

Fatigue Driving Detection Based On Multimodal EEG Feature Fusion

Qiong Hong

0622830@jsei.edu.cn

Jiangsu Vocational College of Electronics and Information

Zhonghua Xu

Huaiyin Institute of Technology

Lianshan Pan

Huaiyin Institute of Technology

Peng Wang

Huaiyin Institute of Technology

Hao Zhang

Huaiyin Institute of Technology

Jie He

Southeast University

Research Article

Keywords: Deep Learning, Fatigue Driving Detection, EEG Signals, Multi-Modal Feature Fusion

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8385637/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Fatigue Driving Detection Based On Multimodal EEG Feature Fusion

Qiong Hong^{1,*}, Zhonghua Xu², Lianshan Pan², Peng Wang², Hao Zhang², Jie He³

¹ Business School, Jiangsu Vocational College of Electronics and Information, Huaian 223003, China.

² School of Transportation Engineering, Huaiyin Institute of Technology, Huaian 223003, China.

³ School of Transportation, Southeast University, Nanjing 211189, China.

Corresponding authors: Qiong Hong; E-mail: 062283@jsei.edu.cn.

Abstract:

Preventing and reducing traffic accidents is an urgent necessity for safeguarding public safety and societal development. Fatigued driving, as one of the core contributing factors, necessitates effective monitoring. Therefore, addressing issues such as weak model generalisation capability, complex parameter tuning, and difficulties in extracting EEG signal features within existing fatigued driving detection technologies, this paper proposes a multimodal EEG feature fusion-based driving fatigue detection model (MEFD-Net). This model comprises three sub-networks: (1) A time-domain feature extraction network utilises an enhanced BiLSTM with GeLU activation to capture temporal dynamics; (2) A time-invariant feature extraction network employs weighted adaptive multi-scale convolutions combined with residual connections and self-attention mechanisms; (3) The spatial-frequency domain feature extraction network employs a pure convolutional architecture to extract spatial-frequency domain features. By integrating multi-dimensional EEG features, MEFD-Net achieves end-to-end classification without manual feature selection, enhancing both classification accuracy and fault tolerance. Experiments demonstrate that the proposed model exhibits superior performance in processing complex, unstable EEG signals, achieving average improvements of 7.09%, 5.56%, and 4.38% in detection accuracy, recall, and F1 score respectively, outperforming existing mainstream algorithms.

Keywords: Deep Learning; Fatigue Driving Detection; EEG Signals; Multi-Modal Feature Fusion

1 Introduction

The rapid expansion of global vehicle ownership has prioritized road traffic safety as a critical public concern. Driver fatigue is a primary cause of severe accidents and necessitates the development of efficient detection technologies. While Electroencephalography (EEG) is a reliable indicator of neural activity, its application is hindered by inherent non-stationarity and multi-domain coupling. Current methodologies often rely on manual feature engineering or single-modality deep learning and fail to capture the multidimensional complexity of EEG signals, resulting in limited robustness in dynamic environments.

To address these challenges, this study proposes a multimodal EEG feature fusion network. The primary contributions are as follows: (1) A WGAN-GP framework addresses data scarcity and class imbalance by generating balanced training samples to mitigate overfitting; (2) An enhanced BiLSTM network incorporating a Hybrid Attention Mechanism (HAM) and GeLU activation captures temporal dynamics while circumventing gradient degradation; (3) A Weighted Adaptive Multi-scale Convolutional Network (WACNet) utilizes a Saliency Encoder and self-attention to optimize feature weight distribution across varying scales; (4) A Pure Convolutional Neural Network (PCNN) extracts spatial-frequency features via multi-layer blocks and downsampling modules to ensure training stability; and (5) By integrating temporal, time-invariant, and spatial-frequency features, the proposed MEFD-Net demonstrates superior detection accuracy on the SEED-VIG dataset and significantly outperforms mainstream baselines.

The paper is organized as follows: Section 2 reviews related work; Section 3 details the proposed architecture and fusion strategy; Section 4 presents experimental validation; and Section 5.

2. Literature Review

Driver fatigue, characterized by diminished psychophysiological function due to prolonged operation, severely compromises traffic safety. Detection strategies are primarily classified into subjective and objective methods. Subjective assessments, such as the KSS and SSS [1], are cost-effective but limited by inherent bias and lower reliability. Conversely, objective approaches, utilizing physiological (e.g., EEG, ECG) and behavioral metrics (e.g., blink rate, trajectory), provide a more accurate and robust quantification of the driver's state [2].

To obtain more objective and reliable assessment results, numerous researchers have turned to objective detection methods.

Firstly, research on non-contact fatigue detection primarily focuses on the analysis of facial visual features and vehicle operational behaviors. Methodologies range from monitoring specific maneuvers like acceleration and braking [3] and standardized driving patterns [5], to extracting facial indicators via skin tone and orbital analysis [4]. Algorithmic approaches include applying k-nearest neighbors to real-world data [6], fusing

image and vehicle features via clustering [7], and deploying lightweight frameworks for multi-task detection [8]. Additionally, enhanced deep learning architectures, such as YOLOv3-tiny, have been utilized for ocular and mouth feature analysis [9]. Collectively, existing research predominantly employs traditional machine learning or modified deep learning algorithms to process single or limited multimodal features.

Secondly, vehicle-based fatigue detection indirectly assesses driver state via lateral and longitudinal control metrics. Methodologies include microcontroller-based sensor systems for parameter monitoring [10], DMS/GPS integration for duration analysis [11], and non-linear SVMs for lane departure prediction [12]. Additionally, entropy features derived from steering angles have been employed for real-time binary state discrimination [13]. Collectively, these approaches leverage vehicle telemetry integrated with traditional machine learning or embedded systems to identify anomalous driving dynamics.

Furthermore, acoustic-based detection predicates on the hypothesis that fatigue-induced neuromuscular impairments manifest as discernible alterations in speech patterns. Methodologies have progressed from utilizing NLD and spectral features for ensemble classification [14] and remote estimation [15], to optimizing openSMILE feature sets for specific domains like air traffic control [16]. Advanced computational strategies include employing deep learning for latent feature extraction [17] and utilizing adaptive quantum genetic algorithms to optimize heterogeneous ensemble learners [20]. Notably, multimodal fusion has emerged as a key trend, integrating vocal signals with ocular metrics (e.g., PERCLOS) via Fuzzy SVMs [18] and Bi-LSTM networks [19] to maximize detection robustness. Overall, existing literature primarily focuses on preprocessing signals for MFCC and NLD extraction, subsequently processing them through machine learning or deep learning frameworks to infer fatigue states.

Finally, physiological monitoring is regarded as the "gold standard" for fatigue detection, directly quantifying CNS and ANS activity via intrinsic fluctuations in biological markers. Methodologies have diversified across signal modalities: Cardiovascular approaches utilise hierarchical transformers for ECG/HRV analysis [21] and integrate ECG with video [22] or EEG [23] for multimodal assessment. Ocular-based (EOG) studies focus on extracting blink dynamics [24,25] and monitoring long-term alertness via wearable sensors [26]. Muscular (EMG) research has advanced through portable acquisition systems [27],

non-contact fusion with ECG using complexity analysis [28], and optimized support vector classification [29]. Furthermore, Neural (EEG) analysis remains pivotal, employing dry electrodes with power spectral density and entropy metrics [30], flexible wavelet transforms [31], and specific frequency band evaluations [32] to robustly distinguish fatigue states.

Despite these advancements, significant limitations remain: (1) The predominant reliance on manual feature extraction or single modalities fails to capture the multidimensional complexity of EEG signals, limiting generalization in dynamic driving scenarios; (2) Existing fusion strategies often lack deep integration across time-domain, time-invariant, and spatial-frequency features, resulting in suboptimal information utilization and reduced robustness; (3) Current algorithms inadequately address the inherent non-stationarity and inter-subject variability of EEG data, leading to classification biases; and (4) The scarcity of fatigue samples and resulting class imbalance are rarely mitigated by systematic data augmentation, rendering models susceptible to overfitting.

To address these challenges, this study proposes a multimodal EEG feature fusion framework. By integrating an enhanced bidirectional LSTM, weighted adaptive multi-scale convolutions, and pure convolutional spatial-frequency feature extraction, the model achieves precise, end-to-end fatigue detection. This approach effectively enhances feature extraction and generalization for non-linear, non-stationary signals, establishing a robust theoretical foundation for real-time, adaptive driving safety systems.

3 Experiment

3.1 Experimental data sets

3.1.1 Dataset

The experimental data utilised in this study were obtained from the SEED-VIG dataset [33], a widely validated benchmark for fatigue detection research. This dataset comprises EEG and electrooculogram (EOG) recordings from 23 subjects in a simulated driving environment. Data were acquired across 17 channels at a sampling rate of 200 Hz over a duration of 118 minutes per subject, yielding a cumulative 7,080 minutes of recording. Fatigue states are annotated using the PERCLOS index, a robust objective metric updated every 8 seconds, resulting in 885 labels per subject.

To validate model performance, ground truth calibration is essential. This study adopts the PERCLOS method proposed by Gao X Y [34], which quantifies fatigue based on the percentage of eye closure over a specific interval. The formula is defined as:

$$PERCLOS = \frac{blink + CLOS}{interval} \quad (1)$$

$interval = blink + CLOS + fixation + saccade(2)$

Where: $blink$ is blink duration; $CLOS$ is Closing eyes time; $interval$ is total time; $fixation$ is fixation state of the eye; $saccade$ is saccade state of the eye.

3.1.2 Data preprocessing

In this paper, band-pass filters with different frequency ranges are used to filter EEG signals to obtain δ -wave, θ -wave, α -wave, β -wave and γ -wave EEG signals. Secondly, in order to improve the stability of model training and the balance of samples, this paper uses WGAN-GP for data enhancement. WGAN-GP enforces Lipschitz constraints by introducing gradient penalty, thereby reducing instability during training. Different from the adversarial training of generator G and discriminator D in traditional GAN, WGAN-GP adds an additional penalty term to the loss function to optimize the training process. The specific form of the loss function is :

$$\min_{\theta_G} \max_{\theta_D} L(X_r, X_g) = \mathbb{E}_{x_r \sim X_r} [\log(D(x_r))] \quad (3)$$

$$- \mathbb{E}_{x_g \sim X_g} [\log(D(x_g))] + \lambda \mathbb{E}_{\hat{x} \sim \hat{X}} [(||\nabla_{\hat{x}} D(\hat{x})||_2 - 1)^2] \quad (4)$$

Where: X_r is real data distribution; X_g is generated data distribution respectively; θ_G is trainable parameters of the generator; θ_D is trainable parameters of discriminator; λ is a hyperparameter that controls the trade-off between the original target and the gradient penalty ; $\hat{x}$ is interpolation between real data and generated data; $\nabla_{\hat{x}} D(\hat{x})$ is the gradient of the discriminator output interpolation.

In addition, this paper also calculates the differential entropy feature, which is a feature that represents the logarithm of the energy spectrum of a specific frequency band. It is assumed that the EEG signal corresponding to the channel of each frequency band conforms to the Gaussian distribution. The calculation formula of differential entropy is :

$$DE = - \int_b^a p(x) \log(p(x)) dx \quad (5)$$

$$DE = - \int_b^a \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right) dx \quad (6)$$

Where, x is EEG signal characteristics; $p(x)$ is probability density function, respectively; $[a, b]$ is range of values.

Substituting the probability density function into DE, the formula is as follows :

$$DE = \frac{1}{2} \log(2\pi e \sigma_i^2) \quad (7)$$

3.2 MEFD-Net model design

3.2.1 Network architecture

As illustrated in **Fig. 1**, the proposed MEFD-Net orchestrates three specialized sub-networks to synthesize multidimensional EEG features. First, an enhanced BiLSTM incorporating GeLU activation

captures temporal signal dynamics. Second, a time-invariant network employs Weighted Adaptive Multi-scale Convolution (WAMC) coupled with residual connections and self-attention mechanisms to extract robust multi-scale representations. Finally, a pure convolutional architecture isolates spatial-frequency characteristics. This synergistic fusion enables precise driver fatigue assessment and enhances operational safety.

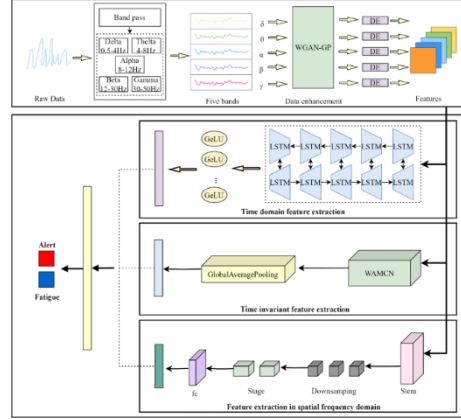


Fig.1. Structure diagram of multimodal EEG feature fusion network

3.2.2 Time domain feature extraction

Time-domain features refer to the changing characteristics of signals on the time axis. For EEG signals, these features can reveal the temporal dynamics of brain activity and are crucial for understanding and analyzing brain states. When extracting these features, this paper proposes an improved time-domain feature extraction network HG-BiLSTM of bidirectional long short-term memory network, which strengthens the feature extraction ability of the network through the mixed attention mechanism, and optimizes the feature processing by combining the GeLU activation function.

The traditional recurrent neural network (RNN) has limitations in processing time series data, and GRU and LSTM [35], as their improved versions, can capture the characteristics of time series more effectively. It solves the problem of gradient disappearance when the recurrent neural network processes long sequence data. The structure of the LSTM network is shown in **Fig. 2**.

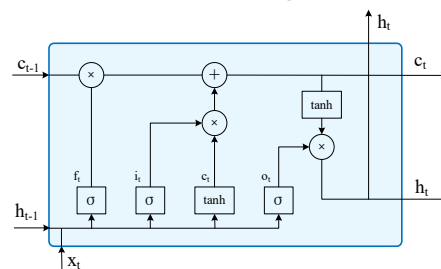


Fig. 2 LSTM network structure diagram

However, although LSTM can remember long-term information, it cannot fully understand the

context of the entire data sequence. Compared with traditional LSTM, BiLSTM processes the input sequence through two independent LSTM layers to capture features from two directions. The forward LSTM processes the input from left to right, while the reverse LSTM processes the input from right to left, so that the network can comprehensively consider the information of the sequence. Specifically, the forward-propagating LSTM unit processes the input sequence in chronological order. At each time step t , the forward LSTM calculates the hidden state h_t and the unit state c_t through the following formula :

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \quad (8)$$

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f) \quad (9)$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \quad (10)$$

$$\tilde{c}_t = \tanh(W_{cx}x_t + W_{ch}h_{t-1} + b_c) \quad (11)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (12)$$

$$h_t = o_t \odot \tanh(c_t) \quad (13)$$

Where: i_t is input gate; f_t is forgetting gate; o_t is output gate; $\tilde{c}_t$ is candidate state of current information; σ is Sigmoid function; $\tanh$ is hyperbolic tangent functions.

The reverse LSTM uses the opposite time sequence to analyze the input sequence, and its hidden state h'_t and unit state c'_t are calculated by the gating mechanism of the same structure. Finally, the hidden states of the two directions of the network connection form a feature representation $y_t = [h_t; h'_t]$, which combines the forward and reverse information. The structure of BiLSTM is shown in Fig. 3.

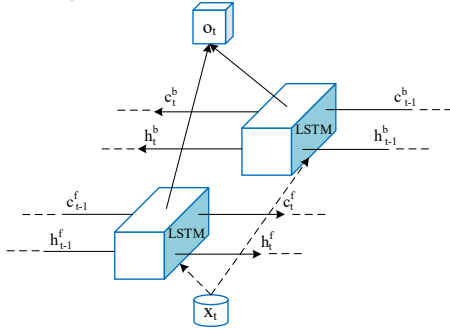


Fig. 3 BiLSTM structure diagram

In order to further strengthen the feature extraction ability of the network, this paper introduces a hybrid attention mechanism to capture the interaction between features. The purpose is to pay attention to the time dimension and feature dimension of the data at the same time, and improve the ability of the model to capture multivariate time series information. The structure of HAM is shown in Fig. 4.

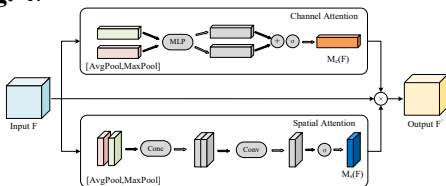


Fig. 4 HAM structure diagram

The Hybrid Attention Mechanism (HAM) integrates channel and spatial attention modules to optimize multi-dimensional time-series analysis. The channel attention module derives feature weights via average and maximum pooling, processed through a multi-layer perceptron and sigmoid activation. Concurrently, the spatial attention module emphasizes temporal significance utilizing pooling and convolutional layers. This parallel architecture minimizes inter-dimensional interference and enhances feature representation. The mathematical formulation is defined as follows:

$$Mc(F) = \sigma(MLP(AvePool(F)) + MLP(MaxPool(F))) \quad (14)$$

$$Ms(F) = \sigma(Conv(Conn[AvePool(F), MaxPool(F)])) \quad (15)$$

$$F' = Mc(F) \otimes F \otimes Ms(F) \quad (16)$$

Where: σ is sigmoid activation function; MLP is multi-layer perceptron; $Conv$ is convolution operation; $Conn$ is connection operation; Mc is output of channel attention; Ms is output of spatial attention; $AvePool$ is average pooling; $MaxPool$ is maximum pooling.

When using BiLSTM, the gradient sometimes disappears, which significantly affects the detection effect. The $GeLU$ activation function is used to calculate and process the output features, which can effectively avoid the disappearance of the gradient. Therefore, the output result is finally realized by the $GeLU$ activation function G . The formula is as follows:

$$GeLU(x) = xP(X \leq x) \quad (17)$$

$$GeLU(x) = 0.5x \left(1 + \tanh \left[\sqrt{\frac{2}{\pi}} (x + 0.044715x^3) \right] \right) \quad (18)$$

$$out = G(P_i + P'_i) \quad (19)$$

Where: x is input value; X is Gaussian random variable with zero mean and unit variance; $P(X \leq x)$ is the probability that X is less than or equal to the given value x ; P_i is output sequence of the forward propagation layer; P'_i is output sequence of the back propagation layer.

3.2.3 Time invariant feature extraction

This study introduces the Weighted Adaptive Multi-scale Convolutional Network (WACNet) to extract time-invariant EEG characteristics. As illustrated in Fig. 5, the architecture integrates a Weighted Adaptive Multi-scale Convolution Module (WAMCM), three parallel enhanced Fully Convolutional Networks (FCNs), and a self-attention mechanism. The WAMCM leverages diverse kernel sizes to capture multi-scale features, while the parallel FCN branches utilise distinct receptive fields and Squeeze-and-Excitation (SE) modules for feature refinement. Structurally, each processing block comprises sequential convolution, batch normalization, and Rectified Linear Unit (ReLU) layers to optimize feature extraction.

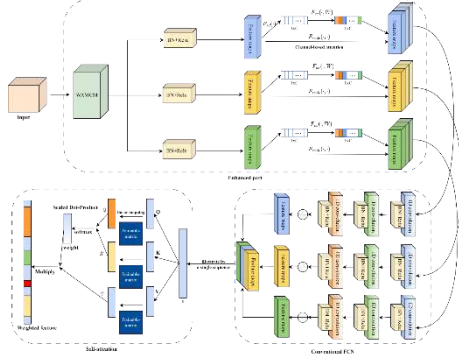


Fig. 5 Structure diagram of weighted adaptive multi-scale convolutional network

The WAMCM module captures multi-scale features by applying convolution kernels of different sizes. The size of these convolution kernels gradually increases from 2 to 16, and is divided into three groups : small, medium and large, each group contains five convolution kernels of different sizes. The feature maps extracted by these convolution kernels are given different weights, and the final output is obtained by weighted summation, as shown in Fig. 6.

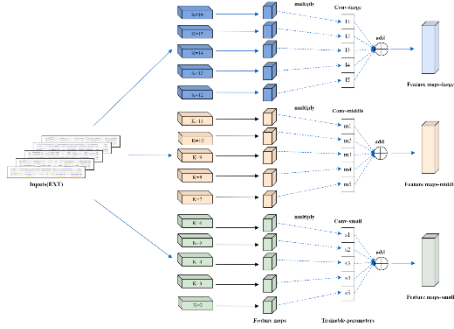


Fig. 6 WAMCM module structure diagram

Let $F = [f_1, f_2, \dots, f_c] \in R^{e \times t \times c}$ denote the feature graph, where c denotes the number of feature channels, and $e \times t$ denotes the spatial dimension. The weighted adaptive multi-scale convolution calculation formula is as follows :

$$F_0 = \sum_{i=1}^5 w_i \cdot F_i \quad (20)$$

$$\sum_{i=1}^5 w_i = 1 \quad (21)$$

Where: F_i is the feature map obtained by each individual convolution in each convolution group; F_0 is the output feature map of each convolution layer; $w = [w_1, w_2, \dots, w_5] \in R^5$ is trainable weight vector.

Each enhanced FCN integrates a traditional architecture with an enhancement module, employing three distinct convolutional kernel sizes across parallel pathways to extract features from varying receptive fields. Additionally, a Squeeze-and-Excitation (SE) module is incorporated to dynamically recalibrate channel weights by explicitly modeling channel interdependencies. The

SE mechanism consists of three primary operations.

Firstly, the compression operation F_{sq} uses global average pooling to compress global spatial information into channel statistics. Suppose that $U = [u_1, u_2, \dots, u_c] \in R^{H \times W \times C}$ denotes the input feature map, where C denotes the number of feature channels and $H \times W$ denotes the spatial dimension. The calculation formula of compression operation is as follows :

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W u_c(i, j) \quad (22)$$

Where: $z = [z_1, z_2, \dots, z_c] \in R^C$ is one-dimensional channel descriptor.

Second, the excitation operation F_{ex} is a gating mechanism with S-shaped activation to capture the channel dependence from z . The formula is as follows :

$$s = F_{ex}(z, W) = \sigma(W_2 \delta(W_1 z)) \quad (23)$$

Where: σ is *sigmoid* function; δ is *ReLU* function, respectively; $W_1 \in R_r^{C \times C}$ is the weight matrix of the dimension reduction layer; $W_2 \in R_r^{C \times C}$ is the weight matrix of the dimension increase layer; R is dimension reduction ratio.

Thirdly, the rescaling operation F_{scale} scale uses the channel multiplication between the feature map u_c and the scale s_c to restore the weighted channel descriptor to the original dimension :

$$\tilde{U}_c = F_{scale}(u_c, s_c) = s_c \cdot u_c \quad (24)$$

Where: $\tilde{U}_c = [\tilde{u}_1, \tilde{u}_2, \dots, \tilde{u}_c] \in R^{H \times W \times C}$ is channel-wise reweighted feature maps.

The convolution operation in the convolution block uses three one-dimensional convolutions with kernel sizes of 8、5 and 3, respectively. In addition, the number of filters in the three cores is 128、256 and 128, respectively. The residual connection is applied between the output of the SE block and the output of the traditional FCN to improve the performance of the model. The calculation formula of convolution block is as follows :

$$y = W \otimes x + b \quad (25)$$

$$s = BN(y) \quad (26)$$

$$h = ReLU(s) \quad (27)$$

Where: W is weight matrix; b is deviation; BN = batch normalization operation; $ReLU$ is corrected linear unit activation function.

Finally, the self-attention module is used to re-weight the output of the final feature. The calculation formula is as follows :

$$Attention(\tilde{Q}, \tilde{K}, \tilde{V}) = softmax\left(\frac{\tilde{Q}\tilde{K}^T}{\sqrt{d_k}}\right)\tilde{V} \quad (28)$$

$$\tilde{Q} = QW_Q, \tilde{K} = KW_K, \tilde{V} = VW_V \quad (29)$$

Where: $Q = K = V \in R^C$ is a subset of $I \in R^C$; I is output vector of the global average pooling layer; W_Q is trainable linear mapping matrix; W_K is trainable linear mapping matrix; $W_V \in R^{C \times C}$ is

trainable linear mapping matrix; $\tilde{Q}$ is mapping vector; $\tilde{K}$ is mapping vector; $\tilde{V} \in R^C$ represent the mapping vector; d_k is the dimension of I .

3.2.4 Spatial frequency domain feature extraction

This paper proposes a Pure Convolutional Neural Network (PCNN) to extract spatial-frequency features from EEG signals. The architecture comprises input, feature extraction, and output modules. The input module expands 5-channel features to 96 channels via a 4×4 convolutional block, utilizing batch processing and LayerNorm for stabilization. The feature extraction module incorporates sequential down-sampling and processing stages to capture key features while mitigating overfitting. Specifically, down-sampling employs initial 4×4 kernels followed by 2×2 spatial reductions, integrated with BatchNorm and LayerNorm to ensure training stability.

The input feature $X = (X_1, X_2, \dots, X_N)$ is defined, and the size of each input feature is (C, H, W) , where N represents the batch size, C represents the channel size of the input, H represents the height of the feature, and W represents the width of the feature. The convolution layer calculation formula is as follows :

$$f_i = G(W_c * X) \quad (30)$$

Where: G = activation function ; W_c = weight value.

The downsampling module consists of n ($1 \leq n \leq 3$) convolution blocks, as shown in **Fig. 7**. The size of the first convolution kernel is 4×4 , and the size of the subsequent convolution kernel is 2×2 , which is used for spatial downsampling. The establishment of the down-sampling layer in the feature extraction process can avoid model over-fitting and reduce the data size as much as possible to extract useful feature information. In addition, adding BatchNorm and LayerNorm can stabilize the model training time.

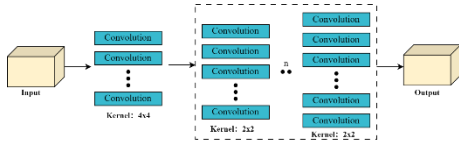


Fig. 7. Downsampling module

The Stage module is stacked by multiple network layers, and each network layer consists of multiple block modules. These block modules are connected by residuals to keep the number of input and output channels consistent. For example, the first network layer contains n Block modules, and the input and output channels of the convolutional layer of each module are 192 ; the second network layer has $n \times n$ Block modules, and the number of input and output channels of the convolutional layer is 384. The third network layer contains $n \times n \times n$

Block modules, and the number of input and output channels of the convolution kernel is 96 / 192 / 384. As shown in **Fig. 8**.

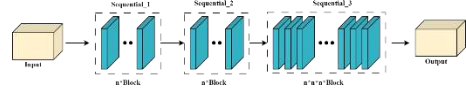


Fig. 8. Stage module

Each block utilizes depthwise separable convolutions to reduce model complexity, reinforced by BatchNorm and LayerNorm for training stability. Furthermore, the architecture integrates a ResNeXt module (**Fig. 9**) to facilitate identity mapping, thereby enhancing deep adaptability and preventing performance degradation.

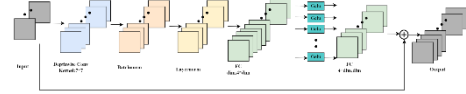


Fig. 9. Block module

3.2.5 Feature fusion

This paper proposes a multi-modal feature fusion network. For time series features, the input sequence is defined as $Z = (Z_1, Z_2, \dots, Z_N) \in R^{N \times 1 \times Len}$, N is the batch size, and Len represents the length of the time series features. For time-invariant features, input features $Y = (Y_1, Y_2, \dots, Y_N)$, and then use the time-invariant feature extraction network to calculate the input features. For spatial frequency domain features, the input feature is $X = (X_1, X_2, \dots, X_N)$. Then, the input features are calculated by the spatial frequency domain feature extraction network to obtain another set of output results. Here, this paper combines time series features, time invariant features and spatial frequency domain features. The formula is as follows :

$$C = Concat(O_1, O_2, O_3) \in R^{N \times 2 \times H \times W} \quad (31)$$

Where: O_1 is feature output of the temporal feature extraction network; O_2 is feature output of the time-invariant feature extraction network; O_3 is feature output of the spatial frequency domain feature extraction network; $Concat$ is sum of the feature outputs of the two networks.

In this paper, mean square error (MSE) is used as the loss function. The formula is as follows.

$$MSE = \frac{1}{m} \sum_{i=1}^m (y_i - y'_i)^2 \quad (32)$$

Where: y_i is actual value; y'_i is predicted value; m is data size.

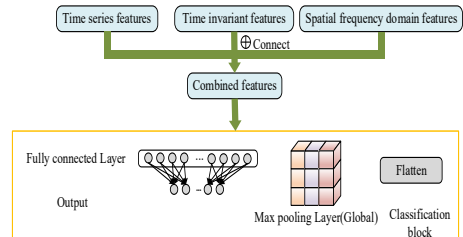


Fig.10. Schematic diagram of feature fusion

4 Experimental result analysis

4.1 Experimental environment configuration

To rigorously evaluate the MEFD-Net model, a series of comparative experiments were conducted against mainstream algorithms, verifying performance and stability under varying conditions. The experimental configuration is detailed in **Table 1**.

Optimal hyperparameter settings were determined through iterative testing. The Adam optimizer with a learning rate of 0.001 yielded superior training results. The training protocol specified 200 iterations and a batch size of 32 to balance efficiency with performance. To mitigate overfitting, a dropout rate of 0.8 was applied. Furthermore, 64 LSTM units were employed to capture complex temporal patterns. The Mean Squared Error (MSE) served as the loss function to quantify the divergence between predicted and ground truth values.

Table 1 Experiment environment configuration

Operating Environment	Detailed Configuration
System	Windows 11
Graphics Card	NVIDIA GeForce GTX 2080 Ti
RAM	40G
VRAM	11G
CUDA Version	10.2
Development Language	Python 3.8
Deep Learning Framework	PyTorch 1.8

4.2 Experimental Evaluation Indicators

In order to accurately evaluate the performance of the MEFD-Net model, this paper uses five indicators to evaluate the performance of the model : Accuracy, Precision, Recall, Missrate and F1-score. The calculation formulas of these indicators are as follows :

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (33)$$

$$Precision = \frac{TP}{TP + FP} \quad (34)$$

$$Recall = \frac{TP}{TP + FN} \quad (35)$$

$$Missrate = 1 - Recall \quad (36)$$

$$F1_{score} = \frac{2 * Precision * Recall}{Precision + Recall} \quad (37)$$

Where: TP is correct recognition of normal state; TN is correct recognition of fatigue state; FP is incorrect recognition of normal state; FN is incorrect recognition of fatigue state.

4.3 Experimental contrast analyses

To evaluate the efficacy of the temporal feature extraction network, a comparative analysis against standard time-series models including BiRNN, BiLSTM, BiGRU, tanh-BiLSTM, and elu-BiLSTM was conducted as illustrated in **Fig. 11**. Results indicate that advanced BiLSTM and BiGRU architectures outperform traditional BiRNNs. Notably, the elu-BiLSTM exhibited strong non-linear processing capabilities. However, the

proposed HG-BiLSTM achieved the highest accuracy, validating that the integration of the Hybrid Attention Mechanism and GeLU activation function significantly enhances fatigue detection performance.

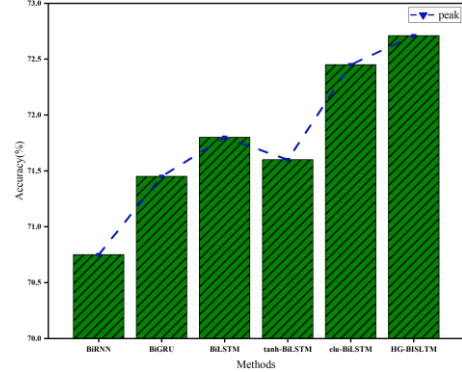


Fig. 11. Comparison experiment results of temporal feature extraction network

In order to verify the effect of time-invariant feature extraction network (WACNet) and analyze its influence on the accuracy of fatigue state detection, this paper designs a series of comparative experiments. These experiments combine time series models to evaluate the performance of time-invariant feature extraction networks in fatigue state detection. The experimental results are shown in **Fig. 12**. In the experiment, several commonly used time series processing models and the improved HG-BiLSTM are selected and combined with the time-invariant feature extraction model proposed in this paper. Based on the analysis results of **Fig. 11** and **Fig. 12**, it is found that all the time series feature networks combined with time-invariant features have improved classification performance. Among them, the combination of HG-BiLSTM and WACNet improved in this paper performs best in these models, which shows that the time-invariant feature extraction network WACNet proposed in this paper is effective. At the same time, the combination of the temporal feature extraction network and the time-invariant feature extraction network proposed in this paper can play the best effect and improve the detection effect of the overall model on the fatigue state.

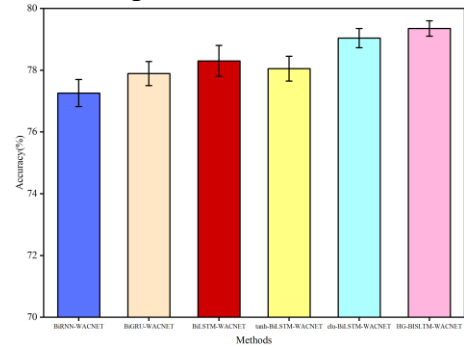


Fig. 12. The influence of time-invariant feature extraction network on fatigue state detection accuracy

4.3 Ablation experimental analysis

In order to verify the influence of spatial frequency domain feature extraction network on model performance and further understand the improvement effect of time series features and time invariant features on model accuracy. The ablation experiment was designed to analyze its influence on the accuracy of fatigue state detection. The experimental results are shown in **Table 2**.

Table 2 MEFD-Net Ablation Experiments

Model	HG-BiLSTM	WACNet	PCNN	Accuracy	Precision	Recall	Miss rate	F1-score
1	★			72.71%	65.12%	71.8%	28.1%	68.3%
2		★		78.74%	73.45%	77.9%	25.6%	73.8%
3			★	80.47%	75.04%	80.5%	20.9%	77.8%
4	★	★		79.35%	71.03%	77.4%	22.5%	72.6%
5	★		★	81.19%	72.81%	81.1%	19.7%	76.8%
6	★	★	★	82.59%	77.60%	80.7%	19.3%	78.7%

Note : ★ indicates that the corresponding strategy is applied.

The incorporation of spatial-frequency domain feature extraction universally enhanced model metrics while reducing the missed diagnosis rate. Specifically, Model 6, which synthesizes HG-BiLSTM, WACNet, and PCNN, achieved optimal performance with an accuracy of 82.59%, precision of 76.60%, recall of 80.71%, a missed diagnosis rate of 19.38%, and an F1 score of 78.72%. These results corroborate the efficacy of the multi-feature extraction strategy.

To further validate the MEFD-Net, a comparative analysis against existing deep learning methods on the SEED-VIG dataset is presented in **Table 3**. Benchmarks included ST Encoding CNN, CNN-attention, CNN-lstm, BiLSTM, ESTCNN, EEGNet, and Interpretable CNN. The proposed MEFD-Net outperformed these established methodologies and demonstrated its capability to effectively analyze non-linear EEG data for precise binary classification. Consequently, the integration of time-domain, time-invariant, and spatial-frequency features is proven effective and establishes the model as a robust tool for fatigue detection.

Table 3. Performance comparison of this algorithm with other mainstream algorithms

Model	Accuracy	Precision	Recall	Miss rate	F1-score
ST_Encoding_CNN	69.98%	77.13%	72.60%	27.40%	74.80%
CNN+LSTM	75.66%	78.27%	76.38%	23.62%	77.31%
BiLSTM	69.55%	63.96%	69.15%	30.85%	66.45%
ESTCNN	78.95%	75.93%	78.45%	21.55%	77.17%

EEGNet	79.23%	71.39%	75.82%	24.18%	73.53%
Interpretable_CNN	79.62%	75.17%	78.49%	21.51%	76.79%
MEFD-Net	82.59%	77.60%	80.71%	19.38%	78.72%

5. Conclusion

To address limitations in feature extraction and generalization regarding non-linear and non-stationary EEG signals, this study proposes MEFD-Net, a multimodal feature fusion framework. The architecture synergizes an enhanced BiLSTM, a Weighted Adaptive Multi-scale Convolutional Network, and a Pure Convolutional Neural Network to comprehensively capture temporal, time-invariant, and spatial-frequency patterns. Additionally, WGAN-GP is employed to mitigate class imbalance and enhance robustness. Empirical evaluation on the SEED-VIG dataset demonstrates superior performance with 82.59% accuracy, 80.71% recall, and an F1 score of 78.72%, significantly surpassing established baselines including ST-Encoding-CNN and EEGNet. Ablation studies further validate the efficacy of the multimodal fusion strategy. Future research will focus on lightweight architecture design and dynamic adaptive mechanisms to optimise real-time performance and individual adaptability in practical scenarios.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

CRediT authorship contribution statement

Zhonghua Xu: Writing - review & editing, Writing - original draft, Visualization, Software, Resources, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. Qiong Hong: Supervision, Project administration, Formal analysis, Conceptualization. Lianshan Pan: Writing - review & editing, Data curation. Peng Wang: Data curation. Hao Zhang: Funding acquisition, Methodology, Data curation. Jie He: Formal analysis, Data curation.

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.

Funding

This work was supported by the National Natural Science Foundation of China (71874067), High-level talent project of "Six Talent Peaks" in Jiangsu Province (JFW049) and "Qing Lan Project" in Jiangsu Province.

References

- [1] Tian Y, Cao J. Fatigue driving detection based on electrooculography: a review[J]. E URASIP Journal on Image and Video Pro

- cessing, 2021, 2021(1): 33. <https://doi.org/10.1186/s13640-021-00575-1>
- [2] Gao D, Liu S, Gao Y, et al. A comprehensive adaptive interpretable Takagi–Sugeno–Kang fuzzy classifier for fatigue driving detection[J]. *IEEE Transactions on Fuzzy Systems*, 2024, 33(1): 108-119. <https://doi.org/10.1109/TFUZZ.2024.3399400>
- [3] Peng Z, Rong J, Wu Y, et al. Exploring the different patterns for generation process of driving fatigue based on individual driving behavior parameters[J]. *Transportation research record*, 2021, 2675(8): 408-421. <https://doi.org/10.1177/0361198121998351>
- [4] Huang A C, Yuan C, Meng S H, et al. Design of fatigue driving behavior detection based on circle hough transform[J]. *Big data*, 2023, 11(1): 1-17. <https://doi.org/10.1089/big.2021.0166>
- [5] Hu J, Xu L, He X, et al. Abnormal driving detection based on normalized driving behavior[J]. *IEEE Transactions on Vehicular Technology*, 2017, 66(8): 6645-6652. <https://doi.org/10.1109/tvt.2017.2660497>
- [6] Xi J, Wang S, Ding T, et al. Detection model on fatigue driving behaviors based on the operating parameters of freight vehicles[J]. *Applied Sciences*, 2021, 11(15): 7132. <https://doi.org/10.3390/app11157132>
- [7] Yu W, Deng L, Wu Y H. Fatigue driving behavior recognition method based on image and vehicle feature change[J]. *Advances in transportation studies*, 2022, 2. <https://doi.org/10.53136/97912599481132>
- [8] Hou M, Wang M, Zhao W, et al. A light weight framework for abnormal driving behavior detection[J]. *Computer Communications*, 2022, 184: 128-136. <https://doi.org/10.1016/j.comcom.2021.12.007>
- [9] Li K, Gong Y, Ren Z. A fatigue driving detection algorithm based on facial multi-feature fusion[J]. *IEEE Access*, 2020, 8: 101244-101259. <https://doi.org/10.1109/access.2020.2998363>
- [10] Jia H, Xiao Z, Ji P. Real-time fatigue driving detection system based on multi-module fusion[J]. *Computers & Graphics*, 2022, 108: 22-33. <https://doi.org/10.1016/j.cag.2022.09.001>
- [11] Kang Z, Yang X. Fatigue Driving Characteristics Analysis Based on Data of Driver Monitoring System[M]//*CICTP 2020*. 2020: 4312-4323. <https://doi.org/10.1061/9780784482933.370>
- [12] Albousefi A A, Ying H, Filev D, et al. A two-stage-training support vector machine approach to predicting unintentional vehicle lane departure[J]. *Journal of Intelligent Transportation Systems*, 2017, 21(1): 41-51. <https://doi.org/10.1080/15472450.2016.1196141>
- [13] Li Z, Li S E, Li R, et al. Online detection of driver fatigue using steering wheel angles for real driving conditions[J]. *Sensors*, 2017, 17(3): 495. <https://doi.org/10.3390/s17030495>
- [14] Krajewski J, Schnieder S, Sommer D, et al. Applying multiple classifiers and non-linear dynamics features for detecting sleepiness from speech[J]. *Neurocomputing*, 2012, 84: 65-75. <https://doi.org/10.1016/j.neucom.2011.12.021>
- [15] Greeley H P, Berg J, Friets E, et al. Fatigue estimation using voice analysis[J]. *Behavior research methods*, 2007, 39(3): 610-619. <https://doi.org/10.3758/BF03193033>
- [16] Vekkot S, Chavali S T, Kandavalli C T, et al. Continuous speech-based fatigue detection and transition state prediction for air traffic controllers[J]. *IEEE Access*, 2024. <https://doi.org/10.1109/access.2024.3524452>
- [17] Chen S, Sun Y, Zhang H, et al. Speech fatigue detection based on deep learning[C]//*Journal of Physics: Conference Series*. IOP Publishing, 2022, 2224(1): 012023. <https://doi.org/10.1088/1742-6596/2224/1/012023>
- [18] Xu L, Ma S, Shen Z, et al. Analyzing multi-mode fatigue information from speech and gaze data from air traffic controllers[J]. *Aerospace*, 2023, 11(1): 15. <https://doi.org/10.3390/aerospace11010015>
- [19] Huang Z, Tang W, Tian Q, et al. Air traffic controller fatigue detection based on facial and vocal features using long short-term memory[J]. *IEEE Access*, 2024, 12: 56663-56682. <https://doi.org/10.1109/access.2024.3391427>
- [20] Wu N, Sun J. Fatigue detection of air traffic controllers based on radiotelephony communications and self-adaption quantum genetic algorithm optimization ensemble learning[J]. *Applied Sciences*, 2022, 12(20): 10252. <https://doi.org/10.3390/app122010252>
- [21] Mu S, Liao S, Tao K, et al. Intelligent fatigue detection based on hierarchical multi-scale ECG representations and HRV measures[J]. *Biomedical Signal Processing and Control*, 2024, 92: 106127. <https://doi.org/10.1016/j.bspc.2024.106127>
- [22] Zhao L, Li M, He Z, et al. Data-driven learning fatigue detection system: A multi-modal fusion approach of ECG (electrocar

- diagram) and video signals[J]. *Measurement*, 2022, 201: 111648. <https://doi.org/10.1016/j.measurement.2022.111648>
- [23] Wang F, Yao W, Lu B, et al. ECG-based real-time drivers' fatigue detection using a novel elastic dry electrode[J]. *IEEE Transactions on Instrumentation and Measurement*, 2023, 73: 1-16. <https://doi.org/10.1109/tim.2023.3343799>
- [24] Kołodziej M, Tarnowski P, Sawicki D J, et al. Fatigue detection caused by office work with the use of EOG signal[J]. *IEEE Sensors Journal*, 2020, 20(24): 15213-15223. <https://doi.org/10.1109/jsen.2020.3012404>
- [25] Song M, Li L, Guo J, et al. A new method for muscular visual fatigue detection using electrooculogram[J]. *Biomedical signal processing and control*, 2020, 58: 101865. <https://doi.org/10.1016/j.bspc.2020.101865>
- [26] Tag B, Vargo A W, Gupta A, et al. Continuous alertness assessments: Using EOG glasses to unobtrusively monitor fatigue levels In-The-Wild[C]//*Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019: 1-12. <https://doi.org/10.1145/3290605.3300694>
- [27] Kinugasa R, Kubo S. Development of consumer-friendly surface electromyography system for muscle fatigue detection[J]. *Ieee Access*, 2023, 11: 6394-6403. <https://doi.org/10.1109/access.2023.3237557>
- [28] Wang L, Wang H, Jiang X. A new method to detect driver fatigue based on EMG and ECG collected by portable non-contact sensors[J]. *Promet-Traffic&Transportation*, 2017, 29(5): 479-488. <https://doi.org/10.7307/ptt.v29i5.2244>
- [29] Wu Q, Xi C, Ding L, et al. Classification of EMG signals by BFA-optimized GSV CM for diagnosis of fatigue status[J]. *IEEE Transactions on Automation Science and Engineering*, 2016, 14(2): 915-930. <https://doi.org/10.1109/tase.2016.2564419>
- [30] Wang H, Dragomir A, Abbasi N I, et al. A novel real-time driving fatigue detection system based on wireless dry EEG[J]. *Cognitive neurodynamics*, 2018, 12(4): 365-376. <https://doi.org/10.1007/s11571-018-9481-5>
- [31] Subasi A, Saikia A, Bagedo K, et al. EEG-based driver fatigue detection using FA WT and multiboosting approaches[J]. *IEEE Transactions on Industrial Informatics*, 2022, 18(10): 6602-6609. <https://doi.org/10.1109/tii.2022.3167470>
- [32] Jap B T, Lal S, Fischer P, et al. Using EEG spectral components to assess algorithms for detecting fatigue[J]. *Expert Systems with Applications*, 2009, 36(2): 2352-2359. <https://doi.org/10.1016/j.eswa.2007.12.043>
- [33] Zheng W L, Lu B L. A multimodal approach to estimating vigilance using EEG and forehead EOG[J]. *Journal of neural engineering*, 2017, 14(2): 026017. <https://doi.org/10.1088/1741-2552/aa5a98>
- [34] Gao X Y, Zhang Y F, Zheng W L, et al. Evaluating driving fatigue detection algorithms using eye tracking glasses[C]//*2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2015: 767-770. <https://doi.org/10.1109/ner.2015.7146736>
- [35] Graves A, Graves A. Long short-term memory[J]. *Supervised sequence labelling with recurrent neural networks*, 2012: 37-45. https://doi.org/10.1007/978-3-642-24797-2_4