Driving Fatigue Detection Based on Hybrid Electroencephalography and Eye Tracking

Zequan Lian, Tao Xu, Zhen Yuan, Junhua Li, *Senior Member, IEEE*, Nitish Thakor, *Fellow, IEEE*, Hongtao Wang, *Senior Member, IEEE*

***Abstract*— EEG-based unimodal method has demonstrated significant success in the detection of driving fatigue. Nonetheless, data from a single modality might be not sufficient to optimize fatigue detection due to incomplete information. To address this limitation and enhance the performance of driving fatigue detection, a novel multimodal architecture combining hybrid electroencephalograph (EEG) and eye tracking data was proposed in this work. Specifically, the EEG and eye tracking data were separately input into encoders, generating two one-dimensional (1D) features. Subsequently, these 1D features were fed into a cross-modal predictive alignment module to improve fusion efficiency and two 1D attention modules to enhance feature representation. Furthermore, the fused features were recognized by a linear classifier. To evaluate the effectiveness of the proposed multimodal method, comprehensive validation tasks were conducted, including intra-session, cross-session, and cross-subject evaluations. In the intra-session task, the proposed architecture achieves an exceptional average accuracy of 99.93%. Moreover, in the cross-session task, our method demonstrates an average accuracy of 88.67%, surpassing the performance of EEG-only approach by 8.52%, eye tracking-only method by 5.92%, multimodal deep canonical correlation analysis (DCCA) technique by 0.42%, and multimodal deep generalized canonical correlation analysis (DGCCA) approach by 0.84%. Similarly, in the cross-subject task, the proposed approach achieves an average accuracy of 78.19%, outperforming EEG-only method by 5.87%, eye tracking-only approach by 4.21%, DCCA method by 0.55%, and DGCCA approach by 0.44%. The experimental results conclusively illustrate the superior effectiveness of the proposed method compared to both single modality approaches and canonical correlation analysis-based multimodal methods.**

***Index Terms*— electroencephalograph, eye tracking, fatigue detection, multi-modality, cross-modal alignment**

This work was supported in part by Special Projects in Key Fields Supported by Wuyi University and Hong Kong & Macao joint Research Project under Grant 2019WGALH16, in part by Projects for International Scientific and Technological Cooperation of Guangdong Province under Grant 2023A0505050144. This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Committee of Jiangmen Central Hospital under Application No. [2021]8A, and performed in line with hospital requirements.

Zequan Lian and Hongtao Wang are with the School of Electronics and Information Engineering, Wuyi University, Jiangmen 529020, China (e-mail: hongtaowang@wyu.edu.cn).

Tao Xu is with the Department of Biomedical Engineering at Shantou University, Shantou, China.

Zhen Yuan is with the Faculty of Health, University of Macau, Macau 999078, China, and also with the Centre for Cognitive and Brain Sciences, University of Macau, Macau 999078, China.

Junhua Li is with the School of Computer Science and Electronic Engineering, University of Essex, CO4 3SQ Colchester, U.K.

Nitish Thakor is with the Department of Biomedical Engineering, Johns Hopkins University, United States.

I. Introduction

DRIVING fatigue has emerged as a significant contributor to the increasing number of accidents in contemporary society, incurring financial costs and endangering human lives. In fact, driving fatigue has become a leading cause of road fatalities in various countries [1]. Numerous researchers have dedicated their efforts to developing efficient and robust methods for detecting fatigue and mitigating the consequences of fatigue-related incidents [2].

In the driving fatigue detection methods, various features are adopted as the input. Recently, researchers have explored physiological and behavioral features for modeling, including electroencephalogram (EEG) [3]–[9], electrooculogram (EOG) [10], eye closure features [11], eye-tracking-based features [12], yawn/nod motions [13], etc. Among these, physiological features have received the most attention due to their ability to directly reflect mental states [14]. The fatigue detection methods typically utilize one of these features to train a high-performance classifier. However, physiological methods, particularly EEG-based approaches, sometimes perform poorly when transferring well-trained classifiers to different subjects (cross-subject) or different sessions with the same subject (cross-session). This implies that well-trained classifiers today may not be effective tomorrow. Experts attribute this phenomenon to domain shift, which is likely caused by distribution variances among different sessions and subjects [15].

There are two primary approaches to mitigating distribution variances across different sessions and subjects. One approach involves utilizing transfer learning techniques including domain adaptation (DA) methods and domain generalization (DG) methods [16], [17]. Another approach involves leveraging data from hybrid modalities, commonly referred to as the multimodal approach. The multimodal approach provides multiple perspectives that capture various aspects of a specific object or phenomenon. Each modality contributes unique information, allowing them to complement one another. Furthermore, multimodal methods have shown significant promise in enhancing the performance of brain-computer interface (BCI) systems, including SSVEP-BCI [18] , MI-BCI [19], [20] , P300-BCI [21] , and others [22], [23].

In the past decade, there have been preliminary works on detecting fatigue using hybrid modalities. Zheng *et al.* proposed a hybrid approach for estimating vigilance by combining EEG and forehead EOG data, with their data labeled using the percentage of eye closure (PERCLOS) index calculated from eye tracking data [24]. The limitation of their study involves: i) their PERCLOS index was used for labeling, while the generation of this index was based on the missed records of eye-tracking glasses, which could occur due to issues such as tracking device overheating or other failures. Consequently, the direct representation of the subject's mental fatigue state by their index is uncertain; ii) their eye-movement features were generated based on the EOG data, which could not capture the dynamic amplitude and velocity of eye movements. Nonetheless, their study demonstrated an improvement in adopting a hybrid modalities approach. Some recent studies have also demonstrated the effectiveness of EEG- and EOG-based multimodal approaches [25], [26] and the efficacy of other hybrid modalities methods [27], [28]. Moreover, numerous studies focusing on eye-tracking unimodal features have shown that, in addition to PERCLOS, metrics related to saccades and fixations are effective indicators for fatigue detection [29], [30]. Therefore, we propose leveraging a combination of EEG and eye-tracking data to improve the detection of driving fatigue, utilizing eye-tracking glasses to capture the eye-movement features.

A typical architecture for multimodal fatigue detection comprises four primary modules: preprocessing of unimodal data, encoding of unimodal features, multimodal fusion and encoding, and classification. The encoding of unimodal features may be omitted in some architectures [24], [31]. Among these modules, multimodal fusion and encoding are crucial as they directly impact the architecture's performance. Previous studies predominantly employed feature-level fusion strategies, including direct fusion [25], [32], subspace feature fusion [31], empirical weight fusion [33], canonical correlation analysis fusion [6], [34], and neural network fusion [26]. Neural network fusion approaches are specifically designed for deep learning methods, utilizing additional network modules to fuse multimodal features, with efficacy dependent on module design. In other multimodal BCI paradigms, deep canonical correlation analysis (DCCA) and deep generalized canonical correlation analysis (DGCCA) have proven to be effective fusion methods [35], [36]. However, CCA-based fusion strategies may limit the model's ability to learn separable representations. Therefore, we proposed a novel non-CCA cross-modal predictive alignment module to enhance fusion efficacy.

In this study, we proposed a driving fatigue detection architecture by integrating hybrid EEG and eye tracking modalities. This approach considers both brain-derived information and eye-tracking states comprehensively. To improve the fusion efficiency and enhance the feature representation, a cross-modal predictive alignment module and one-dimensional (1D) attention module are proposed respectively. This approach significantly bolsters the architecture's robustness and further elevates the detection performance in intra-session, cross-session, and cross-subject tasks.

The arrangement of this study is organized as follows. Section II shows the materials and methods of our study, which introduces the experimental protocol, the EEG data acquisition, the data preprocessing and labeling, and the multimodal architecture. In Section III and IV, we present the results and discussion. The conclusion is presented in Section V.

II. Materials and Methods

A. Experimental Protocol

We experimented with a hybrid dataset comprising multiple modalities. The experiment was carefully designed to mimic a real driving scenario, aiming to acquire more effective multimodal experimental data. Fourteen healthy participants (9 males and 5 females) with an age of 21.6 ± 1.4 were recruited. All participants were newcomers to the experiment and had prior driving experience. These participants were screened to ensure they had normal or corrected-to-normal vision and had no history of neurological illness.

The driving simulation platform comprises three main components, as illustrated in Fig. 1(b): a Thrustmaster T300 GT vehicle suite, a 32-channel Brain Products (BP) LiveAmp cap for EEG data collection, and a pair of Tobii Glasses 2 for eye-tracking data acquisition. The 32-channel EEG cap efficiently captures signals from crucial brain regions while offering ease of wear. All participants were required to obey traffic regulations, avoiding collisions with other vehicles. Additionally, they were instructed to minimize any movements unrelated to driving operations to prevent interference during data collection. The driving experiment lasts 90 minutes for a session, with each subject undergoing two sessions within a week, totaling 28 sessions across 14 subjects. The experiments were conducted between 2−5 p.m., a period during which drivers are more likely to experience a state of fatigue. Written informed consent was obtained from each participant. The Institutional Review Committee of Jiangmen Central Hospital approved this study ([2021]8A).

B. Data Processing and Labeling

1) EEG: The EEG data was sampled at a rate of 250 Hz, with the FCz channel as the reference. Electrode impedance across all EEG channels was kept below 20 kΩ throughout the driving simulation to prevent interference. EEGLab was utilized for preprocessing [37], the preprocessing procedure included bad-channel interpolation, applying a 1-40Hz band-pass filter, and removing eye artifacts using ICA. The processed data was segmented into non-overlapping 1-second time windows and normalized by Z-score method.

2) Eye-tracking Data: The sample rate of eye-tracking data was set at 100 Hz. In most cases, the eye-tracking data was processed by the given software suite (Tobbi Pro Lab in this study). However, the exported data of the software suite lacks several important indices (e.g., saccade amplitude and saccade velocity). Therefore, the raw eye-tracking data was further processed through several steps to derive some crucial indices and finally generate 8 eye-tracking characteristics. The overall processing procedure is presented in Fig. 2. Two types of data from eye-tracking were adopted, distinguished by red and blue colors in the figure: physical data and tracking data. Among

(a) Experimental scenario (b) Close-up details

Fig. 1. The system of multimodal driving fatigue detection. (a) experimental scenario. (b) a subject wearing an EEG cap and a pair of Tobbi Glasses was driving a simulated vehicle in the simulation platform.

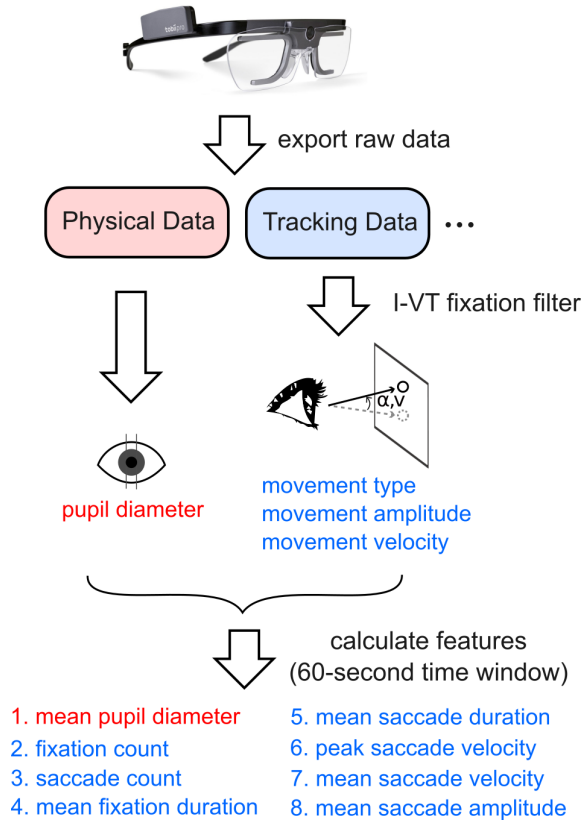


Fig. 2. The overall framework of eye-tracking data processing. The data was first recorded on an SDCard of Tobbi Glasses and then exported by Tobbi Pro Lab. Among the exported data, the tracking data was further processed to obtain several critical indices by a specially designed I-VT fixation filter. Finally, a 60-second sliding window with a 1-second time step was applied to generate 8 characteristics.

them, eye-tracking data was first filtered by a special Velocity-Threshold Identification (I-VT) fixation filter [38] to extract three key indices and further calculated to obtain features. We adopt an optimal 60-second time window (results of using various time windows are shown in Section III-C) for feature generation with a time step of 1 second. The time step matches that of the EEG data, allowing for a multimodal combination. The 8 eye-tracking features are listed at the bottom of Fig. 2. Before sending to the multimodal model, the features were standardized by Z-score method.

3) Labeling: In this study, a labeling scheme based on the driving duration was employed [2]. Specifically, the initial 10 minutes of each driving session were assigned the label of "alert" state, and the final 10 minutes were designated as the "fatigue" state. To validate the appropriateness of the adopted labeling scheme, we conducted post-experimental questionnaires to evaluate participants' subjective experiences and perceptions of fatigue toward the end of the driving experiment. The analysis indicated that the majority of participants reported experiencing fatigue, affirming the efficacy of the labeling method.

C. Multimodal Architecture

A novel multimodal architecture for fatigue detection was proposed in this study. As demonstrated in Fig. 3(a), the primary components include: i) an EEG convolutional network for EEG encoding, ii) a Multi-Layer Perceptron (MLP) network for eye-tracking data, iii) a novel non-CCA cross-modal predictive alignment for feature fusion, and iv) a compact one-dimensional module for feature enhancement.

1) EEG Convolutional Network: The structure of the proposed EEG convolutional network is illustrated in Fig. 3(b). The network consists of three convolutional blocks serving as temporal filters, spatial filters, and deep temporal filters, respectively. DepthWise and Separable convolutional layers are adopted to reduce the network's parameters [39]. The hyperparameters F_1, D, and F_2 control the filter size and were set to 8, 2, and 16, respectively. Here, D represents the depth amplification factor for the DepthWise convolution. Notably, the two pooling layers in the network result in a total downsampling of 32 times, meaning that the sample rate of the EEG signal is reduced to approximately 8Hz before embedding generation.

In this study, the EEG encoder is tasked with processing input signals of dimension 32×250 within a brief one-second temporal window, while the eye-tracking encoder operates on

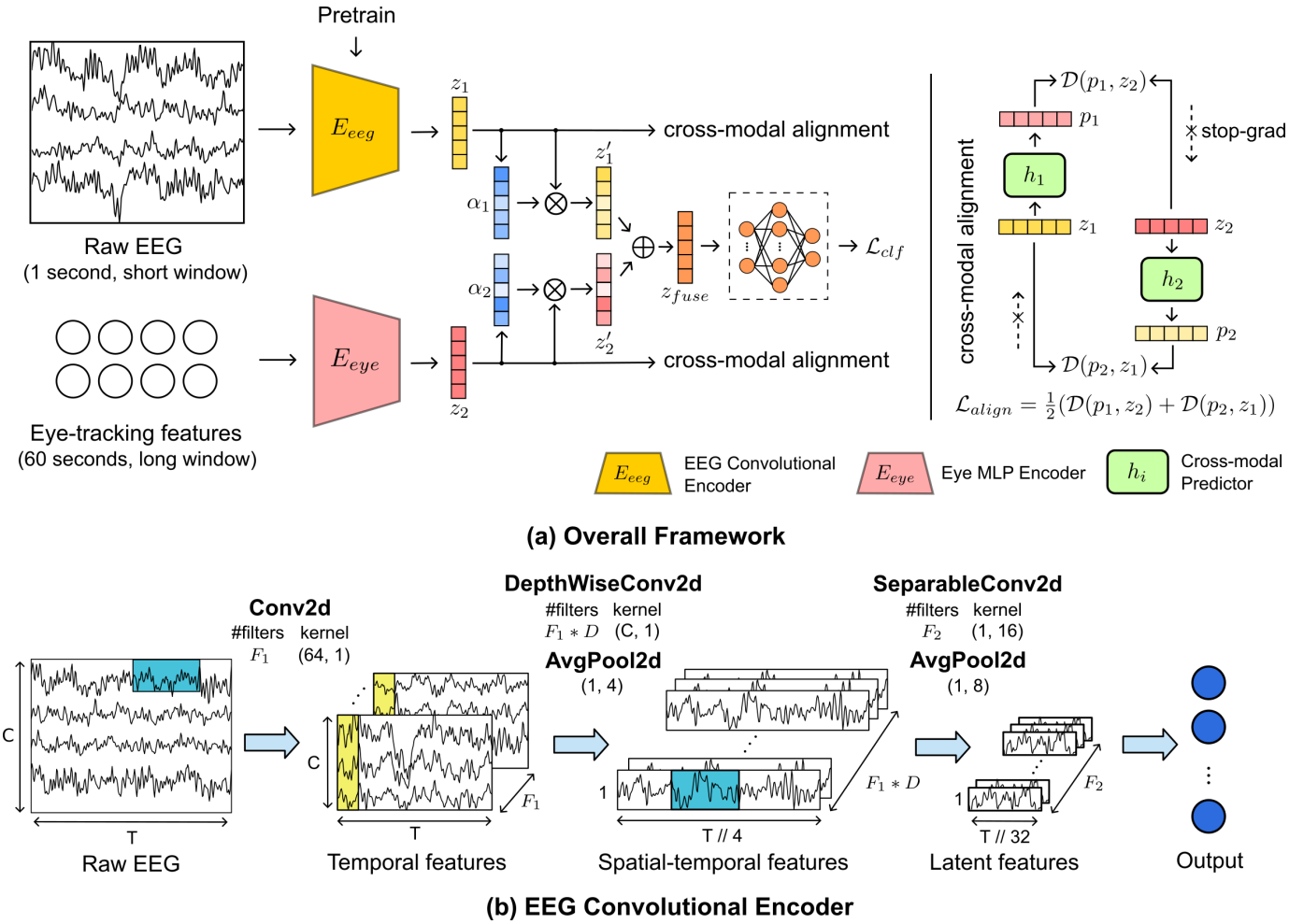


Fig. 3. The framework of the proposed multimodal approach. (a) illustrates the overall framework. (b) presents the details of the EEG convolutional encoder. Our approach leverages two distinct modalities of data, namely EEG and eye tracking, as inputs. By incorporating 1-second and 60-second time windows, both short and long temporal contexts are effectively integrated. The EEG data is encoded by a convolutional network, while the eye-tracking features are encoded using an MLP network. To enhance the representation of features, a 1D attention module is employed. In contrast to CCA-based multimodal approaches, we proposed a cross-modal predictive alignment module to align hybrid features.

1×8 features over an extended 60-second window. The adoption of varying window lengths facilitates feature extraction across both short and long temporal contexts. Furthermore, the uniform output dimensions of 1×64 for both encoders facilitate seamless multimodal fusion.

2) Cross-modal Predictive Alignment: We proposed a novel non-CCA alignment method, enhancing the fusion efficacy through the optimization of cross-modal predictive tasks. As shown at the right of Fig. 3(a), this approach improves the relevance of unimodal features without excessive parameters. The alignment of unimodal features is achieved through the utilization of a similarity loss function, comparing predicted results with their corresponding targets across both modalities.

Suppose there are two modalities, where x_1 and x_2 represent the input data of these modalities. Let $z_1 \triangleq f_1(x_1)$ and $z_2 \triangleq f_2(x_2)$ be the encoded one-dimensional features of x_1 and x_2, respectively. An MLP was employed to predict the features of z_2 from z_1, and vice versa:

$$\begin{aligned} p_1 &= h_1(z_1) \\ p_2 &= h_2(z_2), \end{aligned} \tag{1}$$

here, h_1 and h_2 represent the MLP predictors, p_1 and p_2 are the predicted results from z_1 and z_2, respectively. p_1 is intended to closely match z_2, and similarly, p_2 is intended to closely match z_1 (the predicted results and their target features are visually depicted in the same color family in Fig. 3(a)). Subsequently, the negative cosine similarity was computed between the predicted results and the target features using the following equation:

$$\mathcal{D}(p_1, z_2) = -\frac{p_1}{\|p_1\|_2} \cdot \frac{z_2}{\|z_2\|_2}, \tag{2}$$

$\mathcal{D}(p_1, z_2)$ represents the negative cosine similarity function, and the predictive alignment loss is defined as follows:

$$\mathcal{L}_{align} = \frac{1}{2}(\mathcal{D}(p_1, z_2) + \mathcal{D}(p_2, z_1)), \tag{3}$$

$\mathcal{L}_{align}$ denotes the alignment loss, the minimum value of which is -1. Inspired by Simsiam [40], the stop-gradient (stopgrad) operation was adopted to stop the gradient of the target feature vector while optimizing, so that the alignment task will only affect the parameters alongside the predictor's gradient propagation lane, avoiding the optimizer from quickly

accessing to a degenerated solution. Equation (2) was modified as follows:

$$\mathcal{D}(p_1, \text{stopgrad}(z_2)), \tag{4}$$

this means the z_2 here is treated as a constant vector with no gradient. Finally, the predictive alignment loss is changed as:

$$\mathcal{L}_{align} = \frac{1}{2}(\mathcal{D}(p_1, \text{stopgrad}(z_2)) + \mathcal{D}(p_2, \text{stopgrad}(z_1))), \tag{5}$$

with the iterations of backpropagation, the features of two modalities would be much more able to predict each other, thus more relative to each other in the latent space.

3) One-dimensional Attention Module: A one-dimensional attention module was adopted to enhance the representation of encoded features. As shown in the middle of Fig. 3(a), the proposed attention module is applied to the features encoded by the primary backbone of each modality.

Let $O_f \in \mathbb{R}^D$ represent the one-dimensional output feature of a modality, where D is the length of the feature vector. The attentional weights α are calculated as follows:

$$\begin{aligned} O_s &= f_s(O_f) \\ \alpha &= softmax(O_s), \end{aligned} \tag{6}$$

here, $O_s \in \mathbb{R}^D$ represents the score of D elements obtained by a score function f_s, where the score function applied is a 2-layer MLP. Once the attentional weights are computed, the enhanced feature is obtained as follows:

$$O'_f = \alpha \otimes O_f, \tag{7}$$

$\otimes$ denotes the Hadamard product. O'_f represents the enhanced feature after the attention module.

4) Training: In this architecture, the final loss function consists of two parts: a cross-modal alignment loss ($\mathcal{L}_{align}$) and a cross-entropy classification loss ($\mathcal{L}_{clf}$). The overall loss function is expressed as follows:

$$\mathcal{L} = \mathcal{L}_{align} + \mathcal{L}_{clf}. \tag{8}$$

Given the substantial variation in data magnitude across different modalities, a two-stage training approach was adopted for this multimodal network to mitigate the risk of potential overfitting. The first stage involves the pretraining of the EEG convolutional network using exclusively EEG unimodal data to acquire appropriate initial parameters. The second stage comprises the training of the entire multimodal network.

The architecture was implemented using PyTorch 1.12.1 (Python 3.8.16) on a workstation equipped with an Intel(R) Core(TM) i5-9400F CPU and an NVIDIA GeForce RTX 2080 Ti GPU. The AdamW optimizer with a learning rate of 0.001 and a weight decay of 0.01 was employed in this work. Model training extended for 200 epochs, with accuracy serving as the evaluation metric for model selection. All validation tasks and the pretraining stage followed the same setup. To mitigate potential variations, five runs was conducted for the performance evaluation tasks.

D. Evaluation Method

To comprehensively evaluate the proposed method, three tasks were conducted [41], [42]: i) intra-session evaluation task, which evaluates the performance of each session, the result of each subject is the average of two sessions. The data pertaining to each label within a singular session is partitioned chronologically into training, validation, and testing sets, with allocations of 50%, 25%, and 25%, respectively; ii) cross-session evaluation task, which evaluates the performance across two sessions for every subject. A leave-one-session-out cross-validation strategy was adopted for this task. Specifically, one complete session was used as the training set, while the other session was utilized for validation and testing; iii) cross-subject evaluation task, which evaluates the performance across all subjects. Inspired by [43], a leave-one-subject-out cross-validation strategy was adopted. Specifically, in each validation fold, the data of a specific subject was adopted for validation and testing, and the data of the rest subjects were used for training [44], [45].

The performance of the proposed multimodal architecture is compared with other methods, including: i) EEG method, an unimodal method that utilizes only EEG signals as input, with the same structure as the proposed EEG Convolutional Encoder in Fig. 3(b); ii) eye tracking method, another unimodal method that uses only eye-tracking features as input; iii) DCCA-based multimodal method, a multimodal method employs DCCA as the fusion strategy, with the backbone encoders for both modalities data identical to the proposed architecture [46]; iv) DGCCA-based multimodal method, another multimodal method that utilizes DGCCA as the fusion strategy [36].

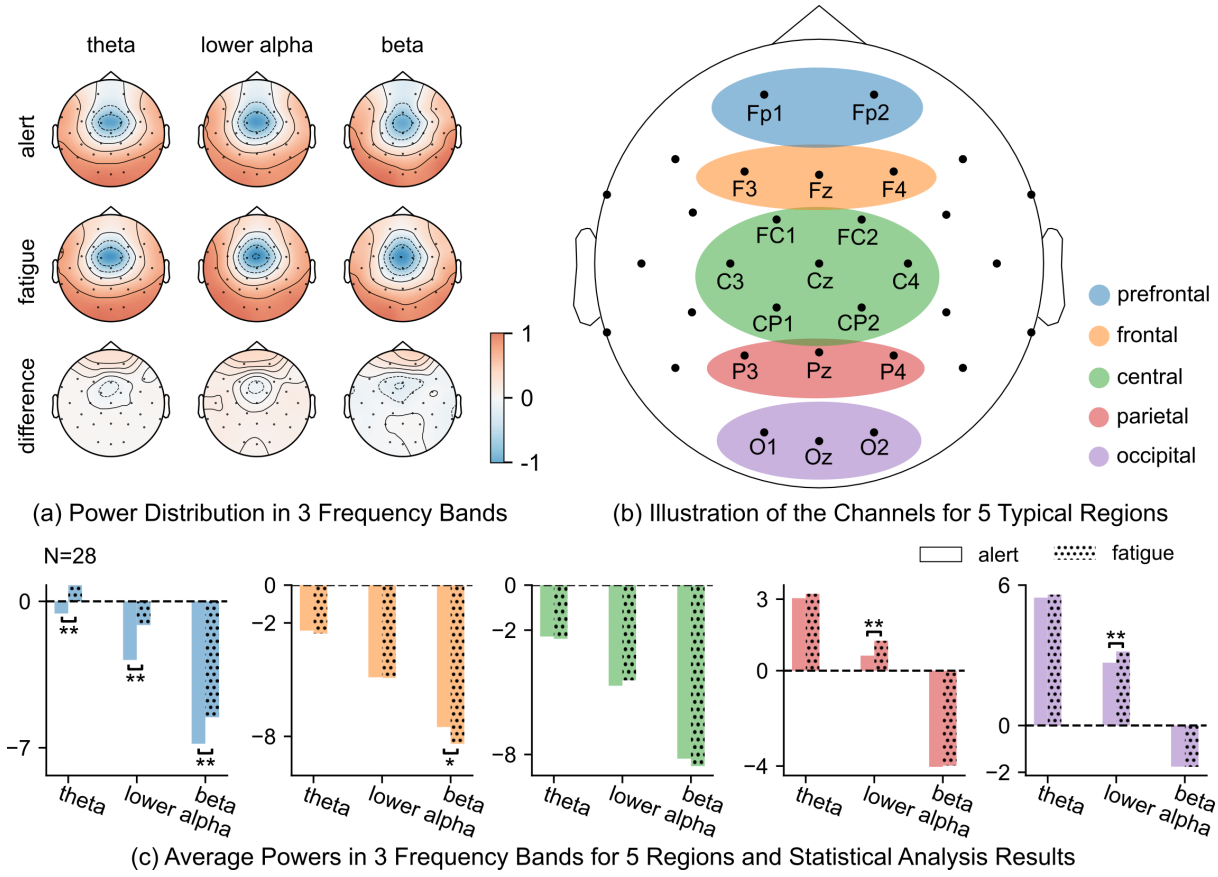
III. Experimental Results

This section presents the experimental results of the study. First, we analyzed the data from the two modalities to assess the feasibility of using multimodal data and the effectiveness of the adopted labeling method. Second, fatigue detection performance was evaluated across intra-session, cross-session, and cross-subject scenarios, respectively. Third, we investigated the impact of different eye-tracking time windows. Fourth, an ablation study was conducted to assess the effectiveness of each module in the proposed multimodal architecture. Finally, we conducted cross-subject evaluation experiments using the publicly available SEED-VIG data.

A. Analyses of Multimodal Data

1) Visualization and Analysis of EEG Data: Firstly, we conducted a visual analysis of EEG signal power distribution across three prevalent frequency bands: theta, lower alpha, and beta. Statistical analyses were then performed for five specific brain regions. The visualization results are depicted in Fig. 4. The statistical analyses, as shown in Fig. 4(b), targeted the prefrontal cortex, frontal lobe, central area, parietal lobe, and occipital lobe.

The power distribution changes across three frequency bands are illustrated in Fig. 4(a). The top and middle rows display the average band power during the alert state and the

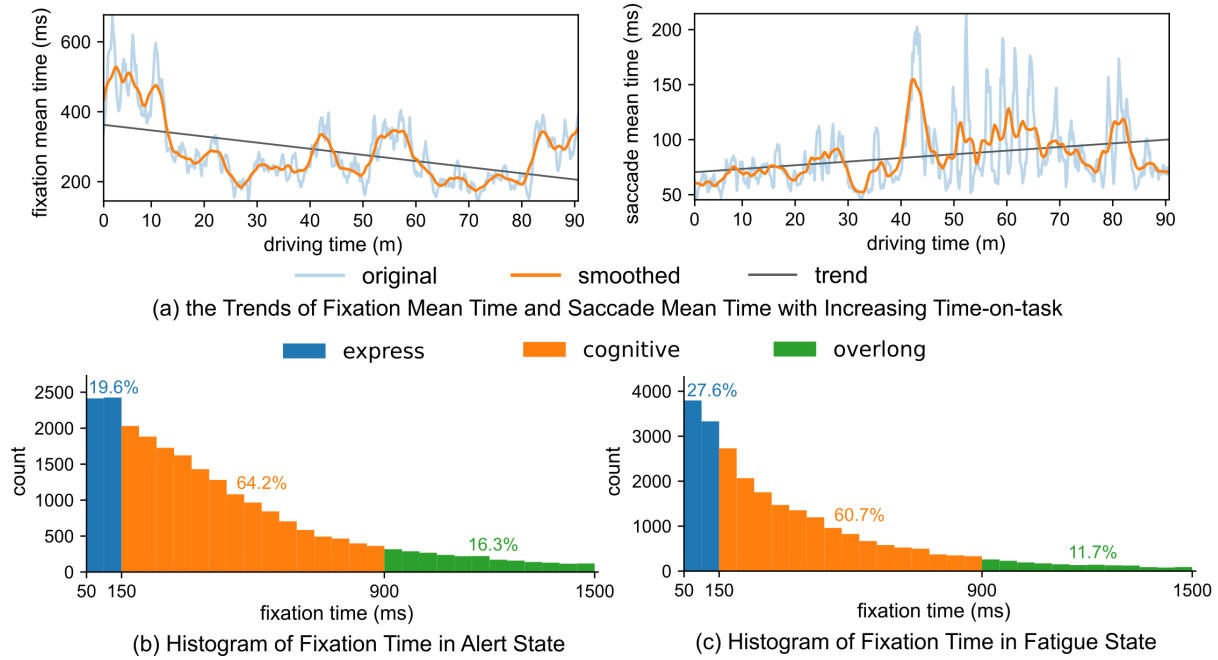


(a) Power Distribution in 3 Frequency Bands

(b) Illustration of the Channels for 5 Typical Regions

(c) Average Powers in 3 Frequency Bands for 5 Regions and Statistical Analysis Results

Fig. 4. The EEG topographies of power distribution in three frequency bands and the subsequent statistical analysis across five brain regions are presented. These topographies are organized in a 3x3 grid, as depicted in (a), with each column representing a distinct frequency band and each row signifying different cognitive states: the alert state, fatigue state, and the subtraction of the fatigue state from the alert state. Notably, the color representations in (b) and (c) are consistent. In (b), the channels chosen for analysis in the five brain regions are shown. Meanwhile, (c) illustrates the average power across three frequency bands for the five regions in two mental states, accompanied by the corresponding significance by Wilcoxon test. The Wilcoxon test was performed on 28 driving sessions ("*" indicates $p<0.05$, and "**" indicates $p<0.01$).



(a) the Trends of Fixation Mean Time and Saccade Mean Time with Increasing Time-on-task

(b) Histogram of Fixation Time in Alert State

(c) Histogram of Fixation Time in Fatigue State

Fig. 5. The trends of 2 eye-tracking features for a subject (Subject 7) throughout the entire 90-minute driving session and the histograms of fixation time in 2 mental states for all subjects. (a) illustrates the trend of fixation mean time and saccade mean time for a subject. (b) and (c) depict histograms of fixation times in both the alert and fatigue states for all subjects. The fixation times were categorized into three classes: express (<150ms), cognitive (150ms—900ms), and overlong(>900ms). The counts of express, cognitive, and overlong fixations were 4840, 15873, and 4029 in the alert state, and 7127, 15681, and 3030 in the fatigue state, respectively.

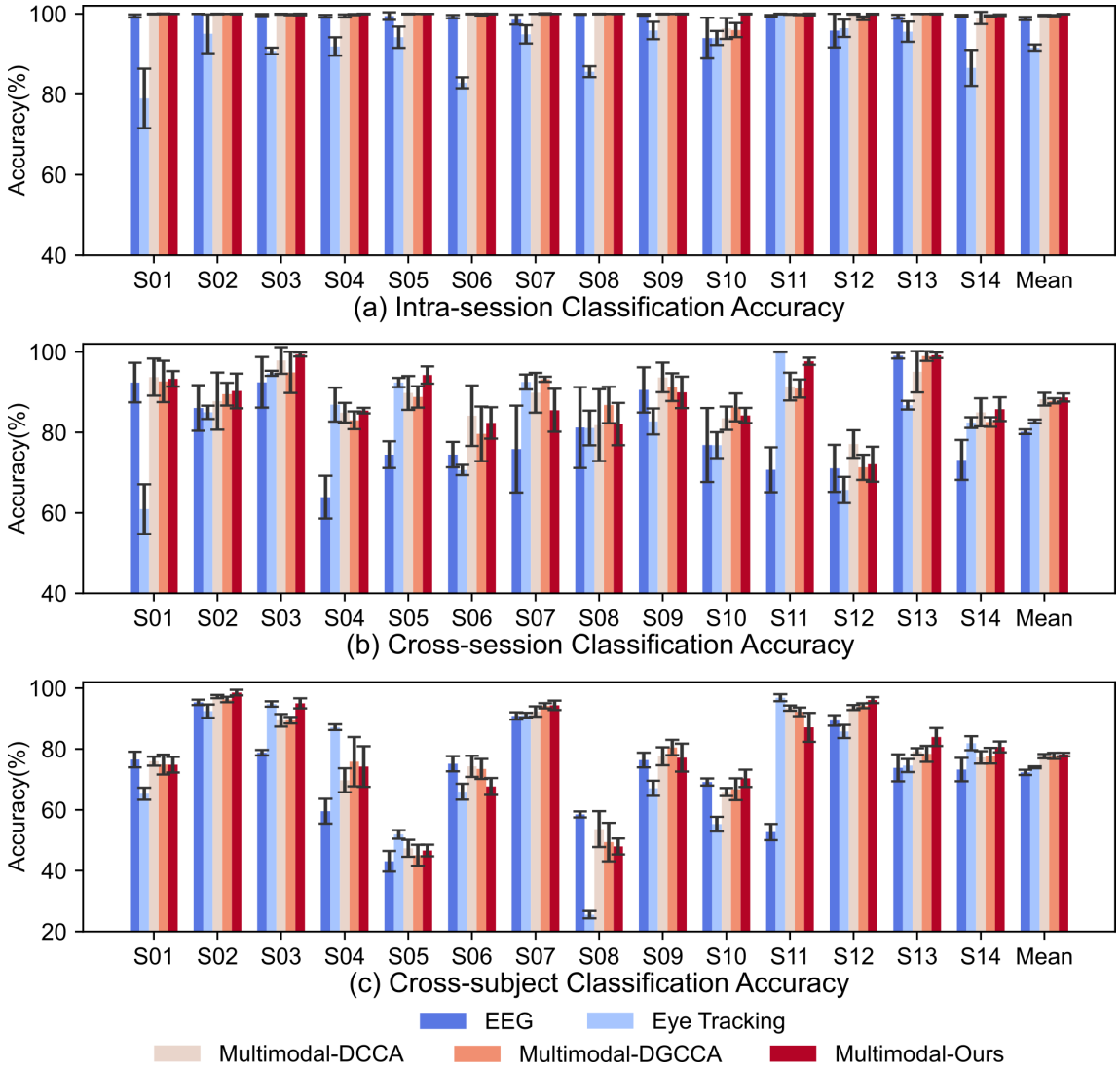


Fig. 6. Classification results of methods in intra-session, cross-session, and cross-subject evaluation tasks. (a) depicts the intra-session results. The average intra-session accuracy of EEG, eye tracking, DCCA multimodal, DGCCA multimodal and our multimodal approaches is 98.84±0.39%, 91.61±0.76%, 99.60±0.16%, 99.72%±,0.15%, and 99.93±0.05%, respectively. (b) depicts the cross-session results. The average cross-session accuracy of the five methods is 80.15±0.58%, 82.75±0.44%, 88.25±1.58%, 87.83%±0.63%, and 88.67±1.00%, respectively. (c) depicts the cross-session results. The average cross-subject accuracy of the five methods is 72.32%±0.86%, 73.98%±0.32%, 77.64%±0.69%, 77.75%±1.00%, and 78.19%±0.56%, respectively. Our multimodal method performs consistently well among the five approaches in three evaluation tasks.

fatigue for all subjects, respectively. The bottom row presents the results of subtracting the power of the fatigue state from that of the alert state. According to the bottom topography, an increase was observed in the prefrontal cortex across all three frequency bands, while a decrease was noted in the other frontal channels.

Fig. 4(c) depicts the average power of 5 regions on different mental states with the difference significance by Wilcoxon test. A non-overlapping 4-second time window was employed to compute the power spectral density (PSD). Subsequently, the PSD values were averaged across brain regions and driving sessions to generate statistical samples. Therefore, each sample provides information about the band power within a specific brain region across a frequency band for a given session. The statistical result revealed a significant increase within the prefrontal cortex across three frequency bands for all sessions. Furthermore, the parietal and occipital lobes exhibited a significant increase in the lower alpha band. A significant decrease was evident in the beta band within the frontal lobe.

2) Visualization and Analysis of Eye-tracking Feature: The eye-tracking features were visualized to investigate the trend of eye-tracking features with subject's increasing driving time and conducted statistical analyses for all eight features.

The trends of 2 eye-tracking features for a subject (Subject 7) throughout the entire 90-minute driving session and the histograms of fixation time in two mental states for all subjects are presented in Fig. 5. Fig. 5(a) illustrates the trend of mean fixation time and saccade mean time for a subject, the mean saccade time exhibits an increasing trend as the time-on-task increases, consistent with findings in [29]. Meanwhile, the mean fixation time demonstrates a decreasing trend. Fig. 5(b) and 5(c) depict histograms of fixation times in the fatigue and alert state, respectively. The fixation times were categorized into three classes: express (<150ms), cognitive

TABLE I
AVERAGE VALUE AND WILCOXON TEST RESULTS FOR 8 EYE-TRACKING FEATURES IN ALERT AND SOBER STATE (N=28)

Feature Name	Alert	Fatigue	P-value
Mean pupil diameter	4.44	4.35	–
Fixation count	87.5	92.12	–
Saccade count	134.85	177.25	**0.007**
Mean fixation duration	615.95	533.1	0.063
Mean saccade duration	58.25	68.38	**0.022**
Peak saccade velocity	158.89	158.48	–
Mean saccade velocity	98.73	96.43	–
Mean saccade amplitude	2.79	2.69	–

(150ms−900ms), and overlong(>900ms) [29]. The counts of express, cognitive, and overlong fixations were 4840, 15873, and 4029 in the alert state, and 7127, 15681, and 3030 in the fatigue state, respectively. The ratio of cognitive fixations decreased from 64.2% to 60.7%, the ratio of overlong fixations decreased from 16.3% to 11.7%, while the ratio of express fixations increased from 19.6% to 27.6%.

The results of the statistical analyses are presented in Table I. Among the 8 eye-tracking features, a significant increase was observed in both the "saccade count" and "saccade mean duration". Additionally, while not statistically significant, there appears to be a decrease in "mean fixation duration" during the fatigue state.

B. Performance of Fatigue Detection

Intra-session, cross-session, and cross-subject evaluation tasks were conducted to assess the performance of methods, respectively. The results are presented in Fig. 6.

1) Intra-session Performance: The results of intra-session classification are presented in Fig. 6(a). The mean accuracy for all subjects in the EEG unimodal, eye tracking unimodal, DCCA multimodal, DGCCA multimodal, and the proposed multimodal methods is 98.84%±0.39%, 91.61%±0.76%, 99.60%±0.16%, 99.72%±,0.15% and 99.93%±0.05%, respectively. The proposed method achieved the highest performance among the five methods (99.93%) with the lowest standard deviation compared to the other methods.

2) Cross-session Performance: The results of cross-session classification are presented in Fig. 6(b). The mean accuracy for the EEG unimodal, eye tracking unimodal, DCCA multimodal, DGCCA multimodal, and the proposed multimodal methods is 80.15%±0.58%, 82.75%±0.44%, 88.25%±1.58%, 87.83%±0.63%, and 88.67%±1.00%, respectively. Our multimodal method achieved the highest accuracy among the five methods (88.67%±1.00%).

3) Cross-subject Performance: Cross-subject evaluation is a more challenging task as it evaluates the model's performance across different subjects. The results are demonstrated in Fig. 6(c). The mean accuracy of 5 runs for the EEG unimodal, eye tracking unimodal, DCCA multimodal, DGCCA multimodal, and the proposed methods is 72.32%±0.86%, 73.98%±0.32%, 77.64%±0.69%, 77.75%±1.00%, and 78.19%±0.56%, respectively. The proposed multimodal method consistently performed well, achieving the highest accuracy among the five approaches in three validation tasks.

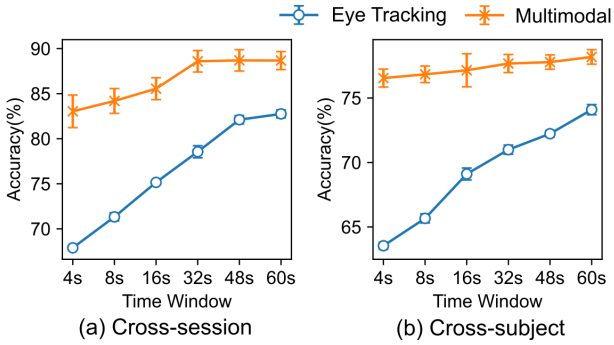


Fig. 7. The performance influence of the eye-tracking time window for both the eye-tracking method and the proposed multimodal method. (a) demonstrates the trend of accuracy as the time window increases in cross-session evaluation task. (b) demonstrates the trend of accuracy in cross-subject evaluation task.

C. Influence of Eye-tracking Time Window

An interesting phenomenon in the selection of the time window for eye-tracking feature generation (Section II-B.2) was found. We examined the relationship between time window length and classification performance by testing various time windows ranging from 4 seconds to 60 seconds in the cross-session and cross-subject classification tasks, with the time window for EEG data consistently set at 1 second. The results are presented in Fig. 7. Both the eye tracking unimodal method and the multimodal method were tested, with each configuration run five times. In the eye-tracking unimodal method, we observed a consistent increase in accuracy in both cross-session(from 67.90% to 82.75%) and cross-subject(from 63.54% to 74.10%) tasks as the time window duration increased. Meanwhile, in the cross-session task, the performance of the multimodal method improved from 4 seconds to 32 seconds (from 83.03% to 88.59%), after which it stabilized between 32 seconds and 60 seconds. In cross-subject task, the accuracy of the proposed multimodal method increased from 76.54% to 78.19%.

D. Ablation Study

The proposed architecture encompasses a cross-modal predictive alignment task, a 1D attention module, and a pretraining procedure for the EEG encoding module. To systematically evaluate different combinations of the proposed modules and the effectiveness of pretraining procedure, ablation studies were conducted. In the ablation experiments, the random seed was consistently set to 42, the division of data and the input order of data remains identical in all configurations for the same validation task.

1) Ablation of Pretraining: The ablation result of pretraining procedure was presented in Table II. With the pretraining

TABLE II
ABLATION RESULT (IN PERCENT) OF PRETRAINING

Method	Cross-session	Cross-subject
w/o pretraining	83.44	76.28
with pretraining	**86.48**	**77.39**

"w/o" represents the abbreviation of "without".

TABLE III
ABLATION RESULT (IN PERCENT) OF ALIGNMENT AND ATTENTION

Methods		Cross-session	Cross-subject
Alignment	Attention		
×	×	81.38	78.64
✓	×	85.17	**80.16**
×	✓	82.34	76.86
✓	✓	**86.48**	77.39

TABLE IV
PERFORMANCE (IN PERCENT) ON SEED-VIG DATASET

Method	Accuracy
Multimodal-DCCA	86.99
Multimodal-DGCCA	87.83
Multimodal-Ours	**96.62**

procedure, the performance of model increase from 83.44% to 86.48% and 76.28% to 77.39% in the cross-session and cross-subject tasks, respectively.

2) Ablation of Alignment and Attention: The ablation results for cross-modal predictive alignment and one-dimensional attention are presented in Table II. In the cross-session task, both the alignment and attention modules improve the model's performance by 3.79% and 0.96%, respectively. Combining both modules yields a 4.62% performance enhancement. However, in the cross-subject task, the adoption of the attention module resulted in a 1.78% performance decrease. Although combining both modules leads to performance loss, utilizing only the alignment module will result in a 1.52% performance improvement.

E. Performance on SEED-VIG Dataset

To assess the performance of the proposed method on the other dataset, cross-subject evaluation experiments using the publicly available SEED-VIG dataset was conducted in this work [24]. The results are summarized in Table IV. The proposed approach achieved an accuracy of 96.92%, outperforming both DCCA-based and DGCCA-based multimodal methods by 9.97% and 9.13%, respectively. These results further underscore the efficacy of the proposed method.

IV. DISCUSSION

In this study, we conducted a series of driving simulation experiments utilizing EEG and eye tracking. Visual and statistical analyses were executed on EEG data in five brain regions, focusing on three typical frequency bands, as shown in Figure 4. Our findings consistently revealed a significant increase in activity within the prefrontal cortex across all three frequency bands. This aligns with prior research [47], given the prefrontal cortex's role in guiding motor and cognitive behaviors over time [48]. Additionally, a significant increase in power activity was observed in the lower alpha bands within the parietal and occipital lobes, a phenomenon in line with previous studies [24]. The elevation in lower alpha bands likely reflects the heightened effort required to maintain alertness [49]. Conversely, a significant decrease was detected in beta band activity within the frontal region, consistent with observations in [3]. This decrease suggests that subjects might be losing focus while driving.

Simultaneously, we conducted a visualization of the trends in two eye-tracking features, the fixation time histograms in two states (Fig. 5), and a statistical analysis for eight eye-tracking features (Table I). The majority of subjects displayed an increasing trend in mean saccade time, in line with previous research [29]. However, the mean fixation time for most subjects decreased for the entire driving process. This decline in fixation time can be attributed to the growing occurrence of eye wandering as the time-on-task increases. Additionally, the histograms of fixation time revealed a reduction in the proportion of cognitive and overlong fixations, coupled with an increase in express fixations. It is important to note that the change in proportion is primarily driven by an increase in the count of express fixations and a decrease in the count of overlong fixations. This phenomenon may suggest that the subjects are gradually losing interest in the driving scene as drowsiness sets in. Meanwhile, the statistical findings presented in Table I indicate that the saccade count and saccade mean duration exhibit significant differences between the alert and fatigue states. These results offer substantial analytical support for the adoption of eye-tracking features.

To address the information limitation of unimodal methods and enhance the performance of driving fatigue detection, we introduced an innovative multimodal architecture, which outperforms the EEG unimodal method, eye-tracking unimodal method, DCCA multimodal method, and DGCCA multimodal method (Fig. 6). Notably, the performance of the multimodal methods surpasses that of the unimodal ones. To delve into how the multimodal features enhance the feature representation in the latent space, Uniform Manifold Approximation and Projection (UMAP) [50]–[52] was employ to visualize the latent features for both multimodal and unimodal models in the cross-subject task, as illustrated in Fig. 8. The points in the visualization represent the feature samples. The points cover almost the entire 2D projection space, indicating the higher complexity of EEG features. On the other hand, although the eye-tracking feature appears relatively simpler compared to EEG, it lacks the compactness required for an effective classification boundary to be established by a classifier. However, when considering the multimodal features, a noticeable improvement in adherence to the principle of maintaining high cohesion within the same class and low coupling between different classes is observed in the 2D projection, particularly in the case of the multimodal method utilizing additive fusion. This suggests the integration of multiple modalities by additive fusion allows for a more robust and discriminative representation. The superior pattern exhibited by additive fusion may arise from its inherent capacity to effectively preserve distinctive features, while avoiding excessive parameters that could result in model overfitting.

Given the intricate pattern of the 2D projection of EEG features in the unimodal method, there is a curious question regarding whether the proposed multimodal method effectively captures EEG features. To investigate which channels significantly contribute to the modeling process [53], we analyzed the parameters of the channel convolution (DepthWiseConv2d

in Fig. 3(b)) kernel within the EEG convolutional network to visualize the learned importance of EEG channels across three validation tasks. The results related to this analysis are presented in Fig. 9, which illustrates that the prefrontal cortex emerges as the most critical area for fatigue detection, a finding that aligns with the observed increase in the prefrontal power during the analysis of drowsy states (Fig. 4). This underscores the efficacy of the proposed EEG convolutional network.

Additionally, we assessed the impact of varying eye-tracking time windows on classification accuracy, examining different window durations for both unimodal and multimodal methods in cross-session and cross-subject tasks (Fig. 7). It's worth noting that while previous studies employed shorter time windows (e.g., 4 seconds in [35] and 10 seconds in [54]) for eye-tracking feature generation, our results demonstrated improved performance with longer time windows. This observation may be attributed to the substantial variability in eye-tracking data within short time windows, influenced by saccades and rapid eye movements occurring in both alert and fatigue states. In contrast, longer time windows could mitigate these effects, leading to more reliable features.

Moreover, we conducted an ablation study on the proposed multimodal method. Table II highlights the efficacy of our pre-training procedure and Table III demonstrates the significant benefits of incorporating the cross-modal predictive alignment module in both cross-session and cross-subject tasks. However, the performance of the attention module increases in the cross-session task, while it appears to diminish in the cross-subject task. This could be attributed to more pronounced distribution differences in the cross-subject data, which may impede the one-dimensional attention module from learning optimal feature weights effectively. In the application, we suggest incorporating cross-modal alignment and one-dimensional attention in cross-session tasks, and only integrating cross-modal alignment in cross-subject tasks.

To gain insight into how the proposed alignment module enhances cross-modal features, we conducted UMAP visualizations to compare the features of the multimodal method with the alignment module against those of the model without alignment and the model incorporating the DCCA module. The results are illustrated in Fig. 10. All three methods effectively distinguish the features of different classes as the training

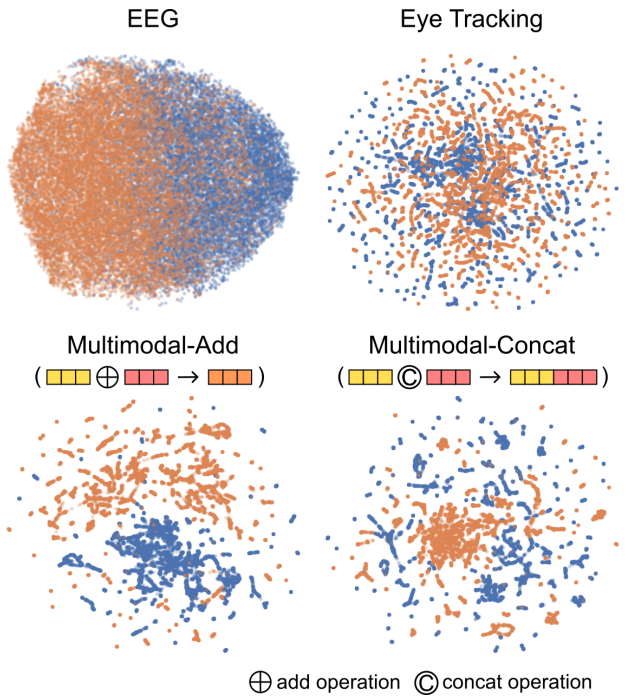


Fig. 8. Feature visualization of unimodal and multimodal methods for all subjects. The visualization represents the features learned by the two unimodal methods, namely "EEG" and "Eye Tracking," as well as the features learned by the proposed multimodal methods using additive fusion and concatenative fusion as fusion methods denoted as "Multimodal-Add" and "Multimodal-Concat," respectively. Notably, the visualization intentionally omits the modules of 1D attention and cross-modal alignment in "Multimodal-Add" and "Multimodal-Concat" to provide a clear and intuitive assessment of the efficacy of adopting multimodal approaches. To avoid the overfitting of models, the data of Subject 2 was used as validation and model selection.

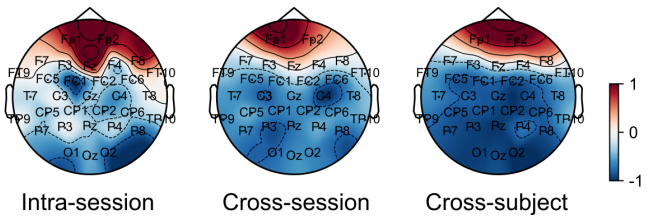


Fig. 9. Channel importance learned through channel convolution within the EEG convolutional network of the proposed multimodal architecture.

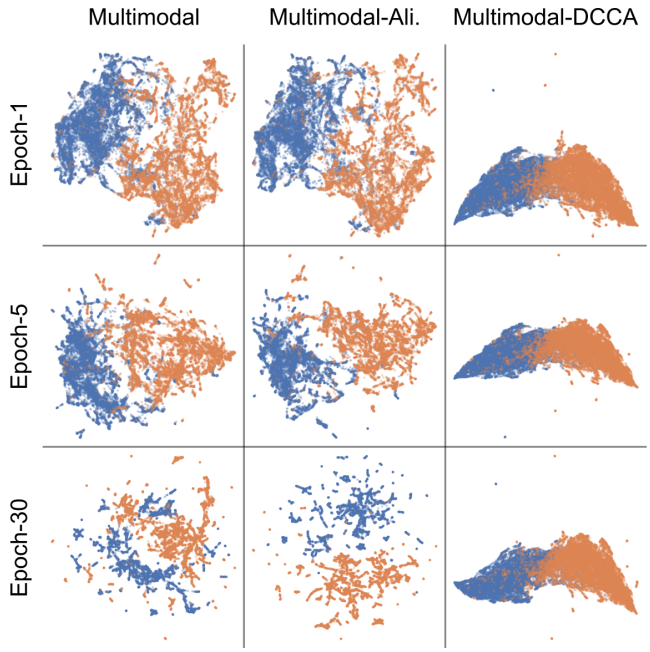


Fig. 10. Impact on features (UMAP) through cross-modal predictive alignment and Deep Canonical Correlation Analysis (DCCA). The visualization encompasses data collected from all subjects. The term "Multimodal" pertains to the proposed multimodal model without cross-modal alignment or DCCA, while "Multimodal-Ali." and "Multimodal-DCCA" refer to the models that incorporate cross-modal alignment and DCCA, respectively. The top, middle, and bottom rows of the visualization correspond to the features extracted from epoch 1, epoch 5, and epoch 30, respectively. Notably, the module of 1D attention was deliberately excluded to facilitate a clear and direct comparison among the methods.

progresses. The key distinction is that the "Multimodal-Ali." method focuses on enhancing feature separability by increasing the average distance between different classes, whereas the "Multimodal-DCCA" method prioritizes preserving the correlation coefficient between modalities, resulting in almost consistent patterns in 2D projection across the first 30 epochs. While preserving inter-modality correlation can be advantageous for certain tasks, it may limit the model's ability to learn highly separable representations. In contrast, the model that incorporates cross-modal alignment does not maintain the correlation in the latent space, it aligns the multimodal features by optimizing the cross-modal predicted similarity, which seems to be more effective in enhancing the representation of multimodal features.

Although the proposed multimodal architecture outperforms other methods, it has several limitations. Firstly, the design of different time windows, although compensating for each other, may require a padding operation for the modality with a longer time window during online applications at the start. Moreover, the multimodal method requires two kinds of devices for data acquisition, which could limit its real-world application. Secondly, the proposed multimodal architecture includes a pretraining stage for the EEG encoding network (which is also adopted in the multimodal comparative method) to enhance detection performance. As a result, our model is somewhat less convenient compared to the end-to-end model. Thirdly, the 1D attention module introduced in this study is employed for 1D features encoded by the backbone networks of different modalities. It's important to note that the weights learned for 1D features may have reduced interpretability compared to those applied to earlier-stage features, such as temporal attention (used for time samples) or spatial attention (used for channels) [55]. Additionally, the performance of the 1D attention module appears to be affected when dealing with data featuring significant distribution differences. In the future, the authors will seek more effective methods for learning highly representative features to address the significant variations present in cross-subject data, and delve deeper into enhancing the interpretability of multimodal models.

V. Conclusion

This study introduced a novel architecture for driving fatigue detection by integrating EEG and eye-tracking modalities, resulting in a hybrid approach. The proposed multimodal neural network architecture consists of a convolutional EEG encoder and an MLP eye feature encoder. During the fusion stage, we proposed a cross-modal predictive task to align features from different modalities and incorporated a one-dimensional attention module to enhance feature representation. The experimental outcomes of this study provide substantial evidence regarding the advantages of utilizing a multimodal approach for fatigue detection. Furthermore, the results demonstrate the superiority of integrating the 1D attention module in the cross-session task and employing the cross-modal predictive alignment in both cross-session and cross-subject tasks. These findings suggest promising potential for the development of improved fatigue detection systems.

References

[1] A. Bener, E. Yildirim, T. Özkan, and T. Lajunen, "Driver sleepiness, fatigue, Careless Behavior and Risk of Motor Vehicle Crash and injury: Population Based Case and Control Study," *Journal of Traffic and Transportation Engineering (English Edition)*, vol. 4, no. 5, pp. 496–502, Oct. 2017, doi: 10.1016/j.jtte.2017.07.005.

[2] T. Xu et al., "E-Key: an EEG-Based Biometric Authentication and Driving Fatigue Detection System," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1–1, 2021, doi: 10.1109/TAFFC.2021.3133443.

[3] B. T. Jap, S. Lal, P. Fischer, and E. Bekiaris, "Using EEG Spectral Components to Assess Algorithms for Detecting Fatigue," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2352–2359, Mar. 2009, doi: 10.1016/j.eswa.2007.12.043.

[4] H. Wang et al., "Dynamic Reorganization of Functional Connectivity Unmasks Fatigue Related Performance Declines in Simulated Driving," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 8, pp. 1790–1799, Aug. 2020, doi: 10.1109/tnsre.2020.2999599.

[5] C. Chen, Z. Ji, Y. Sun, Anastasios Bezerianos, N. Thakor, and H. Wang, "Self-Attentive Channel-Connectivity Capsule Network for EEG-Based Driving Fatigue Detection," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 3152–3162, Jan. 2023, doi: 10.1109/tnsre.2023.3299156.

[6] H. Wang, C. Wu, T. Li, Y. He, P. Chen, and A. Bezerianos, "Driving Fatigue Classification Based on Fusion Entropy Analysis Combining EOG and EEG," *IEEE Access*, vol. 7, pp. 61975–61986, 2019, doi: 10.1109/access.2019.2915533.

[7] H. Wang, L. Xu, Anastasios Bezerianos, C. Chen, and Z. Zhang, "Linking Attention-Based Multiscale CNN with Dynamical GCN for Driving Fatigue Detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, Jan. 2021, doi: 10.1109/tim.2020.3047502.

[8] H. Wang et al., "Driving Fatigue Recognition with Functional Connectivity Based on Phase Synchronization," *IEEE transactions on cognitive and developmental systems*, vol. 13, no. 3, pp. 668–678, Sep. 2021, doi: 10.1109/tcds.2020.2985539.

[9] H. Wang, A. Dragomir, Nida Itrat Abbasi, J. Li, N. V. Thakor, and Anastasios Bezerianos, "A Novel real-time Driving Fatigue Detection System Based on Wireless Dry EEG," *Cognitive Neurodynamics*, vol. 12, no. 4, pp. 365–376, Feb. 2018, doi: 10.1007/s11571-018-9481-5.

[10] M. Kolodziej et al., "Fatigue Detection Caused by Office Work with the Use of EOG Signal," *IEEE Sensors Journal*, vol. 20, no. 24, pp. 15213–15223, Dec. 2020, doi: 10.1109/jsen.2020.3012404.

[11] D. Sommer and M. Golz, "Evaluation of PERCLOS Based Current Fatigue Monitoring Technologies," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, 2010, pp. 4456–4459. doi: 10.1109/iembs.2010.5625960.

[12] J. Xu, J. Min, and J. Hu, "Real-time Eye Tracking for the Assessment of Driver Fatigue," *Healthcare Technology Letters*, vol. 5, no. 2, pp. 54–58, Apr. 2018, doi: 10.1049/htl.2017.0020.

[13] X. Fan, B.-C. Yin, and Y.-F. Sun, "Yawning Detection for Monitoring Driver Fatigue," in *2007 International Conference on Machine Learning and Cybernetics*, 2007, pp. 664–668. doi: 10.1109/ICMLC.2007.4370228.

[14] Z. Zhang, H. Ning, and F. Zhou, "A Systematic Survey of Driving Fatigue Monitoring," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 19999–20020, Nov. 2022, doi: 10.1109/tits.2022.3189346.

[15] Z. Wan, R. Yang, M. Huang, N. Zeng, and X. Liu, "A Review on Transfer Learning in EEG Signal Analysis," *Neurocomputing*, vol. 421, no. 421, pp. 1–14, Jan. 2021, doi: 10.1016/j.neucom.2020.09.017.

[16] Y. Zhao et al., "Label-based alignment multi-source domain adaptation for cross-subject EEG fatigue mental state evaluation," Front. Hum. Neurosci., vol. 15, pp. 1-16, Oct. 2021, doi: 10.3389/fnhum.2021.706270.

[17] B.-Q. Ma, H. Li, Y. Luo, and B.-L. Lu, "Depersonalized Cross-Subject Vigilance Estimation with Adversarial Domain Generalization," in *2019 International Joint Conference on Neural Networks (IJCNN)*, 2019, pp. 1–8. doi: 10.1109/ijcnn.2019.8852347.

[18] K. Wang et al., "Vigilance Estimating in SSVEP-Based BCI Using Multimodal Signals," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine* & Biology Society (EMBC), 2021, pp. 5974–5978. doi: 10.1109/embc46164.2021.9629736.

[19] Ivaylo Ivaylov, Milena Lazarova, and Agata Manolova, "Multimodal Motor Imagery BCI Based on EEG and NIRS," in *2021 56th International Scientific Conference on Information*, Communication and Energy Systems and Technologies (ICEST), 2021, pp. 73–76. doi: 10.1109/icest52640.2021.9483551.

[20] Q. He, L. Feng, G. Jiang, and P. Xie, "Multimodal Multitask Neural Network for Motor Imagery Classification with EEG and fNIRS Signals," *IEEE Sensors Journal*, vol. 22, no. 21, pp. 20695–20706, Nov. 2022, doi: 10.1109/jsen.2022.3205956.

[21] A. Mohammadian, V. Abootalebi, M. H. Moradi, and M.A. Khalilzadeh, "Multimodal Detection of Deception Using Fusion of Reaction Time and P300 Component," in *2008 Cairo International Biomedical Engineering Conference*, 2008, pp. 1–4. doi: 10.1109/cibec.2008.4786064.

[22] S. J. Colloby et al., "Multimodal EEG-MRI in the Differential Diagnosis of Alzheimer's Disease and Dementia with Lewy Bodies," *Journal of Psychiatric Research*, vol. 78, pp. 48–55, Jul. 2016, doi: 10.1016/j.jpsychires.2016.03.010.

[23] H. Cai, Z. Qu, Z. Li, Y. Zhang, X. Hu, and B. Hu, "Feature-level Fusion Approaches Based on Multimodal EEG Data for Depression Recognition," *Information Fusion*, vol. 59, pp. 127–138, Jul. 2020, doi: 10.1016/j.inffus.2020.01.008.

[24] W.-L. Zheng and B.-L. Lu, "A Multimodal Approach to Estimating Vigilance Using EEG and Forehead EOG," *Journal of Neural Engineering*, vol. 14, no. 2, p. 026017, Feb. 2017, doi: 10.1088/1741-2552/aa5a98.

[25] J. Pan, X. Cai, D. Mo, Y. Yu, and Y. Li, "Residual Attention Capsule Network for Multimodal EEG- and EOG-Based Driver Vigilance Estimation," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–12, Jan. 2023, doi: 10.1109/tim.2023.3307756.

[26] J. Shi and K. Wang, "Fatigue Driving Detection Method Based on Time-Space-Frequency Features of Multimodal Signals," *Biomedical Signal Processing and Control*, vol. 84, p. 104744, Jul. 2023, doi: 10.1016/j.bspc.2023.104744.

[27] A. Nemcova et al., "Multimodal Features for Detection of Driver Stress and Fatigue: Review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 6, pp. 3214–3233, Jun. 2021, doi: 10.1109/tits.2020.2977762.

[28] G. Du, L. Zhang, K. Su, X. Wang, S. Teng, and P. X. Liu, "A Multimodal Fusion Fatigue Driving Detection Method Based on Heart Rate and PERCLOS," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 21810–21820, Nov. 2022, doi: 10.1109/TITS.2022.3176973.

[29] R. Schleicher, N. Galley, S. Briest, and L. Galley, "Blinks and Saccades as Indicators of Fatigue in Sleepiness warnings: Looking tired?," *Ergonomics*, vol. 51, no. 7, pp. 982–1010, Jun. 2008, doi: 10.1080/00140130701817062.

[30] L. L. Di, R. S. Renner, A. Catena, J. J. Cañas, B. M. Velichkovsky, and S. Pannasch, "Towards a Driver Fatigue Test Based on the Saccadic Main sequence: A Partial Validation by Subjective Report Data," *Transportation Research Part C-emerging Technologies*, vol. 21, no. 1, pp. 122–133, Apr. 2012, doi: 10.1016/j.trc.2011.07.002.

[31] H. Li, W.-L. Zheng, and B.-L. Lu, "Multimodal Vigilance Estimation with Adversarial Domain Adaptation Networks," in *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018. doi: 10.1109/ijcnn.2018.8489212.

[32] W. Wu et al., "Multimodal Vigilance Estimation Using Deep Learning," *IEEE Transactions on Cybernetics*, vol. 52, no. 5, pp. 3097–3110, May 2022, doi: 10.1109/tcyb.2020.3022647.

[33] Jitender Singh Virk, M. Singh, M. Singh, U. Panjwani, and K. Ray, "A Multimodal Feature Fusion Framework for Sleep-Deprived Fatigue Detection to Prevent Accidents," *Sensors*, vol. 23, no. 8, pp. 4129–4129, Apr. 2023, doi: 10.3390/s23084129.

[34] X. Zhang et al., "Fusing of Electroencephalogram and Eye Movement with Group Sparse Canonical Correlation Analysis for Anxiety Detection," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 1–1, 2020, doi: 10.1109/taffc.2020.2981440.

[35] W. Liu, W.-L. Zheng, Z. Li, S.-Y. Wu, L. Gan, and B.-L. Lu, "Identifying Similarities and Differences in Emotion Recognition with EEG and Eye Movements among Chinese, German, and French People," *Journal of Neural Engineering*, vol. 19, no. 2, pp. 026012–026012, Mar. 2022, doi: 10.1088/1741-2552/ac5c8d.

[36] Y. Lan, W. Liu, and B.-L. Lu, "Multimodal Emotion Recognition Using Deep Generalized Canonical Correlation Analysis with an Attention Mechanism," in *International Joint Conference on Neural Network*, 2020. doi: 10.1109/ijcnn48605.2020.9207625.

[37] A. Delorme and S. Makeig, "EEGLAB: An Open Source Toolbox for Analysis of single-trial EEG Dynamics Including Independent Component Analysis," *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9–21, Mar. 2004, doi: 10.1016/j.jneumeth.2003.10.009.

[38] The Tobii I-VT Fixation Filter, Tobii Technology, 2012, pp. 4–19.

[39] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A Compact Convolutional Neural Network for EEG-based Brain–computer Interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, p. 056013, Jul. 2018, doi: 10.1088/1741-2552/aace8c.

[40] X. Chen and K. He, "Exploring Simple Siamese Representation Learning," in *Computer Vision and Pattern Recognition*, 2021, pp. 15750–15758. doi: 10.1109/cvpr46437.2021.01549.

[41] J. Shen, S. Zhao, Y. Yao, Y. Wang, and L. Feng, "A Novel Depression Detection Method Based on Pervasive EEG and EEG Splitting Criterion," in *IEEE International Conference on Bioinformatics and Biomedicine*, 2017. doi: 10.1109/bibm.2017.8217946.

[42] J. Shen, J. Xu, B. Hu, G. Wang, and Z. Ding, "An Improved Empirical Mode Decomposition of Electroencephalogram Signals for Depression Detection," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 262–271, Jan. 2022, doi: 10.1109/taffc.2019.2934412.

[43] J. Shen et al., "Exploring the Intrinsic Features of EEG Signals via Empirical Mode Decomposition for Depression Recognition," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 356–365, Jan. 2023, doi: 10.1109/tnsre.2022.3221962.

[44] J. Shen et al., "Depression Recognition from EEG Signals Using an Adaptive Channel Fusion Method via Improved Focal Loss," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 7, pp. 3234–3245, Jul. 2023, doi: 10.1109/jbhi.2023.3265805.

[45] J. Shen et al., "A Novel Intelligence Evaluation Framework: Exploring the Psychophysiological Patterns of Gifted Students," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 2, pp. 1–10, Jan. 2023, doi: 10.1109/tcss.2023.3303331.

[46] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep Canonical Correlation Analysis," in *International Conference on Machine Learning*, 2013, pp. 1247–1255. doi: 10.5555/3042817.3043076.

[47] H. J. Foy and P. Chapman, "Mental Workload Is Reflected in Driver behaviour, physiology, Eye Movements and Prefrontal Cortex Activation," *Applied Ergonomics*, vol. 73, pp. 90–99, Nov. 2018, doi: 10.1016/j.apergo.2018.06.006.

[48] B. Kolb, R. Mychasiuk, A. Muhammad, Y. Li, D. O. Frost, and R. Gibb, "Experience and the Developing Prefrontal Cortex," *Proceedings of the National Academy of Sciences*, vol. 109, no. Suppl 2, pp. 17186–17193, Oct. 2012, doi: 10.1073/pnas.1121251109.

[49] W. Klimesch, "EEG Alpha and Theta Oscillations Reflect Cognitive and Memory performance: A Review and Analysis," Brain Research Reviews, vol. 29, no. 2–3, pp. 169–195, Apr. 1999, doi: 10.1016/s0165-0173(98)00056-3.

[50] A. Diaz-Papkovich, L. Anderson-Trocmé, C. Ben-Eghan, and S. Gravel, "UMAP Reveals Cryptic Population Structure and Phenotype Heterogeneity in Large Genomic Cohorts," *PLOS Genetics*, vol. 15, no. 11, p. e1008432, Nov. 2019, doi: 10.1371/journal.pgen.1008432.

[51] D. Yang, V. Wei, Z. Jin, Z. Yang, and X. Chen, "A UMAP-based Clustering Method for multi-scale Damage Analysis of Laminates," Applied Mathematical Modelling, vol. 111, pp. 78–93, Nov. 2022, doi: 10.1016/j.apm.2022.06.017.

[52] K. Chen, Z. Liu, Q. Liu, Q. Ai, and L. Ma, "EEG-based Mental Fatigue Detection Using Linear Prediction Cepstral Coefficients and Riemann Spatial Covariance Matrix," *Journal of Neural Engineering*, vol. 19, no. 6, p. 066021, Nov. 2022, doi: 10.1088/1741-2552/aca1e2.

[53] J. Shen et al., "An Optimal Channel Selection for EEG-Based Depression Detection via Kernel-Target Alignment," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2545–2556, Jul. 2021, doi: 10.1109/jbhi.2020.3045718.

[54] J. Kang, X. Han, J. Song, Z. Niu, and X. Li, "The Identification of Children with Autism Spectrum Disorder by SVM Approach on EEG and eye-tracking Data," *Computers in Biology and Medicine*, vol. 120, p. 103722, May 2020, doi: 10.1016/j.compbiomed.2020.103722.

[55] Z. Wang, C. Chen, J. Li, T. Rades, Y. Sun, and H. Wang, "ST-CapsNet: Linking Spatial and Temporal Attention with Capsule Network for P300 Detection Improvement," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 991–1000, Jan. 2023, doi: 10.1109/tnsre.2023.3237319.