NVIDIA Jetson series and Raspberry Pi series. Under the same hardware conditions, the model achieves a P99 latency of 30.10ms, which has minimal impact in practical vigilance estimation scenarios and issuing low-vigilance warnings. Additionally, in real-world applications, the model's complexity can be further adjusted by modifying the dimensions of the DE Feature Fusion module to balance accuracy and real-time performance requirements when in practical application.

Table 8 Comparison of FLOPs, P99 Latency, and Parameter Counts Across Six baseline models

	FLOPs (M)	P99 (ms)	Params
EEGNet	87.69	0.71	17KB
DeepConvNet	199.95	1.86	584KB
ShallowConvNet	418.06	0.63	131KB
SFT	243.32	34.86	0.58MB
EEG-Conformer	518.89	5.99	1.14MB
TSception	359.46	1.45	45KB
LGG	775.36	5.83	6.92M
ARNN	223.48	36.84	1.67M
STGCN	59.63	3.53	2.59M
Deformer	639.15	6.64	4.23M
GE-LSTM-Net	218.64	30.10	1.52MB

4.9 Limitations and Future Work

The proposed GE-LSTM-Net achieves the best results on SEED-VIG and MMV datasets under the mixed-subject paradigm and some spectral features and associations weights cross multiple brain regions under the state of awake and fatigue are explored. However, in real-world scenarios, subject-independent paradigms, where user calibration data may be unavailable, exhibit greater practical value. A wide array of domain adaptation methods have been developed to align EEG data from different distributions, thereby improving the performance. The method proposed in this study does not involve any domain adaptation techniques, and as such it has not reached the current state-of-the-art performance levels in subject-independent cross-validation. In future work, we will further incorporate domain adaptation algorithms to enhance the model's performance under the subject-independent cross-validation, thereby increasing its applicability in practical scenarios.

5 Conclusion

In this paper, we propose the GE-LSTM-Net, synergizing the Transformer's attention mechanism with LSTM to enhance the extraction of spatiotemporal features in EEG signals. The adopted sample partitioning method combined with the designed DE feature fusion module in GE-LSTM-Net enables the model to effectively extract frequency, spatial, and temporal information from EEG signals. In addition, the proposed GE-LSTM module, serving as the core of the GE-LSTM-Net, can be used to extract both spatial and temporal information and can be easily adopted to other EEG decoding tasks. Furthermore, we prove that the DE features extracted from the finer filtering techniques contains richer vigilance associated information than traditional filtering technique by the visualization of saliency map and can improve the model performance. The comparison with the advanced baseline models on the SEED-VIG and MMV public datasets demonstrates that our model achieves SOTA performance, validating the effectiveness of the model in vigilance estimation. The proposed GE-LSTM-Net can serve as a new benchmark for vigilance estimation.

Dates

Received: 22 August 2025; Revised: 11 Feburary 2026;
Accepted:20 April 2026

References

[1] Raja Parasuraman, Joel S Warm, and Judi E See. Brain systems of vigilance. 1998.

[2] Weidong Dang, Zhongke Gao, Dongmei Lv, Xinlin Sun, and Chichao Cheng. Rhythm-dependent multilayer brain network for the detection of driving fatigue. *IEEE Journal of Biomedical and Health Informatics*, 25(3):693–700, 2020.

[3] Yuanru Guo, Kunping Yang, and Yi Wu. A Multi-Modality Attention Network for Driver Fatigue Detection Based on Frontal EEG, EDA and PPG Signals. *IEEE Journal of Biomedical and Health Informatics*, 2025.

[4] Wei-Long Zheng, Kunpeng Gao, Gang Li, Wei Liu, Chao Liu, Jing-Quan Liu, Guoxing Wang, and Bao-Liang Lu. Vigilance estimation using a wearable EOG device in real driving environment. *IEEE Transactions on Intelligent Transportation Systems*, 21(1):170–184, 2019.

[5] Shengjian Hu, Weining Fang, Haifeng Bao, and Tianlong Zhang. Non-contact detection of mental fatigue from facial expressions and heart signals: A self-supervised-based multimodal fusion method. *Biomedical Signal Processing and Control*, 105:107658, 2025.

[6] Alice Othmani, Aznul Qalid Md Sabri, Sinem Aslan, Faten Chaieb, Hala Rameh, Romain Alfred, and Dayron Cohen. EEG-based neural networks approaches for fatigue and drowsiness detection: A survey. *Neurocomputing*, 557:126709, 2023.

[7] Yibo Zhang, Hui Shen, Ming Li, and Dewen Hu. Brain biometrics of steady-state visual evoked potential functional networks. *IEEE Transactions on Cognitive and Developmental Systems*, 15(4):1694–1701, 2022.

[8] Ming Li, Xue Mei Song, Tao Xu, Dewen Hu, Anna Wang Roe, and Chao-Yi Li. Subdomains within orientation columns of primary visual cortex. *Science Advances*, 5(6):eaaw0807, 2019.

[9] Hui Shen, Zhenfeng Li, Jian Qin, Qiang Liu, Lubin Wang, Ling-Li Zeng, Hong Li, and Dewen Hu. Changes in functional connectivity dynamics associated with vigilance network in taxi drivers. *Neuroimage*, 124:367–378, 2016.

[10] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018.

[11] Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG Conformer: Convolutional Transformer for EEG Decoding and Visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023.

[12] Yi Ding, Neethu Robinson, Su Zhang, Qiuhao Zeng, and Cuntai Guan. Tsception: Capturing temporal dynamics and spatial asymmetry from eeg for emotion recognition. *IEEE Transactions on Affective Computing*, 14(3):2238–2250, 2023.

[13] Dongrui Gao, Pengrui Li, Manqing Wang, Yujie Liang, Shihong Liu, Jiliu Zhou, Lutao Wang, and Yongqing Zhang. CSF-GTNet: A novel multi-dimensional feature fusion network based on Convnext-GeLU-BiLSTM for EEG-signals-enabled fatigue driving detection. *IEEE Journal of Biomedical and Health Informatics*, 2023.

[14] Budi Thomas Jap, Sara Lal, Peter Fischer, and Evangelos Bekiaris. Using EEG spectral components to assess algorithms for detecting fatigue. *Expert Systems with Applications*, 36(2):2352–2359, 2009.

[15] Budi Thomas Jap, Sara Lal, and Peter Fischer. Comparing combinations of EEG activity in train drivers during monotonous driving. *Expert Systems with Applications*, 38(1):996–1003, 2011.

[16] Wei-Long Zheng and Bao-Liang Lu. A multimodal approach to estimating vigilance using EEG and forehead EOG. *Journal of Neural Engineering*, 14(2):026017, 2017.

[17] Wei Wu, Wei Sun, QM Jonathan Wu, Yimin Yang, Hui Zhang, Wei-Long Zheng, and Bao-Liang Lu. Multimodal vigilance estimation using deep learning. *IEEE Transactions on Cybernetics*, 52(5):3097–3110, 2020.

[18] Sheng Dai, Ming Li, Xu Wu, Xiangyu Ju, Xinyu Li, Jun Yang, and Dewen Hu. Contrastive Learning of EEG Representation of Brain Area for Emotion Recognition. *IEEE Transactions on Instrumentation and Measurement*, 2025.

[19] Betul Ay, Ozal Yildirim, Muhammed Talo, Ulas Baran Baloglu, Galip Aydin, Subha D Puthankattil, and U Rajendra Acharya. Automated depression detection using deep representation and sequence learning with EEG signals. *Journal of Medical Systems*, 43:1–12, 2019.

[20] Theerawit Wilaiprasitporn, Apiwat Ditthapron, Karis Matchaparn, Tanaboon Tongbuasirilai, Nannapas Banluesombatkul, and Ekapol Chuangsuwanich. Affective EEG-based person identification using the deep learning approach. *IEEE Transactions on Cognitive and Developmental Systems*, 12(3):486–496, 2019.

[21] Heng Cui, Aiping Liu, Xu Zhang, Xiang Chen, Jun Liu, and Xun Chen. EEG-based subject-independent emotion recognition using gated recurrent unit and minimum class confusion. *IEEE Transactions on Affective Computing*, 14(4):2740–2750, 2022.

[22] Yang Li, Lei Wang, Wenming Zheng, Yuan Zong, Lei Qi, Zhen Cui, Tong Zhang, and Tengfei Song. A novel bi-hemispheric discrepancy model for EEG emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 13(2):354–367, 2020.

[23] Xiaobing Du, Cuixia Ma, Guanhua Zhang, Jinyao Li, Yu-Kun Lai, Guozhen Zhao, Xiaoming Deng, Yong-Jin Liu, and Hongan Wang. An efficient LSTM network for emotion recognition from multichannel EEG signals. *IEEE Transactions on Affective Computing*, 13(3):1528–1540, 2020.

[24] Dongrui Gao, Kejie Wang, Manqing Wang, Jiliu Zhou, and Yongqing Zhang. SFT-Net: A Network for Detecting Fatigue From EEG Signals by Combining 4D Feature Flow and Attention Mechanism. *IEEE Journal of Biomedical and Health Informatics*, 28(8):4444–4455, 2024.

[25] Kuldeep Singh and Jyoteesh Malhotra. Two-layer LSTM network-based prediction of epileptic seizures using EEG spectral features. *Complex & Intelligent Systems*, 8(3):2405–2418, 2022.

[26] Ming Li, Yadong Liu, Fanglin Chen, and Dewen Hu. Including signal intensity increases the performance of blind source separation on brain imaging data. *IEEE Transactions on Medical Imaging*, 34(2):551–563, 2014.

[27] Shurui Li, Miao Tian, Ren Xu, Andrzej Cichocki, and Jing Jin. Decoding continuous motion trajectories of upper limb from EEG signals based on feature selection and nonlinear methods. *Journal of Neural Engineering*, 21(6):066039, 2024.

[28] Yuhao Zhang, Hanying Guo, Yongjiang Zhou, Chengji Xu, and Yang Liao. Recognising drivers' mental fatigue based on EEG multi-dimensional feature selection and fusion. *Biomedical Signal Processing and Control*, 79:104237, 2023.

[29] Yunhe Liu, Zirui Xiang, Zhixin Yan, Jianxiu Jin, Lin Shu, Lulu Zhang, and Xiangmin Xu. CEEMDAN fuzzy entropy based fatigue driving detection using single-channel EEG. *Biomedical Signal Processing and Control*, 95:106460, 2024.

[30] Brahim Hamadicharef, Haihong Zhang, Cuntai Guan, Chuanchu Wang, Kok Soon Phua, Keng Peng Tee, and Kai Keng Ang. Learning EEG-based spectral-spatial patterns for attention level measurement. In *2009 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1465–1468. IEEE, 2009.

[31] Chin-Teng Lin, Ruei-Cheng Wu, Tzyy-Ping Jung, Sheng-Fu Liang, and Teng-Yi Huang. Estimating driving performance based on eeg spectrum analysis. *EURASIP Journal on Advances in Signal Processing*, 2005(19):521368, 2005.

[32] Haitham M Al-Angari and Alan V Sahakian. Use of sample entropy approach to study heart rate variability in obstructive sleep apnea syndrome. *IEEE Transactions on Biomedical Engineering*, 54(10):1900–1904, 2007.

[33] Enrique Molina, Daniel Sanabria, Tzyy-Ping Jung, and Angel Correa. Electroencephalographic and peripheral temperature dynamics during a prolonged psychomotor vigilance task. *Accident Analysis & Prevention*, 126:198–208, 2019.

[34] Li-Chen Shi, Ying-Ying Jiao, and Bao-Liang Lu. Differential entropy feature for EEG-based vigilance estimation. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6627–6630. IEEE, 2013.

[35] Kangning Wang, Shuang Qiu, Wei Wei, Yukun Zhang, Shengpei Wang, Huiguang He, Minpeng Xu, Tzyy-Ping Jung, and Dong Ming. A multimodal approach to estimating vigilance in SSVEP-based BCI. *Expert Systems with Applications*, 225:120177, 2023.

[36] Wei Zheng and Bo Pan. A spatiotemporal symmetrical transformer structure for EEG emotion recognition. *Biomedical Signal Processing and Control*, 87:105487, 2024.

[37] Zequan Lian, Tao Xu, Zhen Yuan, Junhua Li, Nitish Thakor, and Hongtao Wang. Driving fatigue detection based on hybrid electroencephalography and eye tracking. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[38] Yi Ding, Yong Li, Hao Sun, Rui Liu, Chengxuan Tong, Chenyu Liu, Xinliang Zhou, and Cuntai Guan. EEG-Deformer: A Dense Convolutional Transformer for Brain-Computer Interfaces. *IEEE Journal of Biomedical and Health Informatics*, 29(3):1909–1918, 2025.

[39] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017.

[40] Peiliang Gong, Ziyu Jia, Pengpai Wang, Yueying Zhou, and Daoqiang Zhang. ASTDF-Net: Attention-Based Spatial-Temporal Dual-Stream Fusion Network for EEG-Based Emotion Recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 883–892, New York, NY, USA, 2023. Association for Computing Machinery.

[41] Peiliang Gong, Pengpai Wang, Yueying Zhou, and Daoqiang Zhang. A Spiking Neural Network With Adaptive Graph Convolution and LSTM for EEG-Based Brain-Computer Interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:1440–1450, 2023.

[42] Fangyao Shen, Guojun Dai, Guang Lin, Jianhai Zhang, Wanzeng Kong, and Hong Zeng. EEG-based emotion recognition using 4D convolutional recurrent neural network. *Cognitive Neurodynamics*, 14:815–828, 2020.

[43] Xiangyu Ju, Xu Wu, Sheng Dai, Ming Li, and Dewen Hu. Domain adversarial learning with multiple adversarial tasks for EEG emotion recognition. *Expert Systems with Applications*, 266:126028, 2025.

[44] Mohammed Alghanim, Hani Attar, Khosro Rezaee, Mohamadreza Khosravi, Ahmed Solyman, and Mohammad A Kanan. A hybrid deep neural network approach to recognize driving fatigue based on EEG signals. *International Journal of Intelligent Systems*, 2024(1):9898333, 2024.

[45] Kun Yang, Keze Zhang, Yubin Hu, Jing Xu, Bing Yang, Wanzeng Kong, and Jianhai Zhang. Adaptive multi-branch CNN of integrating manual features and functional network for driver fatigue detection. *Biomedical Signal Processing and Control*, 102:107262, 2025.

[46] Chin-Teng Lin, Chun-Hsiang Chuang, Yu-Chia Hung, Chieh-Ning Fang, Dongrui Wu, and Yu-Kai Wang. A driving performance forecasting system based on brain dynamic state analysis using 4-D convolutional neural networks. *IEEE Transactions on Cybernetics*, 51(10):4959–4967, 2020.

[47] Kefa Wang, Xiaoqian Mao, Yuebin Song, and Qiuyu Chen. EEG-based fatigue state evaluation by combining complex network and frequency-spatial features. *Journal of neuroscience methods*, 416:110385, 2025.

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[49] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[50] Ruilin Li, Minghui Hu, Ruobin Gao, Lipo Wang, Ponnuthurai Nagaratnam Suganthan, and Olga Sourina. TFormer: A time–frequency Transformer with batch normalization for driver fatigue recognition. *Advanced Engineering Informatics*, 62:102575, 2024.

[51] Wei Wei, Kangning Wang, Shuang Qiu, and Huiguang He. A MultiModal Vigilance (MMV) dataset during RSVP and SSVEP brain-computer interface tasks. *Scientific data*, 11(1):867, 2024.

[52] Udo Trutschel, Bill Sirois, David Sommer, Martin Golz, and Dave Edwards. PERCLOS: An alertness measure of the past. In *Driving Assessment Conference*, volume 6. University of Iowa, 2011.

[53] Yi Ding, Neethu Robinson, Chengxuan Tong, Qiuhao Zeng, and Cuntai Guan. LGGNet: Learning From Local-Global-Graph Representations for Brain–Computer Interface. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7):9773–9786, 2024.

[54] Salim Rukhsar and Anil K. Tiwari. ARNN: Attentive recurrent neural network for multi-channel EEG signals to identify epileptic seizures. *Neurocomputing*, 620:129203, 2025.

[55] Kangning Wang, Shuang Qiu, Wei Wei, Yukun Zhang, Shengpei Wang, Huiguang He, Minpeng Xu, Tzyy-Ping Jung, and Dong Ming. A multimodal approach to estimating vigilance in SSVEP-based BCI. *Expert Systems with Applications*, 225:120177, 2023.

[56] XinWang Song, DanDan Yan, LuLu Zhao, and LiCai Yang. LSDD-EEGNet: An efficient end-to-end framework for EEG-based depression detection. *Biomedical Signal Processing and Control*, 75:103612, 2022.

[57] Yifan Jiang, Ning Chen, and Jing Jin. Detecting the locus of auditory attention based on the spectro-spatial-temporal analysis of EEG. *Journal of Neural Engineering*, 19(5):056035, 2022.

[58] Mingyi Sun, Weigang Cui, Shuyue Yu, Hongbin Han, Bin Hu, and Yang Li. A dual-branch dynamic graph convolution based adaptive transformer feature fusion network for eeg emotion recognition. *IEEE Transactions on Affective Computing*, 13(4):2218–2228, 2022.

[59] Chin-Teng Lin, Kuan-Chih Huang, Chun-Hsiang Chuang, Li-Wei Ko, and Tzyy-Ping Jung. Can arousing feedback rectify lapses in driving? Prediction from EEG power spectra. *Journal of Neural Engineering*, 10(5):056024, 2013.

[60] Mastaneh Torkamani-Azar, Sumeyra Demir Kanik, Serap Aydin, and Mujdat Cetin. Prediction of reaction time and vigilance variability from spatio-spectral features of resting-state EEG in a long sustained attention task. *IEEE Journal of Biomedical and Health Informatics*, 24(9):2550–2558, 2020.

[61] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32, 2019.