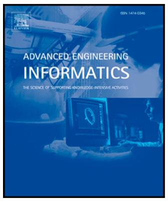
Contents lists available at ScienceDirect

Advanced Engineering Informatics

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



Full length article

TFormer: A time–frequency Transformer with batch normalization for driver fatigue recognition

Ruilin Li [a], Minghui Hu [a], Ruobin Gao [b], Lipo Wang [a,*], P.N. Suganthan [c], Olga Sourina [d]

[a] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[b] School of Civil and Environmental Engineering, Nanyang Technological University, Singapore
[c] KINDI Center for Computing Research, College of Engineering, Qatar University, Doha, Qatar
[d] Fraunhofer, Nanyang Technological University, Singapore

ARTICLE INFO

Keywords:
Advanced human-cybernetic interfaces (HCI)
Driver fatigue
Time-frequency Transformer (TFormer)
Batch normalization (BN)
Cross-subject Electroencephalogram (EEG) decoding

ABSTRACT

Within the framework of the advanced human-cybernetic interfaces (HCI), Cross-subject electroencephalogram (EEG)-based driver fatigue recognition is emerging as a pivotal application in the paradigm of Industry 5.0. Recognizing the importance of ensuring driver safety through proactive monitoring, it is essential to offer a general EEG decoding system to improve road safety. This work investigated the use of Transformers for the challenging cross-subject EEG decoding task due to the great success the Transformers have achieved in various applications. Previous research focused on using Transformers to capture global temporal information, but less work targeted global frequency-domain patterns. Furthermore, in order to leverage a standard Transformer architecture grounded in natural language processing for EEG decoding, it is imperative to account for inherent characteristics in EEG and make pertinent adjustments accordingly. In this work, we proposed a time–frequency Transformer (TFormer) that can automatically learn the global time–frequency patterns from raw EEG data. TFormer consisted of three components: convolutional stems for input embedding, time–frequency multi-head cross-attention (TF-MCA) for integrating time-domain patterns into frequency points, and self-attention to further learn global time-frequency patterns. Moreover, we analyzed TFormer's internal settings and found batch normalization (BN) more suitable for cross-subject EEG decoding than layer normalization (LN). The experiment results demonstrated the superiority of the proposed model compared to existing methods. Overall, our work contributes to the development of Transformer models in EEG decoding and illustrates a different way to leverage Transformers for decoding raw EEG data.

1. Introduction

As we transition into the paradigm of Industry 5.0, ensuring driver safety through proactive monitoring becomes paramount. Among different human mental states, driver fatigue is a serious problem that can result in accidents and fatalities on the road [1]. A study by Byeon et al. [2] found that driver fatigue was responsible for a significant proportion (10%–15%) of traffic accidents. Thus, it is crucial to develop a human-cybernetic interface (HCI) to monitor and further improve road safety.

Leveraging bio-signals and machine learning algorithms [3,4] represents a cutting-edge direction in the evolution of advanced HCIs, where the advantages of objective and real-time responses are beneficial to significantly improve the operator's performance in dynamic systems. In recent years, electroencephalogram (EEG) signals, which can reflect human brain activities, have been widely adopted to develop driver fatigue recognition systems [5,6]. In practice, the EEG-based system can function as a means of signaling impending fatigue states, or as an instructional mechanism aimed at augmenting operators' non-technical skills. In the evolving landscape of advanced HCI, the subject variability [7] impedes the transition of the EEG-based system from the laboratory setting (subject-dependent model) to the practice setting (subject-independent model). In the context of EEG decoding, a subject-dependent setting implies that models are exclusively trained and tested on data from the same subject, which often achieves higher decoding accuracy but may not be practical for widespread application due to the necessity for individual-specific data collection and model training. Conversely, a subject-independent setting, often referred to as a cross-subject setting, characterizes models designed for broad applicability across diverse individuals. These models are trained on datasets from

[*] Corresponding author.
E-mail addresses: ruilin001@e.ntu.edu.sg (R. Li), e200008@e.ntu.edu.sg (M. Hu), gaor0009@e.ntu.edu.sg (R. Gao), ELPWang@ntu.edu.sg (L. Wang), p.n.suganthan@qu.edu.qa (P.N. Suganthan), EOSourina@ntu.edu.sg (O. Sourina).

https://doi.org/10.1016/j.aei.2024.102575
Received 3 November 2023; Received in revised form 15 March 2024; Accepted 25 April 2024
Available online 20 May 2024

multiple subjects to ensure robust generalization, enabling their deployment on unseen subjects without requiring subject-specific calibration. A calibration-free model capable of direct applications to unseen subjects stands out as one of the pivotal challenges in realizing the full potential of advanced HCI.

In the domain of EEG-based fatigue detection, a diverse array of machine and deep learning algorithms have been investigated, encompassing traditional classifiers [8,9], Convolutional Neural Networks (CNNs) [10], and Recurrent Neural Networks (RNNs) [11]. CNNs, despite their widespread use, are constrained by their inherently limited receptive fields, which may hinder their capacity to encapsulate broader contextual information within EEG signals. RNNs, on the other hand, are tailored for sequential data processing but often encounter the challenge of capturing long-range dependencies within the data. Additionally, this sequential processing nature impedes their capacity for parallelization, adversely affecting training efficiency. With the development of deep learning, Transformers have recently gained a lot of attention. This innovative deep-learning architecture was introduced by Vaswani et al. [12] and has since shown remarkable success in different fields such as computer vision [13]. Distinct from previously used models, the Transformer architecture employs a self-attention mechanism, enabling the model to process entire data sequences simultaneously. This attribute is particularly advantageous for EEG decoding, where discerning global dependencies across temporal, spectral, and spatial dimensions is crucial. Given these advantages, there has been an increasing interest in applying Transformer models to various EEG-based applications, such as emotion recognition [14] and motor imagery classification [15]. However, the exploration of Transformer models for driver fatigue detection, especially within cross-subject scenarios, remains an underexplored area of research.

In the usage of Transformers, recent works mostly performed self-attention over the time axis, aiming to extract the global temporal patterns. However, frequency-domain information was usually neglected in the previous works. Frequency-domain information is significant in EEG analysis [16] and previous studies have shown that frequency-domain features were beneficial for EEG decoding [17]. Consequently, there is a need to further explore the use of Transformer architecture for learning global frequency-domain patterns and time-frequency patterns, leveraging its inherent advantages in this context.

Furthermore, although different variants of Transformers have shown outstanding performance in different EEG decoding tasks, the design of specific components in the Transformers for the cross-subject setting remains to be further investigated. The first aspect to consider is the selection of the position embedding function. While learnable 1D position embeddings are currently popular in the field [13], it is crucial to take into account the limited size of EEG datasets when fitting large-scale Transformers. Secondly, layer Normalization (LN) [18], used in original Transformers, was developed to address the challenges posed by the varying sequence lengths in text data, proving to be more effective than batch normalization (BN) [19] for this purpose [18]. Despite efforts to standardize sequence lengths through padding, the intrinsic variability in text length can result in BN's statistical computations failing to accurately capture the true characteristics of the text data due to the distortion introduced by the padding. Moreover, initial attempts to apply BN to NLP tasks encountered significant performance degradation [20]. In other fields such as computer vision, Transformer-based architectures have predominantly adopted LN from the original design. However, the advantages of BN for vision tasks have prompted some researchers to experiment with integrating BN into Transformer models in lieu of LN, yielding positive results [21]. Transitioning to EEG cross-subject decoding, previous works underscored the crucial role of BN's statistical parameters in enhancing both the robustness and accuracy of cross-subject classification tasks [10]. Therefore, the selection of the normalization layer should also be carefully explored for EEG decoding tasks.

The research questions in this work are articulated as follows: (1) enhancing the learning and integration of global time-domain and frequency-domain features by capitalizing on the capabilities of Transformers within the context of cross-subject EEG-based fatigue detection task, thereby surpassing the efficacy of previous methodologies. (2) investigating the optimal choices for the components of Transformers to ensure maximum effectiveness in this specific task. To deal with them, this work proposed a time-frequency Transformer (TFormer) catered to the characteristics of EEG decoding and can automatically learn global time–frequency patterns. Specifically, considering the intrinsic nature of non-stationary and low signal-to-noise ratio (SNR), the Transformers that lack inductive biases inherent to CNNs may not generalize well when trained directly on raw EEG data which are typical of small sizes. Therefore, the convolutional stems were first designed to convert the raw EEG data into feature maps that were well-suited for further learning tasks. Following that, this work proposed to integrate the temporal patterns and frequency-domain patterns by utilizing the benefits of the cross-attention mechanism, which has demonstrated significant potential in the fusion of diverse modalities [22]. Specifically, a time-frequency multi-head cross-attention (TF-MCA) was introduced to learn the global time-frequency features by integrating the time-domain patterns into each frequency point. Then, by passing the obtained time–frequency patterns through self-attentions, the model can further learn global dependencies with enhanced representation capabilities. In addition, we found that batch normalization (BN), which can exploit the statistics of individual subjects, was more suitable for the TFormer used in the task of cross-subject EEG decoding.

The contributions of this work are summarized as follows:

- This work proposed a TFormer, which advanced the extraction of the global time-frequency patterns by introducing a TF-MCA module. Experiment results demonstrated the superior performance of the TFormer over the strong baselines.
- This work designed general convolutional stems for both domains, which were effective in dealing with the intrinsic nature of EEG data.
- This work investigated the internal settings for EEG decoding, including the selection of position embedding functions and normalization layer. We found that using BN layers in Transformers is more suitable for EEG decoding.

The remaining sections of this paper are organized as follows: Related works are presented in Section 2. Section 3 introduces the proposed TFormer in detail. Section 4 presents information about the datasets, comparison results, and ablation study. Then, the advantages of the proposed model are discussed in Section 5. Finally, conclusions are drawn in Section 6.

2. Related work

2.1. EEG signal classification

Various machine learning algorithms have been used to decode EEG signals for fatigue recognition by first identifying hand-engineering features and then applying classifiers. Several hand-engineering features have been investigated for different EEG paradigms, including common spatial filter patterns (CSP) for MI classification [23], differential entropy (DE) for emotion recognition [24], and sample entropy for mental stress recognition [25]. For fatigue detection, power spectral density (PSD) has become a popular feature in recent studies. Ye et al. [26] identified fatigue state by exploiting PSD features. Furthermore, Gao et al. [27] exploited the PSD features to characterize the traits related to driver fatigue states. In their works, CNNs were exploited as the classifier. In addition, attempts were also made to utilize classical classifiers such as support vector machines (SVM) [17] as the classifier. Although widely used, this conventional pipeline can present notable drawbacks, including the need for time-consuming feature engineering and limitations in the information contained within the features.

The success of deep learning in image processing has led to similar advancements in end-to-end EEG decoding using DNNs. Compared to conventional hand-engineering feature-based models, DNNs with stronger feature learning capability from raw EEG data can offer automatic feature extraction from raw data. Among the different architectures of DNNs, deep convolutional neural networks have demonstrated outstanding performance in automatically learning from EEG signals. Schirrmeister et al. [28] studied raw EEG data decoding by using CNN models. They proposed both shallow and deep CNN architectures and achieved superior performance on motor imagery (MI) classification tasks. Following that, Lawhern et al. [29] introduced a compact CNN named EEGNet for raw EEG data decoding, which has been utilized in different EEG paradigms. Although these two works were proposed for other EEG paradigms, they are also applicable for fatigue recognition tasks and widely used as the baseline methods [7]. Regarding recent fatigue recognition works, Cui et al. [30] proposed an InterpretableCNN (ICNN) that performs spatial and temporal convolution operations and achieved superior performance compared to previous models such as EEGNet in drowsiness classification based on EEG signals. Recent literature also showed that ICNN could be applied to other EEG paradigms [31].

In the application of DNNs for fatigue recognition, frequency-domain data as inputs has also been investigated, which mainly employed hand-engineering spectral features such as DE and PSD. For instance, Shi et al. [32] proposed a convolutional autoencoder to learn frequency features from extracted spectral features. Gao et al. [33] transformed the DE of five frequency bands of EEG signals into a 4-dimensional feature tensor. The attention module and long short-term memory (LSTM) network were used to combine the spatial-frequency-temporal features for fatigue recognition. Despite the advancements in DNNs for time–frequency-domain EEG decoding, the learning of the global time-frequency patterns from raw EEG data remains to be further explored.

2.2. Transformers

The Transformer architecture is based on the self-attention mechanism that enables the model to capture long-range dependencies [12]. Motivated by the success of Transformers in other domains [13,34], researchers started to utilize the Transformer model and the self-attention mechanism in various EEG paradigms. Gong et al. [35] leveraged the integration of convolutional layers with Transformer architectures to learn spatial, spectral, and temporal dynamics within DE features, tailored for emotion recognition. Similarly, Zeynali et al. [36] harnessed hand-crafted PSD features to train a Transformer-based model, which was named 'Spectral Transformer'. In the time domain, the Transformer model was used to learn the features directly from the raw EEG data. Ensemble learning was used to combine them to perform the final classification. The end-to-end training of Transformers has also been explored in recent works. Siddhad et al. [37] showcased the Transformers' utility across diverse tasks, underscoring their efficacy in learning raw EEG data. Song et al. [15] proposed a convolutional Transformer (Conformer) to perform end-to-end motor imagery (MI) recognition. For the MI task, Xie et al. [38] also proposed a series of Transformer models that can capture temporal or spatial information separately. Moreover, the application of Transformers extended to EEG-based diagnostics, including Alzheimer detection [39] and seizure prediction [40]. Transformer has also been applied to reduce individual differences by combining with transfer learning techniques. For instance, Song et al. [41] proposed a global adaptive Transformer that used an attention-based adaptor to align source features to the target domain. The model leveraged both adversarial loss and adaptive center loss for enhanced domain feature alignment.

A wide range of alternatives have been investigated in related areas regarding Transformer components, specifically normalization layers and positional embeddings. Shen et al. [20] critically examined the inapplicability of BN in NLP tasks and proposed an enhanced BN variant, PowerNorm, tailored for such applications. Yao et al. [21] investigated the use of BN within computer vision tasks, affirming its viability within Transformer architectures for visual tasks. Regarding position embeddings, fixed sine, and cosine position embeddings [12], as well as learned position embeddings [13], have proven effective in previous studies. These previous works highlight the importance of exploring the most suitable options for these components in EEG-based classification tasks.

2.3. Research gaps

The recent advances in EEG-based fatigue detection have predominantly been occupied by CNN models. However, their constrained receptive fields might impede the generalization capability. With their advantage of capturing extensive dependencies, Transformers present a promising alternative. Yet, the application of Transformers in fatigue detection remains underexplored. Additionally, while some studies have integrated time-domain and frequency-domain features within a Transformer-based model, the frequency-domain insights often stem from hand-engineered features, potentially limiting the depth of learned information. Additionally, the approach of learning time and frequency features separately before combining them only indirectly represents time-frequency information. The direct learning of time–frequency features from data still requires more in-depth exploration. Finally, a critical examination and optimization of Transformer components for EEG-based applications warrant further investigation.

3. Methodology

The architecture of the proposed TFormer is shown in Fig. 1. In this section, the designed convolutional stems and Transformer blocks are presented. In addition, the selection of the normalization layer in the TFormer is introduced.

3.1. Input embedding and position embedding

To exploit the Transformer to learn from the raw time- and frequency-domain data, the input embedding was first performed to generate the 1D sequence of "token" embedding as the input. In EEG decoding of this work, "token" represents the time steps in the time domain and the frequency points in the frequency domain. The fast Fourier transformation (FFT) was utilized to transform the time-domain EEG signals $\boldsymbol{x}_t \in \mathbb{R}^{N\times M\times T}$ into the frequency-domain data $\boldsymbol{x}_f \in \mathbb{R}^{K\times N\times F}$, where N, M, T and F represents the size of mini-batch, EEG channels, time steps and frequency points, respectively.

We designed a convolutional stem for input embedding based on the typical spatial and temporal filters which have been shown to be promising for EEG decoding in recent works [42]. Specifically, when the time-domain EEG data were used as the input, the convolutional stem stacked a $C \times 1$ pointwise convolution (spatial filters) and a $1 \times K$ temporal convolution (temporal filter), where K represents the kernel size. Following that, the Gaussian Error Linear Unit (GELU) activation function and batch normalization (BN) were performed. For the frequency-domain input, the pointwise convolution was only used in the convolutional stem, followed by the GELU action layer and BN layer. Since the frequency-domain data can potentially separate the noisy and beneficial frequency points, there was no operation along the frequency axis. The outputs of the temporal and spectral stem are denoted as $\boldsymbol{k}_t \in \mathbb{R}^{N\times C\times T_{Conv}}$ and $\boldsymbol{q}_f \in \mathbb{R}^{N\times C\times F}$, respectively, where T_{Conv} represents the size along the time axis after the temporal stem. For the learning of the following Transformer blocks along the time axis and

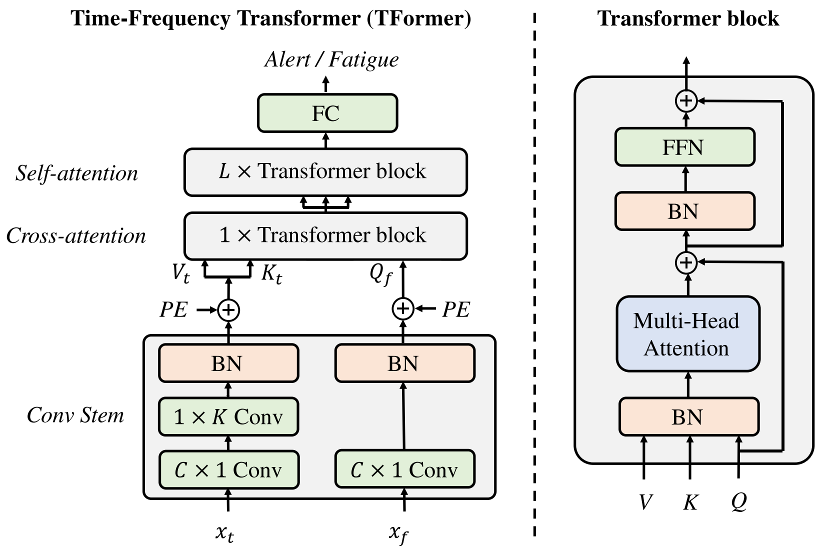


Fig. 1. The overview of the TFormer. The terms 'Q', 'K', and 'V' represent queries, keys, and values, respectively. The cross-attention Transformer block used Q_f, K_t, and V_t as the inputs. For the self-attention Transformer block, the queries, keys, and values all came from the learned time-frequency patterns. L represents the number of self-attention Transformer blocks. The term 'FC' represents a fully-connected layer that was used for classification. The term 'PE' represents the petition embeddings.

frequency axis, the outputs were further reshaped to (N, T_{Conv}, C) and (N, F, C) in the time domain and the frequency domain, respectively.

Upon the input embeddings obtained by the stems, the Transformer was used to learn the global time-frequency patterns along the time and frequency axes. Position embeddings were added to the input embeddings of both domains to exploit the order information of the learned patterns in both domains. There are two types of position embeddings, learned and fixed [43]. Considering the limited EEG data for training the learnable 1D position embeddings, the fixed position embeddings calculated by sine and cosine functions were used. Specifically, the Eqs. (1) and (2) were followed for calculating position embeddings.

$$PE_{(pos,2i)} = sin(\frac{pos}{10000^{2i/C}}), \tag{1}$$

$$PE_{(pos,2i+1)} = cos(\frac{pos}{10000^{2i/C}}), \tag{2}$$

where pos represents the position and i represents the index of the dimension. We also investigated the performance of using learnable 1D position embeddings. The comparison is performed in Section 4.7.

3.2. Transformer blocks

In the present study, the pre-norm Transformer architecture [44] was employed. This particular configuration integrates layer normalization within the residual connections and positions it anterior to the multi-head attention layer. Xiong et al. [45] provided the evidence that training using the pre-norm Transformer architecture can be more efficient, attributed to the well-behaved gradients within the pre-norm Transformer framework. This architectural paradigm has been widely adopted in recent advancements such as Vision Transformers [13].

Frequency-domain analysis has been shown to be promising for EEG processing [16] since the useful patterns can appear in specific frequency points, reducing the impact of inferior information. Therefore, we proposed integrating the powerful time-domain patterns by temporal convolution into each frequency point. The obtained overall time-frequency patterns with global frequency-domain information can have the merits of both domains and be beneficial to EEG decoding. In this work, a time-frequency multi-head cross-attention (TF-MCA) was proposed to establish the connection between each frequency point and the sequential time-domain patterns. In TF-MCA, the queries were obtained by $\boldsymbol{Q_f} = \boldsymbol{q_f} + PE$ and the identical keys and values are obtained by $\boldsymbol{K_t} = \boldsymbol{V_t} = \boldsymbol{k_t} + PE$ as keys and values. Specifically, the TF-MCA is described by Eq. (3).

$$\begin{aligned} &TF - MCA(\boldsymbol{Q_f}, \boldsymbol{K_t}, \boldsymbol{V_t}) = Concat(head_1, \dots, head_H)\boldsymbol{W^O}, \\ &where\ head_h = Attention(\boldsymbol{Q_f}\boldsymbol{W_h^{Q_f}}, \boldsymbol{K_t}\boldsymbol{W_h^{K_t}}, \boldsymbol{V_t}\boldsymbol{W_h^{V_t}}), \end{aligned} . \tag{3}$$

where $\boldsymbol{W^O} \in \mathbb{R}^{T_{Conv} \times T_{Conv}}$, $\boldsymbol{W_h^{Q_f}} \in \mathbb{R}^{F \times F/h}$ and $\boldsymbol{W_h^{K_t}}, \boldsymbol{W_h^{V_t}} \in \mathbb{R}^{T_{Conv} \times T_{Conv}/h}$ represent the weight matrices of linear projections. The term h represents the number of attention heads.

To introduce non-linear mapping and encourage feature quality, the feed-forward network (FFN) was utilized on the learned time–frequency representations. Denoting the outputs of TF-MCA by $\boldsymbol{Y}$, the function of the FFN is described by Eq. (4).

$$FFN(\boldsymbol{Y}) = \sigma(\boldsymbol{Y}\boldsymbol{W_1} + \boldsymbol{b_1})\boldsymbol{W}2 + \boldsymbol{b_2}, \tag{4}$$

where $\boldsymbol{W_1}$, $\boldsymbol{W_2}$ and $\boldsymbol{b_1}$, $\boldsymbol{b_2}$ are the weights and the biases of the fully-connected layers.

Subsequently, multi-head self-attention (MHA) was utilized to further learn the global dependencies of the time–frequency representations, followed by the FFN. It is worth noting that the number of blocks (MHA+FFN) is a hyper-parameter of this work.

3.3. Batch normalization (BN) instead of layer normalization (LN)

Normalization technique is essential in Transformers to stabilize the training process and improve the model performance. This work explored the suitable normalization technique for the Transformers used for EEG decoding. The commonly used BN in CNN models for EEG-based classification and popular LN in standard Transformers were compared.

The main difference between BN and LN lies in the dimension over which the normalization is performed. Taking a Tensor input to the standard Transformers with the shape of (N, L, C), where L represents the sequence length, BN computes the mean and variance of the activations across (N, L) in each training batch, while LN performs the computing across (C) in each training sample. Although LN has become the most popular technique used in Transformers, it mainly considers the normalization for individual samples and neglects the statistics among the training samples which are important for cross-subject EEG decoding [10]. Based on the analysis, this work employed the BN in the TFormer introduced above. Furthermore, in the BN layer, rather than relying on the moving average and moving variance obtained during training, this work automatically used the mean and variance of the testing set to normalize the activations of each layer during testing, which was advantageous in cross-subject EEG decoding where the distribution of the input data during testing differs from that of the training data. This setting is denoted as subject-specific BN. The comparison between the use of BN and LN in the proposed TFormer is presented in Section 4.6.

4. Experiments

4.1. Introduction of sustained-attention driving (SAD) dataset

The sustained-attention driving (SAD) dataset [46] was utilized in this work. In the experiment, lane-departure events were randomly induced to make the car drift from the original cruising lane towards the left or right side (deviation onset). Each participant was instructed to quickly compensate for this perturbation by steering the wheel (response onset) to cause the car to move back to the original cruising lane (response offset). Deviation onset, response onset, and response offset events were all included in a complete trial. During the experiment, the EEG activity of each subject was recorded using a 32-channel Quik-Cap following the International 10–20 system of electrode placement. Processed data provided by Cao et al. [46] were used in this work. Specifically, the pre-processing steps included bandpass filtering and artifact rejection. The bandpass finite impulse response filters of 1-50 Hz were applied to remove the low-frequency direct current drifts and power line noise. For artifact rejection, the apparent eye blink contamination in the EEG signals was manually removed by visual inspection. Following that, the artifacts were removed by the Automatic Artifact Removal plug-in for EEGLAB, which provided automatic correction of ocular and muscular artifacts in the EEG signals.

In this study, we utilized three-second EEG data preceding the deviation onset, as commonly employed in previous research [17,30], to classify upcoming lane-departure events. Following the approach of Wei et al. [17], we labeled data using local reaction time (RT) and global RT. The RT was defined as the duration between the deviation onset and the response onset. For each participant, the RT for each lane-departure event was considered the local RT, while the global RT was determined by averaging the RTs from all trials within a 90-s window preceding the upcoming deviation onset.

The data labeling process was conducted in accordance with the process described by Wei et al. [17]. Specifically, we calculated the 'alert-RT' as the 5th percentile of local RTs for each driving session. If both the local and global RTs were shorter than 1.5 times the alert-RT, the corresponding EEG data was labeled as 'alert'. Conversely, if both RTs were longer than 2.5 times the alert-RT, the data was labeled as 'fatigue'. We excluded transitional states with moderate performance and did not consider the neutral state in this study. When multiple datasets were available for a subject, we selected the most balanced one for filtering. Subsequently, we down-sampled the data to 128 Hz. Ultimately, we obtained a balanced driver fatigue dataset comprising 2022 samples from 11 participants. The data size for a single sample was 30 (channels) × 384 (time steps).

4.2. Introduction of SEED-VIG dataset

In this study, we also utilized the publicly available SEED-VIG dataset, which was obtained from a monotonous driving task using a virtual reality-based simulated driving system [47]. The EEG data were recorded using the Neuroscan system, and electrode placement followed the International 10–20 electrode system. Alongside the EEG recording, the percentage of eye closure (PERCLOS) [48] was measured using Senso-Motoric Instrument eye-tracking glasses, employing a window size of 60 s and a moving step of 10 s.

We further down-sampled the EEG signals to 128 Hz and applied a low-pass filter of 1 Hz. We extracted 3-s EEG samples prior to each PERCLOS evaluation event, adhering to the procedure outlined by Zheng et al. [47]. Following [47], samples were labeled as 'alert' when PERCLOS was below 0.35, and as 'fatigue' when PERCLOS exceeded 0.7. We discarded samples with PERCLOS values within the intermediate range. Moreover, sessions with fewer than 50 samples for either class were discarded, and we balanced the classes in each session by selecting the most representative 'alert' and 'fatigue' samples. Finally, we compiled a balanced driver fatigue dataset containing 3536 samples from 12 participants. The data size for a single sample was 17 (channels) x 384 (time steps).

4.3. Experiment settings and hyper-parameter optimization

The codes were implemented and tested on Python 3.8.16 with a GeForce RTX 2080 Ti. Pytorch framework was employed in this work.

To evaluate the performance of the proposed TFormer on the challenging cross-subject driver fatigue recognition task, the leave-one-subject-out (LOSO) cross-validation (CV) was conducted. The models were run ten times with different random seeds. The final average results were reported. We employed a stratified sampling approach based on individuals and classes to divide the dataset, except for the testing data, into training and validation sets. Specifically, we randomly extracted 20% of the data from each individual and each class as the validation set, while the remaining 80% of data was used as the training set. During training, the model was trained on the training set for 100 epochs and the validation set was evaluated in each epoch. The model was then re-trained on the combination of the training set and the validation set for a fixed number of epochs where the highest validation accuracy was achieved. The rest hyper-parameter settings for back-prop-based models were the same. Specifically, Adam optimizer was set as momentum $\beta_1 = 0.9$ and $\beta_2 = 0.99$.

In the evaluation of the TFormer, the kernel size C for the pointwise convolution layer was set at 30 for the SAD dataset and 17 for the SEED-VIG dataset. All other parameters remain the same across both datasets. The kernel size K in the temporal convolution layer was set as 64. In both MHA and MCA, the number of the heads h was set as 4. The number of self-attention Transformer block L was set to 1. The expansion rate of FFN was set as 4. The GELU activation function was used in the Transformer blocks.

The proposed model was compared with six baselines for the cross-subject driver fatigue recognition: (1) SVM [17] with the extracted PSD features as the input; (2) EEGNet [29]; (3) ShallowCNN [28]; (4) Subject machine (SM) model [7]; (5) ICNN [30]; (6) Conformer [15]. For the parameter setting of the Conformer, the validation set was exploited to optimize the model parameters including the number of heads and the number of Transformer blocks. It is worth noting that ICNN was the state-of-the-art (SOTA) CNN model in cross-subject driver fatigue recognition, while the Conformer was the SOTA Transformer model in EEG decoding.

To ensure a fair comparison, subject-specific batch normalization (BN) was employed in all back-propagation-based models. In the case of the Conformer model, we replaced the LN layers with BN layers. For clarity, we denoted the original version as Conformer_O, while the modified version was labeled as Conformer $\boldsymbol{w}$ BN in the following comparisons.

For SVM, the regularization parameter was optimized by using grid search. The search space was $[2^{-5}, 2^5]$. The PSD features used for training SVM were computed via Fast Fourier Transform on each EEG epoch from these four spectral bands: delta (1–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), and beta (12–30 Hz). The final feature vector was a concatenation of the spectral powers extracted from the four bands and all available channels. In this study, the final feature vector was of 4 (frequency bands) × 30 (channels) = 120 dimensions.

Table 1
Comparison results (%) of LOSO driver fatigue recognition on SAD dataset. 'Avg. Acc.' represents average accuracy. 'Std.' represents standard deviation. The 'Avg. Rank' column does not have a unit.

Methods	Subjects											Avg. Acc.	Std.	Avg. Rank
	1	2	3	4	5	6	7	8	9	10	11			
SVM [17]	77.66	75.76	66.67	66.22	83.04	75.90	59.80	67.80	**88.54**	70.37	59.73	71.95	9.16	6.41
EEGNet [29]	77.66	61.97	**79.47**	79.46	**90.71**	84.46	62.35	75.15	84.39	76.30	76.90	77.17	8.69	3.41
ShallowCNN [28]	78.72	44.39	78.13	68.24	83.30	76.39	64.71	70.08	83.12	75.37	75.93	72.58	11.00	5.23
SM model [7]	78.72	68.18	79.33	68.24	85.27	83.73	64.71	57.20	78.03	**82.41**	71.68	74.32	8.94	4.5
Conformer_O [15]	75.43	61.06	73.60	68.78	85.80	83.61	60.20	69.70	77.64	73.89	74.34	73.10	8.01	6.09
Conformer ***w*** BN [15]	75.85	56.21	78.53	79.19	84.91	84.82	60.78	72.73	85.03	73.70	67.52	74.48	9.70	5.00
ICNN [30]	**81.60**	65.61	77.60	80.00	90.54	84.46	63.14	**77.88**	86.75	66.85	79.03	77.59	8.89	3.14
TFormer	80.21	**79.85**	76.40	**80.27**	89.29	**85.90**	**65.49**	76.14	84.39	78.89	**79.65**	**79.68**	6.16	**2.23**

Table 2
Comparison results (%) of LOSO driver fatigue recognition on SEED-VIG dataset.

Methods	Subjects												Avg. Acc.	Std.	Avg. Rank
	1	2	3	4	5	6	7	8	9	10	11	12			
SVM [17]	92.11	74.07	78.76	60.42	94.85	68.70	77.57	94.37	84.41	87.96	59.35	88.80	80.11	12.48	6.17
EEGNet [29]	92.11	89.14	87.21	86.46	63.81	95.56	80.15	91.69	73.56	89.49	87.27	88.16	85.38	8.88	5.04
ShallowCNN [28]	88.77	92.72	78.76	81.25	72.47	88.59	81.54	90.99	90.54	**92.70**	72.45	88.16	84.91	7.40	5.21
SM model [7]	92.98	91.98	87.39	85.42	92.78	95.37	71.69	**97.18**	91.09	92.34	**91.01**	89.81	89.92	6.55	3.5
Conformer_O [15]	86.14	85.80	75.00	79.43	92.89	85.85	79.49	82.39	79.11	84.67	67.34	82.33	81.70	6.44	7.00
Conformer ***w*** BN [15]	**95.96**	89.01	84.25	86.51	98.25	94.19	78.97	93.66	91.73	92.19	68.13	88.16	88.42	8.32	4.13
ICNN [30]	93.68	92.84	88.10	**87.60**	92.37	95.19	83.97	95.21	**92.82**	90.36	87.12	**90.19**	90.79	3.52	2.75
TFormer	95.88	**93.70**	**88.54**	86.88	**96.49**	**96.00**	**84.04**	95.35	91.58	92.19	85.11	88.54	**91.19**	4.46	**2.21**

4.4. Cross-subject driver fatigue recognition results

The accuracy of the proposed TFormer and the baselines on two datasets are compared in Tables 1 and 2. Additionally, the average ranks based on the accuracy of all subjects are also presented in tables. The comparison results on average accuracy are further highlighted in Fig. 2. The results could be analyzed from two aspects.

- **The comparison with CNN models**. Compared with CNN models, the superior performance of the proposed TFormer could be observed. With the same setting of subject-specific BN, the neglect of the frequency patterns and the global dependencies compared to the TFormer might lead to the suboptimal performance of the CNN models.
- **The comparison with Conformer**. As the SOTA Transformer model proposed for EEG decoding, Conformer_O was primarily designed to extract global temporal information for subject-dependent classification. Even though subject-specific batch normalization was also applied to Conformer ***w*** BN, the other main contributing factor, the learned global time-frequency patterns, could allow the proposed TFormer to achieve superior performance. It was observed that BN layers could also be beneficial in improving the performance of the Conformer in cross-subject EEG decoding.

Overall, the proposed TFormer outperformed the strong baseline methods, achieving the highest average accuracy and ranking first in the comparisons. The enhanced performance of TFormer is primarily due to the innovative TF-MCA component. This cross-attention layer is adept at comprehensively understanding global frequency-domain characteristics and integrating them with temporal features. By explicitly capturing a holistic view of time–frequency features, it surpasses the capabilities of other methods that are limited to analyzing patterns within a single time domain. This broader perspective contributes significantly to TFormer's superior performance. For an in-depth comparison of the time–frequency features against the single-domain features, please refer to the detailed analysis provided in Section 4.5. The exceptional performance of the proposed TFormer highlights its efficacy in cross-subject driver fatigue recognition. The incorporation of BN layers into the Conformer underscores the significance of employing BN in cross-subject EEG decoding tasks.

Table 3
Precision, Sensitivity, Specificity and F1-score (%) of the TFormer and the baseline methods on SAD dataset.

	Precision	Sensitivity	Specificity	F1-Score
SVM [17]	73.76	73.39	73.89	73.57
EEGNet [29]	78.23	79.53	77.86	78.86
ShallowCNN [28]	73.81	75.47	73.21	74.62
SM model [7]	70.66	**82.89**	65.58	76.28
Conformer_O [15]	78.18	67.46	81.17	72.40
Conformer ***w*** BN [15]	76.57	74.80	77.13	75.67
ICNN [30]	79.12	80.65	78.69	79.87
TFormer	**81.02**	79.98	**81.27**	**80.49**

Table 4
Precision, Sensitivity, Specificity and F1-score (%) of the TFormer and the baseline methods on SEED-VIG dataset.

	Precision	Sensitivity	Specificity	F1-Score
SVM [17]	74.00	85.46	69.97	79.32
EEGNet [29]	83.51	89.43	82.24	86.33
ShallowCNN [28]	84.60	84.72	84.59	84.66
SM model [7]	88.57	91.12	88.24	89.82
Conformer_O [15]	83.84	77.30	84.74	80.20
Conformer ***w*** BN [15]	87.09	89.60	86.71	88.32
ICNN [30]	89.65	91.61	89.42	90.61
TFormer	**90.15**	**91.91**	**89.95**	**91.02**

To better understand the classification capabilities of the proposed TFormer for driver fatigue recognition, we compared its performance with the baseline methods in terms of Precision, Sensitivity, Specificity, and F1-score on two datasets. The class 'alert' was set as positive, while the class 'fatigue' was set as negative in the calculation of these metrics. The comparison results are shown in Tables 3 and 4. Notably, the TFormer model outperformed the baseline methods in terms of Precision and F1-score, achieving the highest F1-scores of 80.49% and 91.02% on two datasets, respectively. Additionally, the proposed model exhibited higher Specificity than the baseline methods. This demonstrated that the model with the capabilities of capturing global patterns could ensure a higher probability of correctly identifying 'fatigue' subjects. This is especially valuable in practical applications. Finally, the TFormer also presented higher Sensitivity than most of the baseline methods.

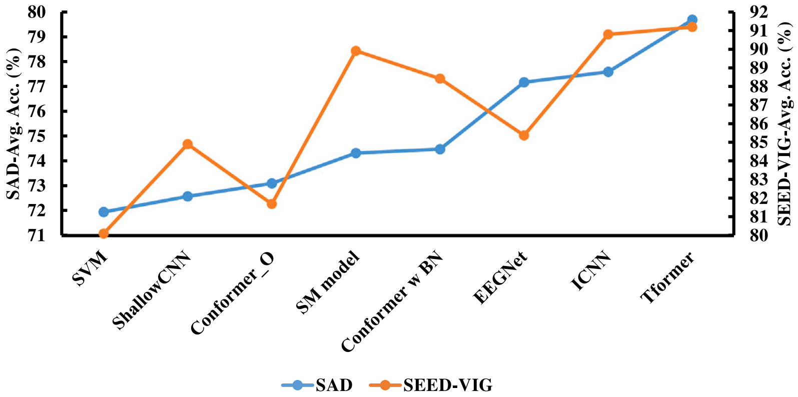


Fig. 2. Average accuracy comparison on SAD dataset and SEED-VIG dataset.

Table 5
p-values of the Wilcoxon results on SAD dataset.

	SVM	EEGNet	ShallowCNN	SM model	Conformer	Conformer w BN	ICNN
TFormer	<0.001	0.015	<0.001	<0.001	<0.001	<0.001	0.030

Table 6
p-values of the Wilcoxon results on SEED-VIG dataset.

	SVM	EEGNet	ShallowCNN	SM model	Conformer	Conformer w BN	ICNN
TFormer	<0.001	0.015	<0.001	0.009	<0.001	<0.001	0.260

To perform a statistical analysis of the comparison results in terms of average accuracy, we conducted a one-tailed Wilcoxon paired signed-rank test. The *p*-values obtained from the Wilcoxon test on both datasets are presented in Tables 5 and 6. The analysis revealed that the average accuracy of the proposed TFormer model was significantly better than that of the baseline methods, with a *p*-value less than 0.05. For the comparison on the SEED-VIG dataset, we concluded that TFormer was slightly better than ICNN.

Based on the above results, we could confidently conclude that the proposed TFormer performed exceptionally well in recognizing driver fatigue under the cross-subject setting, outperforming SOTA methods such as ICNN, SM model, and the Conformer. A higher rank was also observed for the proposed TFormer. Furthermore, comparing the four classification performance metrics demonstrated the superiority of the TFormer. Therefore, the proposed TFormer is a highly competitive classifier for EEG-based cross-subject driver fatigue recognition tasks.

4.5. Ablation study

To validate the proposed TF-MCA in TFormer, it was compared to the time-domain and frequency-domain variants; that is, the cross-attention Transformer block was replaced with the self-attention Transformer blocks. At the same time, only the time-domain or frequency-domain outputs from the convolutional stem were used as the inputs for both variants. Furthermore, the performance of the time-domain convolutional stem and the frequency-domain convolutional stem was also used for comparison, investigating whether the Transformer blocks were beneficial for the classification. The comparison results regarding average accuracy and F1-scores are shown in Fig. 3 It was observed that the proposed TFormer with TF-MCA showed better performance than the Transformer variants, which had two layers of self-attention blocks. This demonstrated the effectiveness of the proposed TF-MCA. The time-frequency patterns learned proved instrumental in achieving superior performance. Furthermore, compared with the basic convolutional stems solely used for classification, the better average accuracy achieved by Transformer models demonstrated the effectiveness of the Transformer blocks.

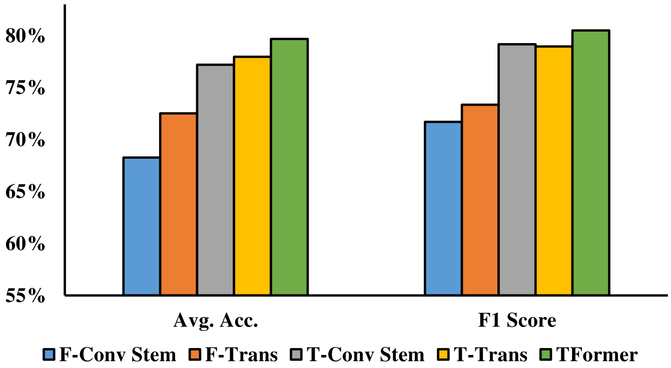


Fig. 3. Analysis of the proposed TF-MCA in the TFormer. The terms 'T-Conv Stem' and 'F-Conv Stem' represents time-domain Convolutional stem and frequency-domain stem, respectively. The terms 'T-Trans' and 'F-Trans' represents the time-domain Transformer and frequency-domain Transformer, respectively.

4.6. Investigation of the normalization layer in the TFormer

The selection of the normalization layer is discussed in this part. The comparison between the BN layer and the LN layer used in Transformer blocks was conducted. The comparison results were shown in Table 7. It was observed that TFormer with BN layers showed substantial improvement over that of the LN layers in terms of all comparison metrics. The possible reason is analyzed as follows. The variation in data statistics across different subjects poses a challenge to the generalization abilities of a well-trained model when applied to new subjects [7]. While LN is a common technique in Transformers to ensure stable training, it normalizes data on a per-token basis - -where a token corresponds to an individual time or frequency point in this context – without aggregating statistics across samples from subjects. In contrast, subject-specific BN offers a more tailored approach by normalizing data across the samples of both training and test subjects within mini-batches. This method is particularly effective in mitigating individual differences in data statistics, thereby enhancing model performance. Therefore, we concluded that the BN layers that could exploit the subject-specific

Table 7
Comparison results (%) of using different normalization layer.

	Avg. Acc.	Precision	Sensitivity	Specificity	F1-Score
TFormer ***w*** LN	75.25	77.65	76.80	77.82	77.20
TFormer ***w*** BN	79.68	81.02	79.98	81.27	80.49

Table 8
Comparison results (%) of using different position embedding methods. The term 'PE' stands for position embedding.

	Avg. Acc.	Precision	Sensitivity	Specificity	F1-Score
No PE	78.53	79.15	79.82	78.97	79.48
Learnable PE	79.01	80.73	78.83	81.17	79.75
Fixed PE	79.68	81.02	79.98	81.27	80.49

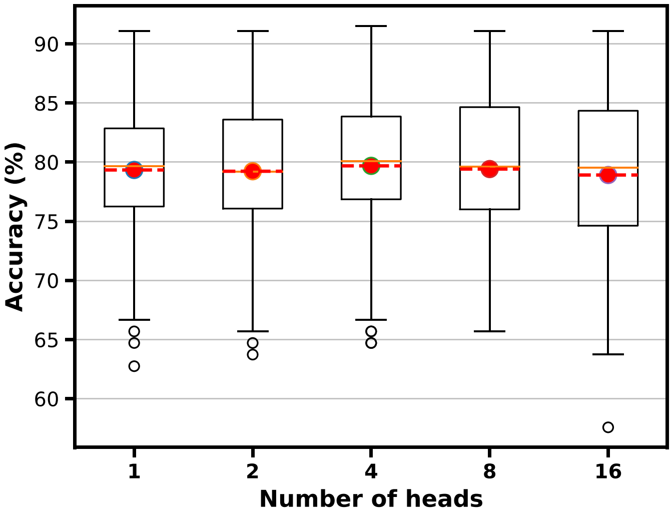


Fig. 4. The box and whisker plot of the accuracy attained by using different numbers of heads in the TFormer on the SAD dataset. The red dotted line and corresponding red dot signify the mean value, while the orange line indicates the median value.

feature statistics to reduce the impact of subject variability were more suitable for the Transformer model used for cross-subject EEG decoding.

4.7. Investigation of using different position embedding methods

Different position embedding methods were analyzed. Specifically, the performance of no position embeddings as well as using fixed position embeddings (*i.e.*, sine and cosine functions) and learnable 1D position embedding was compared as shown in Table 8. Results presented that position embeddings were essential for TFormer and used for cross-subject EEG decoding. Furthermore, slightly better performance was achieved by using fixed-position embeddings. Introducing more trainable parameters through learnable position embeddings can increase the risk of overfitting, which may negatively impact the model's overall performance. Our analysis indicates that, while both embedding methods perform similarly, using fixed-position embeddings could be a better choice, as it may help reduce the risk of overfitting and improve model performance.

4.8. Investigation of using different numbers of heads

To better understand the TFormer, Fig. 4 compares the performance of using various numbers of heads within the TFormer. As the output dimension of the convolutional stems was 16, the experiment assessed TFormer performance with 1, 2, 4, 8, and 16 heads. The results demonstrated similar performance across different head counts, with the 4-head TFormer exhibiting the optimal median and mean accuracy.

5. Discussion

Compared with the SOTA Transformer model used in EEG decoding [15,38], the proposed model focused more on the time-frequency patterns learned by the Transformer blocks. EEG data was commonly analyzed in the frequency domain because they were often found to be associated with behavioral patterns [49]. Hence, incorporating frequency information could enhance the classification performance of the model as demonstrated by the comparison results in Section 4.5. Although previous studies also exploited the time–frequency-domain images as the input of the Transformer to promote the time-frequency feature learning [50], the frequency points were considered as the embeddings of each time step such that the global dependencies along the frequency axis could be neglected. Furthermore, for the task of using multiple EEG channels, solely using time-frequency images may lead to neglect of the spatial information. In contrast, the TFormer exploited the spatial features as the embedding of each frequency point and each time step, and the global frequency information could also be contained in the final time-frequency patterns, which was beneficial in achieving better performance.

The convolutional stem is indispensable in the design of TFormer. The Transformer has a much less inductive bias than CNNs [13]. Therefore, learning beneficial patterns directly from raw EEG data may require a larger data size. However, because of the time-consuming data collection process, the limited EEG dataset may lead to the sub-optimal performance of Transformers directly input with raw EEG data. Therefore, the convolutional stem could be beneficial in reducing the size requirement of the training dataset. The benefits of the convolutional stem have also been demonstrated by a previous study [51]. Early convolutions could help Transformers converge faster and achieve better model accuracy. In addition, the convolutional stem helps to exploit the spatial information, which is also essential for EEG decoding.

Leveraging BN in Transformers has been investigated in different fields [20,21]. Although the standard BN was found not to be a suitable choice for the Transformer model and may lead to frequent crashes in model training, our experiment illustrated that the replacement did not have a negative impact on model learning. Compared with LN, BN has the advantage of considering the batch statistics of different subjects, which has been shown promising for cross-subject EEG decoding [10]. In the TFormer, BN layers also contributed to improving the model performance.

6. Conclusion

In this work, we proposed a Transformer-based model named TFormer that effectively integrates global information from both the time and frequency domains of raw EEG data. Specifically, a convolutional stem was designed for input embedding that operated in both domains. Following that, a TF-MCA was implemented to combine local frequency-domain patterns with global temporal patterns. The resulting time–frequency patterns were then fed into self-attention Transformer blocks to learn the global dependencies further. Moreover, considering the effectiveness of the subject-specific BN for cross-subject EEG decoding, the LN layers were replaced with BN layers in the TFormer. Experiment results demonstrated the effectiveness of the proposed TFormer. Furthermore, the selection of position embedding methods was investigated, which could serve as a reference for future studies. Overall, this work provides insights into developing an EEG-based driver fatigue detection system, further enhancing the human–computer interaction capabilities within driving systems.

6.1. Limitations and future works

While the proposed TFormer can effectively extract time-frequency features from EEG data, it only implicitly captures spatial dimensions through its convolutional stems. Given the significance of spatial information in EEG analysis, future work should explicitly integrate this aspect into our framework. Additionally, the Transformer's larger model size, compared to traditional CNN models, could constrain its generalization ability due to the limited size of our training dataset. To address this, data augmentation techniques to enhance the model's robustness and performance must be further explored. Moreover, this study relied solely on a well-established dataset for evaluation. The impact of using datasets with varying pre-processing and labeling methods on the proposed model and other machine-learning models warrants further exploration.

CRediT authorship contribution statement

Ruilin Li: Conceptualization, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft. **Minghui Hu:** Conceptualization, Formal analysis, Methodology, Writing – review & editing. **Ruobin Gao:** Conceptualization, Formal analysis, Methodology, Writing – review & editing. **Lipo Wang:** Conceptualization, Supervision, Writing – review & editing. **P.N. Suganthan:** Supervision, Writing – review & editing. **Olga Sourina:** Supervision, Writing – review & editing.

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

Data will be made available on request.

References

[1] V. Federico, H.J. S., G.H. G., M.P. McKay, Drowsy driving, Ann. Emerg. Med. 45 (2005) 433–434, http://dx.doi.org/10.1016/j.annemergmed.2005.01.015.

[2] H. Byeon, Exploring the predictors of rapid eye movement sleep behavior disorder for parkinson's disease patients using classifier ensemble, Healthc. 8 (2) (2020) 121, http://dx.doi.org/10.3390/healthcare8020121.

[3] F. Li, C.-H. Chen, G. Xu, L.-P. Khoo, Hierarchical eye-tracking data analytics for human fatigue detection at a traffic control center, IEEE Trans. Hum.-Mach. Syst. 50 (5) (2020) 465–474, http://dx.doi.org/10.1109/THMS.2020.3016088.

[4] F. Li, C.-H. Chen, G. Xu, L.P. Khoo, Y. Liu, Proactive mental fatigue detection of traffic control operators using bagged trees and gaze-bin analysis, Adv. Eng. Inform. 42 (2019) 100987, http://dx.doi.org/10.1016/j.aei.2019.100987.

[5] H.V. Koay, J.H. Chuah, C.-O. Chow, Y.-L. Chang, Detecting and recognizing driver distraction through various data modality using machine learning: A review, recent advances, simplified framework and open challenges (2014–2021), Eng. Appl. Artif. Intell. 115 (2022) 105309, http://dx.doi.org/10.1016/j.engappai.2022.105309.

[6] F. Liu, D. Chen, J. Zhou, F. Xu, A review of driver fatigue detection and its advances on the use of RGB-D camera and deep learning, Eng. Appl. Artif. Intell. 116 (2022) 105399, http://dx.doi.org/10.1016/j.engappai.2022.105399.

[7] R. Li, L. Wang, O. Sourina, Subject matching for cross-subject EEG-based recognition of driver states related to situation awareness, Methods 202 (2022) 136–143, http://dx.doi.org/10.1016/j.ymeth.2021.04.009.

[8] X. Yu, C.-H. Chen, H. Yang, Air traffic controllers' mental fatigue recognition: A multi-sensor information fusion-based deep learning approach, Adv. Eng. Inform. 57 (2023) 102123, http://dx.doi.org/10.1016/j.aei.2023.102123.

[9] Y. Liu, Z. Lan, J. Cui, O. Sourina, W. Müller-Wittig, Inter-subject transfer learning for EEG-based mental fatigue recognition, Adv. Eng. Inform. 46 (2020) 101157, http://dx.doi.org/10.1016/j.aei.2020.101157.

[10] R. Li, R. Gao, P.N. Suganthan, A decomposition-based hybrid ensemble CNN framework for driver fatigue recognition, Inform. Sci. 624 (2023) 833–848, http://dx.doi.org/10.1016/j.ins.2022.12.088.

[11] I. Mehmood, H. Li, Y. Qarout, W. Umer, S. Anwer, H. Wu, M. Hussain, M. Fordjour Antwi-Afari, Deep learning-based construction equipment operators' mental fatigue classification using wearable EEG sensor data, Adv. Eng. Inform. 56 (2023) 101978, http://dx.doi.org/10.1016/j.aei.2023.101978.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, Inc., 2017.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021.

[14] C. Li, Z. Zhang, X. Zhang, G. Huang, Y. Liu, X. Chen, EEG-based emotion recognition via transformer neural architecture search, IEEE Trans. Ind. Inform. 19 (4) (2023) 6016–6025, http://dx.doi.org/10.1109/TII.2022.3170422.

[15] Y. Song, Q. Zheng, B. Liu, X. Gao, EEG conformer: Convolutional transformer for EEG decoding and visualization, IEEE Trans. Neural Syst. Rehabil. Eng. 31 (2023) 710–719, http://dx.doi.org/10.1109/TNSRE.2022.3230250.

[16] A. Craik, Y. He, J.L. Contreras-Vidal, Deep learning for electroencephalogram (EEG) classification tasks: a review, J. Neural Eng. 16 (3) (2019) 031001, http://dx.doi.org/10.1088/1741-2552/ab0ab5.

[17] C.-S. Wei, Y.-T. Wang, C.-T. Lin, T.-P. Jung, Toward drowsiness detection using non-hair-bearing EEG-based brain-computer interfaces, IEEE Trans. Neural Syst. Rehab. Eng. 26 (2) (2018) 400–406.

[18] J. Ba, J. Kiros, G. Hinton, Layer normalization, in: NIPS 2016 Deep Learning Symposium Recommendation, 2016.

[19] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: F. Bach, D. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 37, PMLR, Lille, France, 2015, pp. 448–456.

[20] S. Shen, Z. Yao, A. Gholami, M. Mahoney, K. Keutzer, PowerNorm: Rethinking batch normalization in transformers, in: H.D. III, A. Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 119, PMLR, 2020, pp. 8741–8751.

[21] Z. Yao, Y. Cao, Y. Lin, Z. Liu, Z. Zhang, H. Hu, Leveraging batch normalization for vision transformers, in: 2021 IEEE/CVF International Conference on Computer Vision Workshops, ICCVW, 2021, pp. 413–422, http://dx.doi.org/10.1109/ICCVW54120.2021.00050.

[22] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 10684–10695.

[23] K.K. Ang, Z.Y. Chin, H. Zhang, C. Guan, Filter bank common spatial pattern (FBCSP) in brain-computer interface, in: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008, pp. 2390–2397, http://dx.doi.org/10.1109/IJCNN.2008.4634130.

[24] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, A. Cichocki, EmotionMeter: A multimodal framework for recognizing human emotions, IEEE Trans. Cybern. 49 (3) (2019) 1110–1122, http://dx.doi.org/10.1109/TCYB.2018.2797176.

[25] L.D. Sharma, V.K. Bohat, M. Habib, A.M. Al-Zoubi, H. Faris, I. Aljarah, Evolutionary inspired approach for mental stress detection using EEG signal, Expert Syst. Appl. 197 (2022) 116634, http://dx.doi.org/10.1016/j.eswa.2022.116634.

[26] C. Ye, Z. Yin, M. Zhao, Y. Tian, Z. Sun, Identification of mental fatigue levels in a language understanding task based on multi-domain EEG features and an ensemble convolutional neural network, Biomed. Signal Process. Control 72 (2022) 103360, http://dx.doi.org/10.1016/j.bspc.2021.103360.

[27] Z. Gao, Y. Li, Y. Yang, N. Dong, X. Yang, C. Grebogi, A coincidence-filtering-based approach for CNNs in EEG-based recognition, IEEE Trans. Ind. Inform. 16 (11) (2020) 7159–7167, http://dx.doi.org/10.1109/TII.2019.2955447.

[28] R.T. Schirrmeister, J.T. Springenberg, L.D.J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, T. Ball, Deep learning with convolutional neural networks for EEG decoding and visualization, Hum. Brain Mapp. 38 (11) (2017) 5391–5420, http://dx.doi.org/10.1002/hbm.23730.

[29] V.J. Lawhern, A.J. Solon, N.R. Waytowich, S.M. Gordon, C.P. Hung, B.J. Lance, EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces, J. Neural Eng. 15 (5) (2018) 056013, http://dx.doi.org/10.1088/1741-2552/aace8c.

[30] J. Cui, Z. Lan, O. Sourina, W. Müller-Wittig, EEG-based cross-subject driver drowsiness recognition with an interpretable convolutional neural network, IEEE Trans. Neural Netw. Learn. Syst. 1–13 (2022) http://dx.doi.org/10.1109/TNNLS.2022.3147208.

[31] R. Li, P.N. Suganthan, O. Sourina, Sample-based data augmentation based on electroencephalogram intrinsic characteristics, IEEE J. Biomed. Health. Inform. 26 (10) (2022) 4996–5003, http://dx.doi.org/10.1109/JBHI.2022.3185587.

[32] J. Shi, K. Wang, Fatigue driving detection method based on time-space-frequency features of multimodal signals, Biomed. Signal Process. Control 84 (2023) 104744, http://dx.doi.org/10.1016/j.bspc.2023.104744.

[33] D. Gao, K. Wang, M. Wang, J. Zhou, Y. Zhang, SFT-Net: A network for detecting fatigue from EEG signals by combining 4D feature flow and attention mechanism, IEEE J. Biomed. Health Inf. (2023) 1–12, http://dx.doi.org/10.1109/JBHI.2023.3285268.

[34] H. Wang, T. Fu, Y. Du, et al., Scientific discovery in the age of artificial intelligence, Nature 620 (7972) (2023) 47–60, http://dx.doi.org/10.1038/s41586-023-06221-2.

[35] L. Gong, M. Li, T. Zhang, W. Chen, EEG emotion recognition using attention-based convolutional transformer neural network, Biomed. Signal Process. Control 84 (2023) 104835, http://dx.doi.org/10.1016/j.bspc.2023.104835.

[36] M. Zeynali, H. Seyedarabi, R. Afrouzian, Classification of EEG signals using transformer based deep learning and ensemble models, Biomed. Signal Process. Control 86 (2023) 105130, http://dx.doi.org/10.1016/j.bspc.2023.105130.

[37] G. Siddhad, A. Gupta, D.P. Dogra, P.P. Roy, Efficacy of transformer networks for classification of EEG data, Biomed. Signal Process. Control 87 (2024) 105488, http://dx.doi.org/10.1016/j.bspc.2023.105488.

[38] J. Xie, J. Zhang, J. Sun, Z. Ma, L. Qin, G. Li, H. Zhou, Y. Zhan, A transformer-based approach combining deep learning network and spatial-temporal information for raw EEG classification, IEEE Trans. Neural Syst. Rehabil. Eng. 30 (2022) 2126–2136, http://dx.doi.org/10.1109/TNSRE.2022.3194600.

[39] A. Miltiadous, E. Gionanidis, K.D. Tzimourta, N. Giannakeas, A.T. Tzallas, DICE-Net: A novel convolution-transformer architecture for alzheimer detection in EEG signals, IEEE Access 11 (2023) 71840–71858, http://dx.doi.org/10.1109/ACCESS.2023.3294618.

[40] Z. Deng, C. Li, R. Song, X. Liu, R. Qian, X. Chen, EEG-based seizure prediction via hybrid vision transformer and data uncertainty learning, Eng. Appl. Artif. Intell. 123 (2023) 106401, http://dx.doi.org/10.1016/j.engappai.2023.106401.

[41] Y. Song, Q. Zheng, Q. Wang, X. Gao, P.-A. Heng, Global adaptive transformer for cross-subject enhanced EEG classification, IEEE Trans. Neural Syst. Rehabil. Eng. 31 (2023) 2767–2777, http://dx.doi.org/10.1109/TNSRE.2023.3285309.

[42] Y. Ding, N. Robinson, S. Zhang, Q. Zeng, C. Guan, TSception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition, IEEE Trans. Affect. Comput. (2022) 1, http://dx.doi.org/10.1109/TAFFC.2022.3169001.

[43] J. Gehring, M. Auli, D. Grangier, D. Yarats, Y.N. Dauphin, Convolutional sequence to sequence learning, in: D. Precup, Y.W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 70, PMLR, 2017, pp. 1243–1252.

[44] Q. Wang, B. Li, T. Xiao, J. Zhu, C. Li, D.F. Wong, L.S. Chao, Learning deep transformer models for machine translation, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1810–1822, http://dx.doi.org/10.18653/v1/P19-1176.

[45] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, T. Liu, On layer normalization in the transformer architecture, in: H.D. III, A. Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 119, PMLR, 2020, pp. 10524–10533.

[46] Z. Cao, C.-H. Chuang, J.-K. King, C.-T. Lin, Multi-channel EEG recordings during a sustained-attention driving task, Sci. Data. 6 (1) (2019) 19, http://dx.doi.org/10.1038/s41597-019-0027-4.

[47] W.-L. Zheng, B.-L. Lu, A multimodal approach to estimating vigilance using EEG and forehead EOG, J. Neural Eng. 14 (2) (2017) 026017, http://dx.doi.org/10.1088/1741-2552/aa5a98.

[48] Federal Highway Administration, PERCLOS: A valid psychophysiological measure of alertness as assessed by psychomotor vigilance, 1998.

[49] E.R. Kandel, J.H. Schwartz, T.M. Jessell, S. Siegelbaum, A.J. Hudspeth, S. Mack, et al., Principles of Neural Science, vol. 4, McGraw-hill New York, 2000.

[50] H. Phan, K. Mikkelsen, O.Y. Chén, P. Koch, A. Mertins, M. De Vos, Sleep-Transformer: Automatic sleep staging with interpretability and uncertainty quantification, IEEE Trans. Biomed. Eng. 69 (8) (2022) 2456–2467, http://dx.doi.org/10.1109/TBME.2022.3147187.

[51] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollar, R. Girshick, Early convolutions help transformers see better, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J.W. Vaughan (Eds.), Advances in Neural Information Processing Systems, vol. 34, Curran Associates, Inc., 2021, pp. 30392–30400.