aerospace

MDPI

Article

Remote Tower Air Traffic Controller Fatigue Detection Based on Eye-Tracking and EEG Fusion

Dajiang Song [1], Weijun Pan [2,*], Zirui Yin [3], Boyuan Han [4] and Huafei Gao [5]

[1] CAAC Academy of Flight Technology and Safety, Civil Aviation Flight University of China, Guanghan 618307, China; 17726549663@163.com
[2] Key Laboratory of Flight Techniques and Flight Safety, Civil Aviation Flight University of China, Guanghan 618307, China
[3] School of Transportation and Logistics, Southwest Jiaotong University, Chengdu 611756, China; yin@my.swjtu.edu.cn
[4] Suining Flight College, Civil Aviation Flight University of China, Guanghan 618307, China; hanregulu@gmail.com
[5] College of Air Traffic Management, Civil Aviation Flight University of China, Guanghan 618307, China; doukeyi508@gmail.com
* Correspondence: wjpan@cafuc.edu.cn

Abstract

Remote tower operations require air traffic controllers to maintain continuous visual monitoring and integrate information from panoramic displays, radar data, flight strips, and voice communication. Such screen-mediated and sustained surveillance tasks may lead to covert fatigue, which is difficult to capture using a single physiological or behavioral signal. To address this issue, this study proposes a Gated EEG–Eye Fusion Network (GEEF-Net) for window-level fatigue detection in remote tower controllers. EEG and eye-tracking signals were synchronously collected during simulated remote tower tasks and segmented into 5 s windows with a 2 s step. For each window, 53 EEG features and 47 eye-tracking features were extracted to construct a 100-dimensional multimodal representation. GEEF-Net adopts a lightweight modality-gating mechanism to adaptively weight EEG and eye-tracking representations before fatigue classification. Under the main subject-dependent validation setting, GEEF-Net achieved an Accuracy of 0.883, an F1-score of 0.788, and a ROC-AUC of 0.944, outperforming EEG-only, eye-only, and early-fusion baselines in most overall metrics. The gating analysis indicated that eye-tracking features received a higher average weight than EEG features, suggesting the importance of visual behavior in remote tower fatigue detection. Cross-subject validation showed that individual differences remain a major challenge, while few-shot subject-specific calibration improved model adaptation when limited target-subject samples were available. These findings suggest that EEG–eye-tracking fusion with lightweight modality gating is a feasible approach for fatigue detection in simulated remote tower tasks. However, larger datasets and operationally realistic validation considering shift work, circadian effects, and operational pressure are still required before the approach can be considered operationally reliable.

Keywords: remote tower; air traffic controller; fatigue detection; EEG; eye tracking; multimodal fusion; modality gating; subject-specific calibration

Check for updates

Academic Editor: Miroslav Kelemen

Received: 9 May 2026
Revised: 9 June 2026
Accepted: 10 June 2026
Published: 12 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

In remote tower air traffic management, controllers are required to continuously monitor runways, taxiways, aprons, and surrounding aircraft through multi-screen display

systems, while integrating radar information, flight plan data, and radio communication to support real-time operational decision-making. Compared with conventional tower operations, remote tower control relies more heavily on screen-mediated visual information and sustained attentional regulation. Prolonged visual search and continuous situation monitoring may lead to covert fatigue, increasing the risk of attentional lapses, delayed judgment, and missed information. Previous studies on simulated remote tower operations and air traffic controller fatigue have shown that task duration is associated with changes in eye movement behavior, attentional distribution, and cognitive workload [1,2].

Objective fatigue recognition has increasingly relied on physiological and behavioral signals such as electroencephalography (EEG), eye tracking, and heart rate variability. EEG can reflect neural oscillations, vigilance level, and fatigue-related spectral changes, and has been widely used in driving, flight simulation, and cognitive workload assessment [3,4]. Eye-tracking indicators, including pupil diameter, fixation duration, and saccadic behavior, can capture attentional variation and visual workload changes during sustained visual tasks [2,5]. However, single-modality models are often affected by signal noise, inter-individual variability, and threshold dependence, which limits their robustness in cross-subject applications [3,6].

To improve fatigue detection robustness, multimodal fusion has become an important research direction. Existing studies have attempted to combine EEG, eye tracking, electrocardiography, and other physiological or behavioral signals in driving fatigue, sleep deprivation, and air traffic control fatigue detection tasks [7–9]. These studies suggest that multimodal information can compensate for the limitations of individual modalities. Nevertheless, their applicability to remote tower fatigue detection remains insufficiently examined, particularly under conditions involving sustained visual monitoring, limited sample sizes, and strong individual differences.

Despite the progress in multimodal fatigue detection, three key limitations remain. First, many existing methods rely on direct feature concatenation or fixed fusion strategies, making it difficult to capture state-dependent changes in the relative contributions of EEG and eye-tracking modalities. Second, complex fusion models based on attention mechanisms or high-order cross-modal interactions may improve representation learning, but they can also increase overfitting risk and reduce interpretability in small-sample human-factor datasets. Third, EEG and eye-tracking signals exhibit strong inter-individual variability; therefore, strong performance under a subject-dependent setting does not necessarily imply stable generalization to unseen controllers. For remote tower operations, cross-subject generalization and subject-specific adaptation remain major barriers to practical deployment [1,6,10,11]. Although conventional tower workload and airport-capacity studies provide an important reference, the present study focuses on fatigue detection in a remote tower simulation environment, where visual information is mediated by digital displays rather than direct out-of-window observation.

To address these limitations, this study proposes a lightweight modality-gated fusion network, termed the Gated EEG–Eye Fusion Network (GEEF-Net). The proposed model uses modality-specific encoders to learn EEG and eye-tracking representations separately, and then employs a gating-based weighting mechanism to adaptively adjust the contribution of each modality at the window level. Compared with direct feature concatenation or complex interaction-based fusion structures, GEEF-Net emphasizes sample-level modality weighting, structural compactness, and interpretability, which are particularly important for remote tower fatigue detection with limited samples, signal noise, and pronounced individual differences.

The overall research workflow is illustrated in Figure 1. It includes remote tower task simulation and multimodal data acquisition, EEG and eye-tracking preprocessing, window-

level feature construction, GEEF-Net-based fusion modeling, and multi-level validation and interpretation, including subject-dependent evaluation, strict cross-subject validation, and few-shot subject-specific calibration.

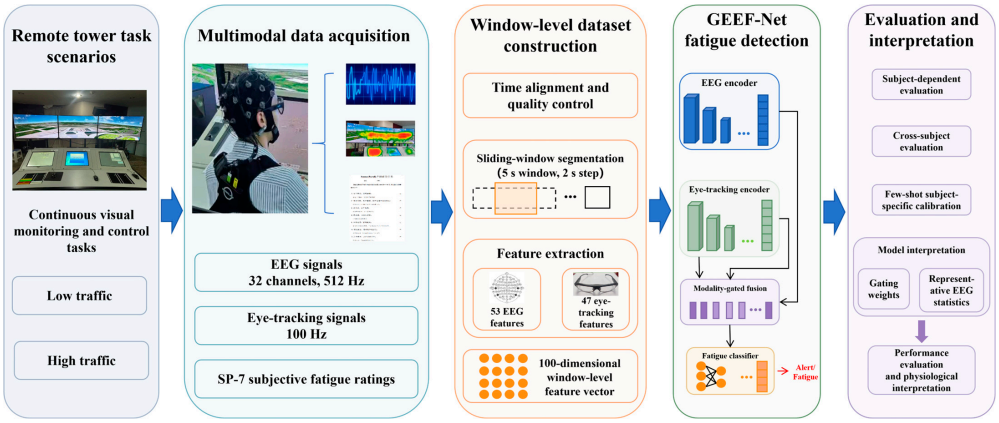

Figure 1. Overall researchworkflow of the proposed fatigue detection framework for remote tower air traffic controllers. The workflow includes remote tower task scenarios, synchronized EEG and eye-tracking data acquisition, window-level dataset construction, GEEF-Net-based fatigue detection, and model evaluation and interpretation.

The main contributions of this study are as follows:

(1) A synchronized EEG–eye-tracking dataset was collected during simulated remote tower tasks, and window-level fatigue labels were constructed for multimodal fatigue recognition.
(2) A lightweight modality-gated fusion network, GEEF-Net, was proposed to capture complementary information between EEG and eye-tracking features through sample-level adaptive weighting.
(3) A multi-level validation framework was designed, including subject-dependent evaluation, strict cross-subject validation, and few-shot subject-specific calibration, to assess both model performance and subject-specific adaptability.
(4) Modality-gating weight analysis and record-level EEG statistical analysis were conducted to support model interpretation and fatigue-related physiological representation.

2. Related Work

Fatigue monitoring for air traffic controllers (ATCOs) has become an important topic in aviation human factors and safety research. Unlike ordinary driving or general surveillance tasks, air traffic control is characterized by long task duration, limited tolerance for error, rapid information updates, and high operational responsibility. EASA has examined ATCO fatigue from the perspectives of regulatory implementation, fatigue risk management, and operational organization, indicating that controller fatigue is not merely an individual physiological state but a human-factor risk closely related to operational safety and organizational management [12]. In remote tower operations, direct out-of-window visual cues are replaced by screen-mediated information from panoramic displays, radar, electronic flight strips, and voice communication, making visual search, sustained attention, and multisource information integration central to fatigue accumulation.

In addition to fatigue detection, air traffic controller workload has long been investigated in conventional tower and airport-capacity studies. In complex conventional tower environments, especially airports with multiple runways and high surface-traffic complexity, controller workload may reach levels comparable to those observed in remote tower monitoring tasks and can directly constrain airport capacity because aircraft movements

must be managed while maintaining acceptable delay and safety margins [13]. Therefore, conventional tower workload and capacity studies provide an important benchmark for interpreting remote tower fatigue detection results. Conventional and remote tower operations share common workload drivers, such as traffic volume, runway configuration, communication load, conflict monitoring, and ground-movement complexity. However, remote tower operations differ because direct out-of-window observation is replaced by camera-based panoramic displays and interface-mediated information. Previous studies on remote tower complexity have indicated that controller workload quantification becomes more critical in multiple remote tower settings, where one controller may need to monitor more than one aerodrome and manage task switching across digital displays [14]. Human-performance assessment of multiple remote tower operations has also shown that, compared with physical tower operations, multiple remote tower operations may significantly affect controllers' mental demand, temporal demand, effort, and frustration [15]. These studies suggest that conventional tower workload and capacity research provides an important reference for remote tower fatigue detection, but its findings cannot be directly transferred without considering the screen-mediated and multi-airport characteristics of remote tower operations. Recent studies have begun to investigate objective fatigue recognition in ATCO and remote tower contexts using physiological and behavioral signals. Yu et al. proposed the RecMF framework, which integrates EEG and eye-tracking information through an attention-enabled CNN-LSTM network for ATCO mental fatigue recognition [1]. Yin et al. further proposed RTFnet for fatigue detection in remote tower controllers using multimodal physiological data fusion [6]. From a single-modality perspective, Hu et al. showed that eye movement indicators such as fixations and saccades are relevant to ATCO fatigue detection [5], while Zhong et al. investigated EEG-based fatigue detection for remote tower controllers using spatio-temporal modeling [16]. These studies provide direct evidence for the feasibility of physiological and behavioral fatigue detection in air traffic control tasks. However, most of them primarily focus on overall classification performance, while the relative contributions of EEG and eye-tracking modalities under different fatigue states, traffic conditions, and individual characteristics remain insufficiently explored.

EEG is one of the most widely used physiological signals for fatigue and drowsiness detection because of its sensitivity to neural oscillations, vigilance changes, and cognitive workload. Othmani et al. reviewed neural-network-based approaches for EEG-based fatigue and drowsiness detection, showing a methodological shift from traditional spectral features and shallow classifiers toward deep neural networks, spatio-temporal learning, and attention-based modeling [3]. Hussein et al. further emphasized that EEG-based fatigue detection remains affected by noise, inter-individual variability, and limited cross-subject generalization [17]. In aviation-related tasks, Hamann and Carstengerdes demonstrated that mental fatigue development during simulated flight can be tracked using neurophysiological measurements [4]. Recent EEG deep learning models, such as SFT-Net, also indicate that fatigue-related information may be distributed across temporal, spectral, and spatial dimensions rather than confined to a single band or channel [18]. In addition, recent EEG fatigue-recognition studies have explored critical-channel identification and optimal feature-subset selection to reduce redundant EEG information and improve fatigue-related representation learning [19]. These findings support the use of EEG for fatigue detection in remote tower tasks, while also highlighting the limitations of EEG-only modeling under strong individual variability.

Eye-tracking signals provide a more direct reflection of visual monitoring behavior, which is particularly relevant to remote tower operations. Eye-tracking features, including pupil-related measures, fixations, saccades, and gaze dynamics, have been shown to reflect

visual fatigue and attentional changes [2,5]. Lian et al. demonstrated that hybrid EEG and eye-tracking fusion can improve fatigue detection compared with single-modality modeling, indicating the complementary value of neural and visual-behavioral information [7]. Multimodal fatigue studies using additional behavioral and physiological cues have also shown that multi-source fusion can improve the robustness of fatigue recognition [8]. Moreover, Vortmann et al. compared early, middle, and late fusion strategies for EEG and eye-tracking features, suggesting that the fusion stage influences how models exploit multimodal information [9]. These studies support EEG–eye-tracking fusion, but they also indicate that the optimal fusion strategy should be determined according to task characteristics, data scale, and interpretability requirements.

From a methodological perspective, multimodal fusion is not equivalent to simple feature concatenation. Deep multimodal fusion methods include encoder-based structures, attention mechanisms, graph neural networks, constrained fusion, and Transformer-based cross-modal interaction models [20,21]. These structures can enhance representation learning, but they often require larger datasets, higher computational resources, and more stable training conditions. This issue is particularly important in human-factor experiments, where EEG and eye-tracking data are usually limited in sample size, affected by signal noise, and characterized by strong individual differences. Studies on EEG–eye-movement fusion have also suggested that fusion methods should balance feature representation and interpretability rather than simply increasing model complexity [22]. Therefore, a lightweight and interpretable fusion mechanism that can dynamically adjust modality contributions may be more suitable for remote tower fatigue detection than highly complex interaction-based architectures.

Cross-subject generalization remains a critical bottleneck for applying physiological fatigue detection models in practice. EEG and eye-tracking signals are strongly subject-dependent, and differences in physiological baselines, fatigue sensitivity, visual search strategies, and task experience can influence feature distributions. Reviews on transfer learning in EEG analysis have noted that EEG data are affected by non-stationarity and inter-individual variability, making target-domain adaptation and limited calibration important strategies for improving generalization [10,11]. For remote tower fatigue detection, reporting only subject-dependent performance is therefore insufficient for assessing deployability. It is necessary to examine model performance on unseen target subjects and to evaluate whether a small amount of subject-specific calibration data can improve adaptation.

Overall, existing studies have demonstrated the value of EEG, eye tracking, and multimodal fusion for fatigue detection, but several gaps remain for remote tower tasks characterized by sustained visual monitoring. First, existing ATCO and remote tower studies have mainly emphasized classification performance, with limited explanation of dynamic modality contributions. Second, many multimodal fatigue detection methods rely on fixed feature concatenation or complex interaction structures, making it difficult to balance performance, interpretability, and deployment feasibility under small-sample human-factor conditions. Third, cross-subject generalization and subject-specific calibration have not been sufficiently integrated into a unified evaluation framework. To address these gaps, this study proposes GEEF-Net and evaluates it under subject-dependent, strict cross-subject, paired-subject, and few-shot calibration settings.

3. Materials and Methods

3.1. Experimental Design and Data Acquisition

Participants were recruited from the School of Air Traffic Management, Civil Aviation Flight University of China. The original experiment included 36 participants, consisting of 6 senior air traffic control instructors and 30 air traffic control trainees. The instructors had

extensive tower control experience, while the trainees had completed relevant air traffic control courses and possessed basic tower control competence. All participants satisfied the Civil Aviation Administration of China Class I medical certificate requirements, had normal or corrected-to-normal vision, and reported no visual dysfunction that could affect task performance or eye-tracking data acquisition.

The study protocol was approved by the Ethics Committee of Civil Aviation Flight University of China. Before the experiment, all participants were informed of the study purpose, procedures, potential risks, and data usage, and written informed consent was obtained. All data were recorded and processed anonymously, and participants were allowed to withdraw at any time.

The experiment was conducted using the Tower Client simulation platform to construct a high-fidelity remote tower operational environment, as shown in Figure 2. The simulated airport was a virtual "Hansha Airport", whose layout and operational characteristics were designed with reference to Wuhan Tianhe International Airport and Changsha Huanghua International Airport. Meteorological conditions were uniformly set as clear daytime conditions to ensure scenario comparability. The platform presented runways, taxiways, aprons, parking stands, and aircraft movement states through three synchronized displays, approximating the out-of-window view of a conventional tower. Participants performed typical tower control tasks, including pushback approval, taxi route planning, takeoff clearance, and landing clearance, using a keyboard, mouse, and voice communication system. To improve ecological validity, a researcher acted as a pseudo-pilot in a separate room and communicated with the participant in real time via radiotelephony.

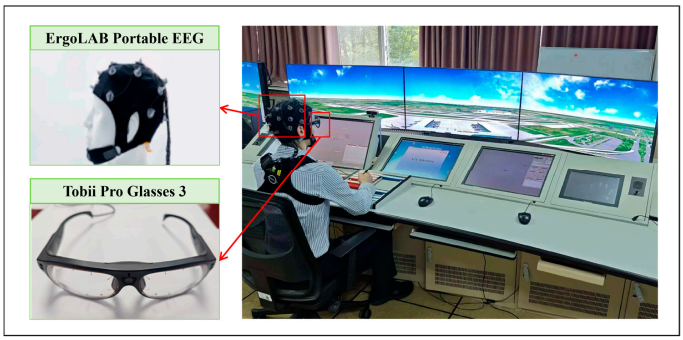

Figure 2. Experimental setup for remote tower control simulation and physiological signal acquisition. The left panels present the ErgoLAB Portable EEG system and Tobii Pro Glasses 3 used for EEG and eye-tracking recording, respectively, and the right panel shows the participant performing the task in the remote tower simulation platform.

Two traffic-flow scenarios were designed. The low-traffic scenario included three inbound and four outbound aircraft, with no more than two aircraft on the taxiways at any given time. The high-traffic scenario included eight inbound and eight outbound aircraft, with up to five aircraft simultaneously appearing on the taxiways during peak periods. This setting increased ground conflict probability and coordination demands, thereby simulating a high-workload remote tower control task. Flight sequences and flight numbers were randomly generated within each scenario type to reduce order effects. The two scenarios also differed in task duration: the low-traffic scenario lasted approximately 20 min, whereas the high-traffic scenario lasted approximately 40 min. Therefore, the traffic-flow condition should be interpreted as a scenario-level workload manipulation involving aircraft number, task duration, and coordination complexity, rather than as an isolated manipulation of traffic volume alone.

Before the formal experiment, participants received standardized instructions and completed a practice session. Approximately 5 min before the task, the physiological

devices were fitted, calibrated, and checked for signal quality; data from this preparation stage were not included in the analysis. During the experiment, participants maintained a standard control posture and completed control instruction and conflict coordination tasks according to the simulated operational information. After each scenario, participants completed the Samn–Perelli 7-point fatigue scale to record subjective fatigue level.

Eye-tracking data, EEG data, and subjective fatigue ratings were collected synchronously. Eye-tracking data were recorded using Tobii Pro Glasses 3 eye-tracking glasses (Tobii AB, Danderyd, Sweden) at 100 Hz, while EEG data were collected using the Ergo-LAB Portable EEG system with 32 channels (Kingfar International Inc., Beijing, China) at 512 Hz. EEG and eye-tracking data were used as the main inputs for the fatigue recognition model, and subjective ratings were used for fatigue-state labeling and result interpretation. After temporal synchronization, signal quality control, and validity screening, synchronized EEG–eye-tracking data from 30 participants were retained for subsequent modeling and analysis.

3.2. Window-Level Dataset Construction and Feature Extraction

To construct a synchronized EEG–eye-tracking dataset suitable for fatigue-state recognition, EEG signals, eye-tracking data, and experimental labels were first temporally aligned and screened for validity. Because EEG and eye-tracking data were recorded at different sampling frequencies, the experimental task timeline was used as a common temporal reference, and both modalities were mapped onto the same task period. Experimental records with evident temporal misalignment, substantial data loss, or abnormal signal quality were excluded. After synchronization matching and quality control, a window-level EEG–eye-tracking dataset was generated for subsequent modeling and analysis.

3.2.1. Window Segmentation and Fatigue Labeling

Fatigue-state labels were determined according to the Samn–Perelli 7-point fatigue scale (SP-7) recorded for each experimental trial [23]. To reduce label uncertainty associated with borderline fatigue states, an extreme-group strategy was adopted to construct binary labels. Trials with SP-7 scores of 1–3 were labeled as the Alert state, whereas trials with SP-7 scores of 5–7 were labeled as the Fatigue state. An SP-7 score of 4 corresponds to the middle of the scale and may reflect an intermediate or ambiguous subjective state rather than a clearly alert or clearly fatigued condition. Therefore, trials with an SP-7 score of 4 were excluded from binary classification to reduce potential label noise and improve the separation between the two target classes. This criterion was intended to improve label reliability for supervised model training, while acknowledging that fatigue is a continuous state rather than a strictly binary condition.

After label assignment, the synchronized EEG and eye-tracking data were segmented using a sliding-window strategy. Considering the need to balance real-time detection capability and short-term signal stability, the window length was set to 5 s, with a sliding step of 2 s. Each time window was treated as an independent sample and inherited the fatigue-state label of its corresponding experimental trial or task scenario. Windows with ambiguous labels, missing data, or insufficient signal quality were excluded from subsequent modeling. Because adjacent sliding windows partially overlapped, data splitting was not performed by random window-level mixing. Instead, window samples were assigned according to the predefined validation strategy to reduce temporal leakage; in the strict cross-subject setting, all windows from the target participant were held out for testing.

Finally, 7242 valid synchronized EEG–eye-tracking window samples were obtained, including 5182 Alert windows and 2060 Fatigue windows. With respect to traffic-flow conditions, 3745 windows were obtained from high-traffic scenarios and 3497 windows

from low-traffic scenarios. Each window contained both EEG and eye-tracking feature inputs and was used to construct the window-level fatigue-state recognition model. It should be noted that the SP-7 scale was used as a practical subjective reference for fatigue-state labeling in this study, while more objective fatigue-related measures were further considered as an important direction for future work.

3.2.2. EEG Feature Extraction

EEG signals are widely used for fatigue and drowsiness assessment because they reflect neural activity, vigilance level, and cognitive workload changes. Previous studies have shown that EEG band power and its variations are related to mental workload, fatigue, and vigilance in safety-critical operators such as drivers and pilots [3,4,24,25]. Therefore, frequency-domain and spatial-distribution-related EEG features were extracted from each time window.

Power spectral density (PSD) was estimated using the Welch method, which reduces spectral estimation variance by segmenting the signal, applying windowing, and averaging segment-wise periodograms [26]. For a given frequency band, the band power was calculated as:

$$P_b = \int_{f_1}^{f_2} PSD(f)df \tag{1}$$

where P_b denotes the power of frequency band b, $PSD(f)$ denotes the power spectral density, and f_1 and f_2 represent the lower and upper frequency limits of the corresponding band, respectively.

To reduce the influence of inter-subject differences in overall EEG energy on feature representation, relative band power was further calculated as follows:

$$RP_b = \frac{P_b}{\sum_{k=1}^{K} P_k} \tag{2}$$

where RP_b denotes the relative power of frequency band b, K is the number of target frequency bands, and $\sum_{k=1}^{K} P_k$ denotes the total power across all target bands.

In addition to absolute and relative band power, band-ratio features were extracted, including theta/alpha, theta/beta, alpha/beta, and $(theta + alpha)/beta$. These ratios have been used to characterize fatigue-related changes, such as relatively enhanced slow-wave activity or reduced fast-wave activity [27,28]. To capture spatial EEG patterns, aggregated band features were calculated over the frontal, central, parietal, and occipital regions, and hemispheric asymmetry features were computed from corresponding channel pairs, including F4–F3, C4–C3, P4–P3, and O2–O1. In total, 53 EEG features were extracted from each time window.

3.2.3. Eye-Tracking Feature Extraction

Eye movement behavior reflects attentional allocation, visual search, and vigilance changes during sustained visual monitoring in remote tower operations. Previous studies have shown that fixation, saccade, pupil-related measures, and gaze dynamics can be used to characterize visual fatigue and attentional state changes [2,5]. Eye-tracking signals are also commonly combined with EEG in multimodal fatigue detection to complement the characterization of overt visual behavior under fatigue [7]. Pupil diameter is associated with cognitive processing load and mental effort [29], pupil fluctuations are related to arousal and noradrenergic activity [30,31], and saccadic dynamics can reflect arousal, attention, and time-on-task effects [32,33].

In this study, window-level eye-tracking features were extracted from raw gaze data and fixation-event data. The extracted features included pupil-diameter statistics, gaze-position statistics, eye movement velocity features, fixation count, fixation rate, mean and total fixation duration, maximum fixation duration, fixation-duration variability, estimated transition distance between adjacent fixation points, and velocity-related statistics from fixation-event files.

Through this procedure, 47 eye-tracking features were extracted from each time window. The 53 EEG features and 47 eye-tracking features were then concatenated to form a 100-dimensional window-level multimodal feature vector. The feature representation of the $i-th$ time window was defined as:

$$\mathbf{x}_i = \left[\mathbf{x}_i^{EEG}, \mathbf{x}_i^{Eye}\right] \tag{3}$$

where $\mathbf{x}_i^{EEG} \in \mathbb{R}^{53}$ denotes the EEG feature vector of the $i-th$ time window, $\mathbf{x}_i^{Eye} \in \mathbb{R}^{47}$ denotes the corresponding eye-tracking feature vector, and $\mathbf{x}_i \in \mathbb{R}^{100}$ represents the fused input feature vector for that time window.

3.2.4. Feature Preprocessing

Before model training, data splitting was first performed according to the corresponding validation strategy. Missing-value imputation and z-score standardization were then fitted only on the training set. Specifically, missing values were replaced using the median values calculated from the training set, and the mean and standard deviation for z-score standardization were also estimated from the training set. The same imputation and standardization parameters were subsequently applied to the validation and test sets. This procedure ensured that no information from the validation or test samples was used during preprocessing [34,35]. In addition, no label-driven or test-set-based feature selection was performed; the EEG and eye-tracking feature sets were predefined according to the signal-processing procedures described above.

Z-score standardization was performed as follows:

$$x' = \frac{x - \mu}{\sigma} \tag{4}$$

where x denotes the original feature value, x' denotes the standardized feature value, and μ and σ represent the mean and standard deviation of the corresponding feature in the training set, respectively. For features with zero standard deviation, a stabilization procedure was applied to avoid numerical instability. After preprocessing, the window-level EEG–eye-tracking feature vectors were used as inputs to the subsequent fatigue detection model.

3.3. Modality-Gated Fusion Network

To achieve adaptive fusion of EEG and eye-tracking features, this study proposes the Gated EEG–Eye Fusion Network (GEEF-Net). The model is designed to dynamically allocate the contributions of EEG and eye-tracking modalities according to the feature state of each time window, and to generate a fused representation for fatigue probability estimation. Compared with early fusion methods based on direct feature concatenation, GEEF-Net can regulate the contribution of different modalities at the sample level, thereby reducing the influence of modality-scale differences, signal noise, and inter-individual variability on the fusion results. The overall network architecture is shown in Figure 3.

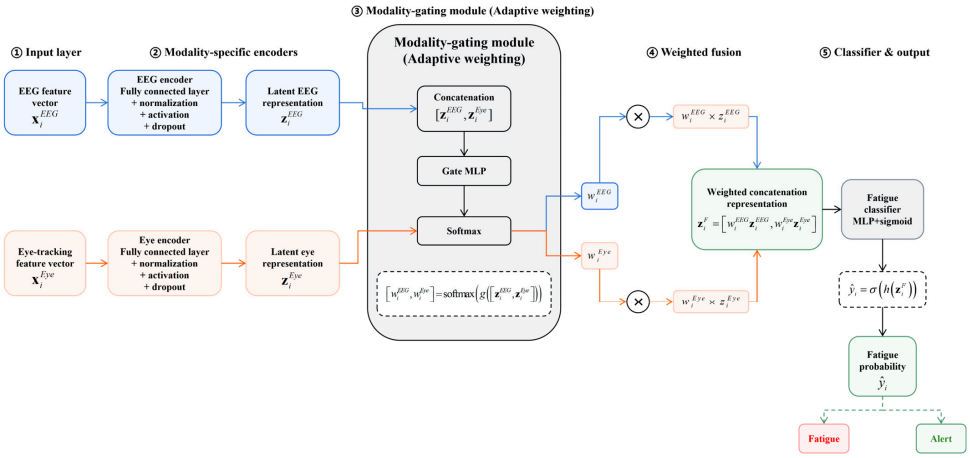

Figure 3. Overall architecture of GEEF-Net.

For the $i-th$ time window, the EEG feature vector and eye-tracking feature vector are denoted as $\mathbf{x}_i^{EEG} \in \mathbb{R}^{53}$ and $\mathbf{x}_i^{Eye} \in \mathbb{R}^{47}$, respectively. The model first employs two independent modality-specific encoders to learn latent representations from the two input modalities, as defined in Equation (5):

$$\mathbf{z}_i^{EEG} = f_{EEG}\left(\mathbf{x}_i^{EEG}\right), \quad \mathbf{z}_i^{Eye} = f_{Eye}\left(\mathbf{x}_i^{Eye}\right) \tag{5}$$

where $f_{EEG}(\cdot)$ and $f_{Eye}(\cdot)$ denote the EEG encoder and the eye-tracking encoder, respectively. $\mathbf{z}_i^{EEG}$ and $\mathbf{z}_i^{Eye}$ represent the latent representations of the two modalities. In the implementation, the EEG encoder and eye-tracking encoder were both designed as lightweight multilayer perceptrons. The EEG branch mapped the 53-dimensional EEG feature vector to a 48-dimensional latent representation through fully connected layers with 96 and 48 neurons, respectively. Similarly, the eye-tracking branch mapped the 47-dimensional eye-tracking feature vector to a 48-dimensional latent representation using the same layer configuration. Each encoder used layer normalization, ReLU activation, and dropout after the hidden layers. The gating MLP received the concatenated 96-dimensional latent representation and generated two modality weights through a 48-neuron hidden layer followed by a softmax output layer. The fatigue classifier used fully connected layers with 96 and 48 hidden units, followed by a sigmoid output unit to estimate the fatigue probability.

After obtaining the two modality-specific representations, $\mathbf{z}_i^{EEG}$ and $\mathbf{z}_i^{Eye}$ are concatenated and fed into the modality-gating module to estimate the adaptive weights of EEG and eye-tracking modalities for the current time window. The gating weights are computed as follows:

$$\left[w_i^{EEG}, w_i^{Eye}\right] = \text{softmax}\left(g\left(\left[\mathbf{z}_i^{EEG}, \mathbf{z}_i^{Eye}\right]\right)\right) \tag{6}$$

where $g(\cdot)$ denotes the modality-gating function, $[\mathbf{z}_i^{EEG}, \mathbf{z}_i^{Eye}]$ denotes the concatenated latent representation, and w_i^{EEG} and w_i^{Eye} represent the weights assigned to the EEG and eye-tracking modalities in the $i-th$ time window, respectively. Because a softmax function is used, the two weights satisfy $w_i^{EEG} + w_i^{Eye} = 1$, which allows them to be interpreted as the relative contributions of the two modalities to the classification decision for the current sample. Figure 4 illustrates the detailed computation process of the modality-gating mechanism, including feature concatenation, Gate MLP, softmax-based weight generation, and modality-wise feature weighting.

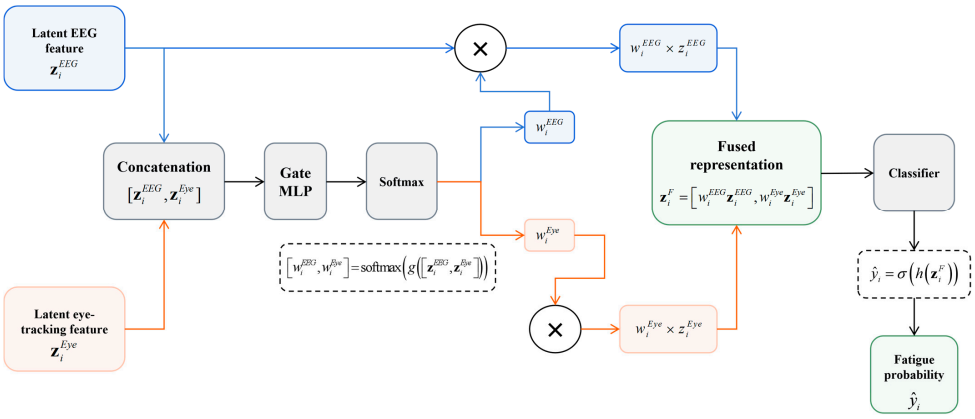

Figure 4. Details of the modality-gating mechanism.

The learned gating weights are then applied to the two latent modality representations to construct the final fused representation, as shown in Equation (7):

$$\mathbf{z}_i^F = \left[w_i^{EEG}\mathbf{z}_i^{EEG}, w_i^{Eye}\mathbf{z}_i^{Eye}\right] \tag{7}$$

where $\mathbf{z}_i^F$ denotes the modality-gated fused representation of the $i-th$ time window. This structure preserves the modality-specific representations of EEG and eye-tracking signals while regulating their relative contributions during the fusion process.

Finally, the fused representation $\mathbf{z}_i^F$ is fed into the fatigue classifier to estimate the probability that the current time window belongs to the Fatigue state, as defined in Equation (8):

$$\hat{y}_i = \sigma\left(h\left(\mathbf{z}_i^F\right)\right) \tag{8}$$

where $h(\cdot)$ denotes the fatigue classifier, $\sigma(\cdot)$ denotes the sigmoid activation function, and $\hat{y}_i \in [0, 1]$ represents the predicted probability that the $i-th$ time window belongs to the Fatigue state. When $\hat{y}_i$ exceeds a predefined decision threshold, the time window is classified as Fatigue; otherwise, it is classified as Alert. The decision threshold is determined based on validation-set performance to reduce the influence of class imbalance on model evaluation.

Overall, GEEF-Net consists of three main components: modality-specific encoders, a modality-gating module, and a fatigue classifier. This architecture preserves complementary information from EEG and eye-tracking modalities while dynamically adjusting modality contributions across time windows. Compared with more complex residual or interaction-based fusion networks, the proposed weighted fusion strategy provides a more compact structure, reducing model complexity and the risk of overfitting while retaining a degree of interpretability.

3.4. Baseline Models and Ablation Variants

To evaluate the effectiveness of the proposed GEEF-Net, this study compared it with several baseline models and ablation variants under the same validation settings. The baseline models were designed to examine the independent contribution of EEG and eye-tracking features as well as the effect of direct feature-level fusion. Specifically, the EEG-only and Eye-only models used only the corresponding single-modality features, while the Early Fusion model directly concatenated the 53 EEG features and 47 eye-tracking features into a 100-dimensional input vector for classification. To further analyze the architectural contribution of GEEF-Net, several ablation variants were constructed by modifying the use of modality-specific encoders, gating weights, residual connections, and explicit cross-modal interaction terms. The compared models are summarized in Table 1.

Table 1. Baseline models and ablation variants used for model comparison.

Category	Model	Main Configuration	Purpose
Baseline	EEG-only	53-dimensional EEG features only	Evaluate the independent discriminative ability of EEG features
Baseline	Eye-only	47-dimensional eye-tracking features only	Evaluate the independent discriminative ability of eye-tracking features
Baseline	Early Fusion	Direct concatenation of EEG and eye-tracking features	Compare direct feature-level fusion with modality-gated fusion
Ablation	Late Concat Fusion	Dual modality-specific encoders followed by direct concatenation	Examine modality-specific representation learning
Ablation	GEEF-Net	Dual encoders with sample-level modality-gated fusion	Evaluate the proposed lightweight gated fusion strategy
Ablation	Gated Fusion w/o Residual	Gated fusion with explicit interaction but without residual connection	Assess the effect of gating without residual preservation
Ablation	Residual Gated w/o Interaction	Residual connection and gated fusion without explicit interaction	Examine residual preservation without explicit interaction
Ablation	Full Residual Gated Fusion	Residual connection, gated fusion, and explicit interaction	Test whether increased fusion complexity improves performance

For the ablation variants, residual connections and explicit cross-modal interactions were defined at the latent-representation level. The basic gated representation was ($z_i^G = [w_i^{EEG} z_i^{EEG}, w_i^{Eye} z_i^{Eye}]$). Residual preservation was implemented by retaining the original latent representations (z_i^{EEG}) and (z_i^{Eye}) in the classifier input. Explicit interaction was implemented using the element-wise product ($z_i^{EEG} \odot z_i^{Eye}$) and absolute difference ($\left| z_i^{EEG} - z_i^{Eye} \right|$) between the two modality representations. Thus, the full residual gated fusion variant combined original latent representations, gated representations, and interaction terms before classification.

3.5. Validation Strategy and Model Training

To systematically evaluate the performance of the GEEF-Net in remote tower controller fatigue detection, this study designed a multi-level validation framework. The subject-dependent main validation was used to assess the window-level fatigue recognition capability of the model when subject-specific historical samples were available. Strict cross-subject validation was performed using a leave-one-subject-out (LOSO) strategy, in which all data from one participant were used as the test set in each iteration, while data from the remaining participants were used for training. This setting was used to examine the generalization capability of the model when applied to unseen participants. Considering that some participants may not contain valid windows from both Alert and Fatigue states, a paired-subject sensitivity analysis was further conducted on the subset of participants containing both state categories, in order to reduce the influence of missing state categories on the interpretation of cross-subject results.

To further simulate the practical situation in which a small amount of target-subject baseline data may be available during deployment, a few-shot subject-specific calibration experiment was conducted. In each round of cross-subject evaluation, 10%, 20%, 30%, and 40% of the target-subject samples were selected as calibration data, while the remaining target-subject samples were used as the test data. The calibration samples were used only

for target-subject adaptation and were not included in the final test set, thereby ensuring that the test results were still evaluated on independent samples. This setting was used to assess the effect of limited target-subject data on model adaptation.

In all validation settings, data preprocessing was strictly confined to the training data. Missing-value imputation and z-score standardization parameters were computed only from the current training set and then applied to the corresponding validation or test set to avoid information leakage. Model training was implemented using the PyTorch 3.9 framework. In GEEF-Net, the hidden and latent dimensions of each modality encoder were set to 96 and 48, respectively. Each encoder used fully connected layers with layer normalization, ReLU activation, and dropout. The Gate MLP used a 48-neuron hidden layer followed by a two-neuron softmax output to generate the EEG and eye-tracking modality weights. The classifier used fully connected layers with 96 and 48 hidden units, followed by a single sigmoid output unit for fatigue-probability estimation. The model was trained using the AdamW optimizer with a learning rate of (3×10^{-4}), a weight decay of (1×10^{-4}), a batch size of 64, and a maximum of 180 epochs. The dropout rate was set to 0.20, and early stopping with a patience of 25 epochs was applied based on validation-set performance to reduce overfitting. For GEEF-Net, the modality-gating module was trained in an end-to-end manner to adaptively learn the relative contributions of EEG and eye-tracking features across different time windows.

All validation strategies generated window-level prediction results, which were evaluated using the metrics described in Section 3.6. For GEEF-Net, window-level modality-gating weights were additionally saved for subsequent analysis of the relative contributions of EEG and eye-tracking modalities under different fatigue states and task conditions. For all validation settings, preprocessing-parameter estimation, threshold selection, and early stopping were performed without using the test set, which was reserved exclusively for final performance evaluation.

3.6. Evaluation Metrics and Statistical Analysis

To systematically evaluate the performance of the GEEF-Net and the baseline models in the window-level fatigue detection task, this study used multiple performance metrics and statistical analysis methods to ensure the reliability and interpretability of the results. The main evaluation metrics included Accuracy, Balanced Accuracy, Recall, Specificity, F1-score, and ROC-AUC. The relationships between the predicted window-level labels $\hat{y}_i$ and the ground-truth labels y_i were quantified using the following metrics:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$Specificity = \frac{TN}{TN + FP} \tag{12}$$

$$Balanced\ Accuracy = \frac{Recall + Specificity}{2} \tag{13}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} \tag{14}$$

where TP, TN, FP, and FN denote the numbers of true-positive, true-negative, false-positive, and false-negative windows, respectively. Since the numbers of Alert and Fatigue windows were moderately imbalanced, the use of Accuracy alone could overestimate

the model's ability to recognize the majority class. Therefore, Balanced Accuracy, Recall, Specificity, F1-score, and ROC-AUC were also reported to provide a more comprehensive assessment of the model's discriminative performance between Alert and Fatigue states. ROC-AUC was used to evaluate the overall classification capability of the model across different decision thresholds.

Model performance calculation and analysis were implemented in Python 3.9 and R 4.3.1 to ensure the reproducibility and verifiability of data processing and result computation. All window-level prediction samples were evaluated using the above standard classification metrics, and the same evaluation protocol was applied to the training, validation, and test sets to support subsequent model performance comparison and statistical analysis.

In addition to model performance evaluation, record-level EEG features were statistically compared between Alert and Fatigue states. Considering the potential non-normality of feature distributions and unequal sample sizes between states, the Mann–Whitney U test was used for between-state comparisons. Cliff's delta was used to quantify effect size, and false discovery rate (FDR) correction was applied to control for multiple comparisons.

4. Results

4.1. Main Subject-Dependent Fatigue Detection Performance

To evaluate the basic recognition capability of GEEF-Net in the window-level fatigue detection task, this study compared four models under the main subject-dependent validation setting: EEG-only, Eye-only, Early Fusion, and GEEF-Net. This setting mainly reflects model performance when subject-specific historical samples are available.

As shown in Table 2, the EEG-only and Eye-only models showed comparable but limited performance, suggesting that each modality contained useful but incomplete fatigue-related information. Early Fusion improved the overall results by directly combining EEG and eye-tracking features, indicating the complementary value of the two modalities. Compared with Early Fusion, GEEF-Net achieved higher Accuracy, Specificity, F1-score, and ROC-AUC, although Early Fusion obtained the highest Recall. This suggests that the modality-gating mechanism mainly improved overall discrimination and reduced false alarms for Alert windows, rather than simply increasing fatigue-window sensitivity. Therefore, GEEF-Net was selected as the main model for subsequent analyses.

Table 2. Window-level performance of baseline and modality-gated fusion models under the main subject-dependent evaluation.

Model	Accuracy	Balanced Accuracy	Recall	Specificity	F1-Score	ROC-AUC
EEG-only	0.801	0.779	0.728	0.831	0.676	0.868
Eye-only	0.791	0.776	0.741	0.811	0.668	0.861
Early Fusion	0.865	0.845	0.801	0.890	0.771	0.931
GEEF-Net	0.883	0.847	0.765	0.930	0.788	0.944

Figure 5 presents the ROC curves of the compared models. GEEF-Net achieved the highest ROC-AUC of 0.944, followed by Early Fusion, EEG-only, and Eye-only, indicating favorable Alert–Fatigue discrimination under the subject-dependent setting.

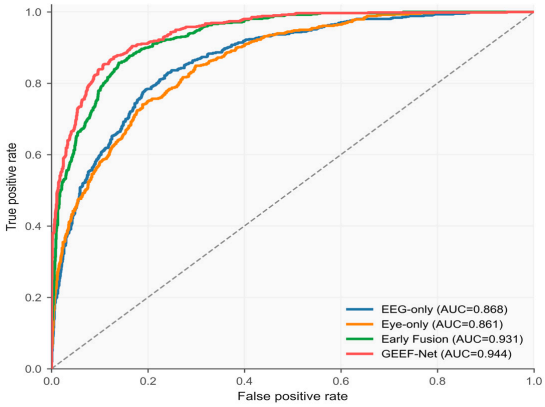

Figure 5. ROC curves of the compared models under the main subject-dependent validation setting.

4.2. Ablation Analysis of Fusion Architecture Components

To further examine the contribution of different architectural components in GEEF-Net to fatigue detection performance, several fusion-architecture variants were constructed for ablation analysis. All variants were trained and tested under the same subject-dependent main validation setting to ensure comparability across models. The ablation study mainly focused on three design factors: the use of modality-specific encoders, the introduction of a modality-gating mechanism, and the inclusion of residual connections or explicit cross-modal interaction terms.

As shown in Table 3, M2 Late Concat Fusion achieved higher Balanced Accuracy, F1-score, and ROC-AUC than M1 Early Fusion, suggesting that modality-specific encoders improved feature representation before multimodal fusion. This indicates that separately learning EEG and eye-tracking latent representations may be more suitable for the current task than directly concatenating the original features.

Table 3. Performance comparison of different GEEF-Net fusion variants in the ablation study.

Variant	Main Configuration	Accuracy	Balanced Accuracy	F1-Score	ROC-AUC
M1 Early Fusion	Direct feature concatenation	0.865	0.845	0.771	0.931
M2 Late Concat	Dual encoders + concatenation	0.871	0.859	0.785	0.937
M3 GEEF-Net	Dual encoders + gated fusion	0.883	0.847	0.788	0.944
M4 Gated Fusion w/o Residual	Gated fusion + interaction	0.855	0.836	0.757	0.933
M5 Residual Gated w/o Interaction	Residual + gated fusion	0.875	0.851	0.783	0.938
M6 Full Residual Gated Fusion	Residual + gated fusion + interaction	0.865	0.845	0.771	0.932

GEEF-Net (M3) achieved the highest Accuracy, F1-score, and ROC-AUC among the ablation variants, with values of 0.883, 0.788, and 0.944, respectively. Compared with M2, GEEF-Net further improved Accuracy, F1-score, and ROC-AUC, indicating that the modality-gating mechanism contributed to overall discriminative performance. This result supports the use of sample-level adaptive modality weighting in EEG–eye-tracking fatigue detection.

Further introducing explicit interaction terms or a full residual structure did not lead to consistent performance gains. The overall performance of M4 and M6 was lower than that of M3, while M5 approached GEEF-Net in some metrics but still showed lower ROC-AUC. These results suggest that, under the current sample size and feature-dimensionality conditions, more complex residual-interaction structures may introduce redundant in-

formation or increase the risk of overfitting. Based on the ablation results, this study selected the relatively compact GEEF-Net architecture as the main model, consisting of dual-branch modality-specific encoders and sample-level modality gating, without adding a full residual-interaction module.

4.3. Modality Contribution and Traffic-Condition Analysis

To examine the modality fusion behavior of GEEF-Net, the modality-gating weights for EEG and eye-tracking features were extracted from the test set and compared across fatigue states and traffic conditions.

As shown in Figure 6, GEEF-Net assigned higher average weights to eye-tracking features than to EEG features across the overall test set. This pattern was also observed under both Alert and Fatigue states, as well as under high- and low-traffic conditions. In particular, the eye-tracking weight was higher in Fatigue windows than in Alert windows, indicating that the model relied more on eye-tracking information for these samples.

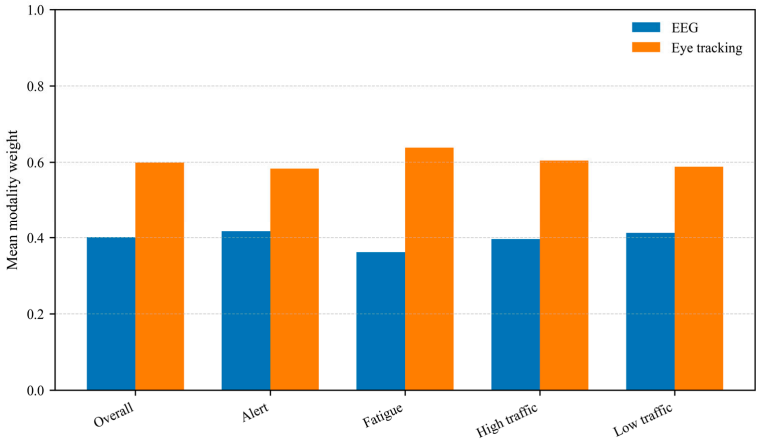

Figure 6. Learned modality-gating weights of GEEF-Net across fatigue states and traffic-flow conditions.

Traffic-specific performance comparisons are shown in Figure 7. GEEF-Net showed slightly better performance under high-traffic conditions than under low-traffic conditions, particularly in Accuracy, Balanced Accuracy, and ROC-AUC.

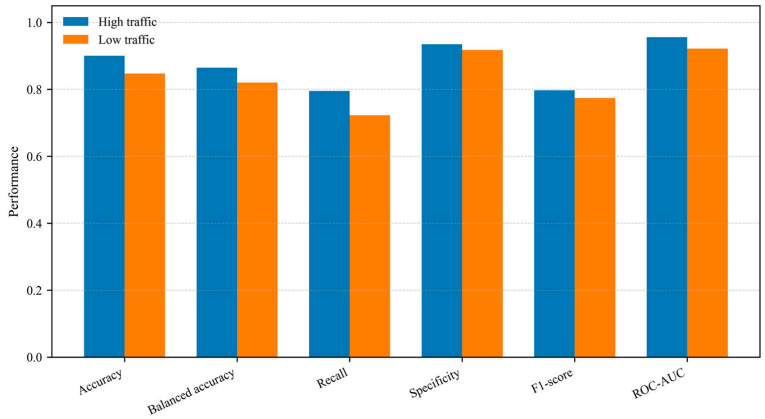

Figure 7. Traffic-specific performance of GEEF-Net under high- and low-traffic scenarios.

Overall, the modality-gating weights and traffic-stratified results provide additional evidence for the interpretability of GEEF-Net and show that the relative contribution of EEG and eye-tracking features varies across task conditions. Because the high-traffic scenario also involved longer task duration and greater operational complexity, the traffic-specific

results should be interpreted as exploratory scenario-stratified findings rather than as isolated effects of traffic volume.

4.4. Cross-Subject Generalization and Paired-Subject Sensitivity Analysis

To evaluate generalization to unseen participants, strict cross-subject validation was conducted using a leave-one-subject-out strategy. In this setting, all windows from the same participant were assigned exclusively to either the training set or the test set, preventing leakage of subject-specific information between training and testing.

As shown in Table 4, all models showed a clear performance decrease under strict cross-subject validation compared with the subject-dependent setting. The EEG-only model achieved the highest Accuracy and ROC-AUC in this setting, with values of 0.572 and 0.532, respectively, but the overall performance remained close to chance level. The Gated fusion model did not retain its advantage under strict cross-subject validation, indicating that fully uncalibrated cross-subject fatigue detection remains challenging.

Table 4. Performance comparison under strict cross-subject and paired-subject validation settings.

Validation Setting	Model	Accuracy	Balanced Accuracy	Recall	Specificity	F1-Score	ROC-AUC
Strict cross-subject	EEG-only	0.572	0.524	0.475	0.573	0.371	0.532
Strict cross-subject	Eye-only	0.506	0.404	0.170	0.638	0.165	0.381
Strict cross-subject	Early Fusion	0.470	0.392	0.228	0.556	0.198	0.311
Strict cross-subject	Gated fusion	0.506	0.404	0.170	0.638	0.169	0.305
Paired-subject subset	EEG-only	0.547	0.554	0.552	0.556	0.509	0.554
Paired-subject subset	Eye-only	0.437	0.514	0.674	0.354	0.454	0.527
Paired-subject subset	Early Fusion	0.386	0.442	0.634	0.251	0.455	0.429
Paired-subject subset	Gated fusion	0.448	0.511	0.719	0.303	0.515	0.512

A paired-subject sensitivity analysis was further conducted using only participants who contained both Alert and Fatigue windows. The results remained limited after this restriction. Although the Gated fusion model achieved relatively high Recall in the paired-subject subset, its Specificity was low, suggesting an increased tendency toward false alarms for Alert windows. These findings indicate that controlling for state availability at the subject level did not fully resolve the cross-subject generalization problem.

Figure 8 summarizes the performance changes across validation settings. Compared with the subject-dependent evaluation, both strict cross-subject validation and paired-subject analysis showed substantial degradation. These results motivate the following few-shot subject-specific calibration analysis, in which a small amount of target-subject data is used to improve individual adaptation.

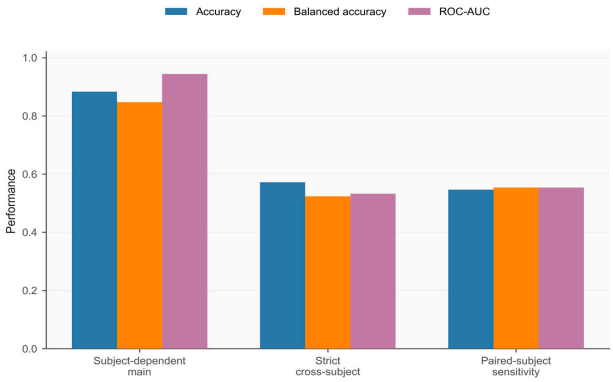

Figure 8. Performance comparison under subject-dependent, strict cross-subject, paired-subject, and calibration-based validation settings.

4.5. Few-Shot Subject-Specific Calibration

Cross-subject validation showed that model performance decreased markedly when no target-subject samples were available. To examine whether limited individual data could improve model adaptation, few-shot subject-specific calibration was conducted using GEEF-Net. In each calibration setting, 10%, 20%, 30%, or 40% of the target-subject samples were used as calibration data, while the remaining target-subject samples were held out for testing.

As shown in Table 5 and Figure 9, increasing the calibration ratio generally improved model performance. Accuracy increased from 0.703 at 10% calibration to 0.803 at 40% calibration, while ROC-AUC increased from 0.642 to 0.804. Balanced Accuracy also improved from 0.614 to 0.746, suggesting that limited target-subject data helped adjust the model to individual feature distributions. However, F1-score remained lower than the other metrics and showed large variability across subjects, indicating that stable recognition of Fatigue windows remained challenging under the few-shot setting.

Table 5. Performance summary of GEEF-Net under different subject-specific calibration ratios. Values are reported as mean ± standard deviation.

Calibration Ratio	Accuracy	Balanced Accuracy	F1-Score	ROC-AUC
10%	0.703 ± 0.151	0.614 ± 0.138	0.301 ± 0.315	0.642 ± 0.210
20%	0.750 ± 0.157	0.678 ± 0.150	0.356 ± 0.352	0.727 ± 0.199
30%	0.785 ± 0.133	0.740 ± 0.141	0.396 ± 0.372	0.792 ± 0.174
40%	0.803 ± 0.151	0.746 ± 0.153	0.410 ± 0.387	0.804 ± 0.170

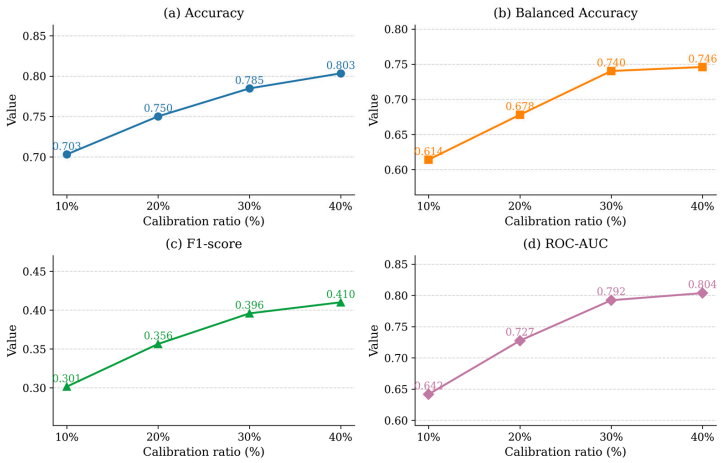

Figure 9. Few-shot subject-specific calibration performance of GEEF-Net under different calibration ratios: (**a**) Accuracy, (**b**) Balanced Accuracy, (**c**) F1-score, and (**d**) ROC-AUC.

Overall, these results indicate that a small amount of target-subject data can improve the subject-specific adaptation of GEEF-Net. Nevertheless, the large standard deviation of F1-score suggests that the calibration benefit was not fully consistent across participants, and fatigue-window recognition may still be affected by class imbalance and heterogeneity in fatigue expression.

4.6. Record-Level Statistical Comparison of Representative EEG Features

To further examine EEG-related fatigue characteristics, record-level differences in representative EEG features were compared between Alert and Fatigue states. The selected features included theta-, alpha-, and beta-related power features, relative band power, band-ratio features, and hemispheric asymmetry indicators. Considering potential non-normality and unequal sample sizes, the Mann–Whitney U test was used for between-state comparisons, Cliff's delta was used to quantify effect size, and FDR correction was applied for multiple comparisons.

As shown in Table 6, none of the representative EEG features reached statistical significance after FDR correction. Although some band-ratio features, such as theta/alpha, theta/beta, and (theta + alpha)/beta, showed higher mean values in the Fatigue state, these differences should be interpreted only as descriptive patterns rather than statistically supported group-level effects. Similarly, the observed effect-size directions of several alpha asymmetry features indicate possible variability between states, but they do not provide sufficient evidence for stable EEG feature differences. Therefore, the record-level EEG analysis should be regarded as an exploratory supplementary analysis, and individual EEG features should not be interpreted as standalone fatigue markers in the present dataset.

Table 6. Record-level statistical comparison of representative EEG features between Alert and Fatigue states. Values are reported as mean ± standard deviation.

EEG Feature	Alert	Fatigue	*p*-Value	FDR-Adjusted *p*	Cliff's Delta
Theta/alpha ratio	3.577 ± 0.685	4.547 ± 1.383	0.076	0.867	0.388
Occipital alpha asymmetry (O2-O1)	0.186 ± 0.580	0.108 ± 0.897	0.089	0.867	−0.372
Theta/beta ratio	4.527 ± 1.706	6.668 ± 3.592	0.141	0.867	0.322
(Theta + alpha)/beta ratio	5.773 ± 1.956	8.013 ± 3.895	0.141	0.867	0.322
Central alpha asymmetry (C4-C3)	0.013 ± 0.203	−0.065 ± 0.147	0.229	0.867	−0.264
Frontal alpha asymmetry (F4-F3)	0.184 ± 1.224	−0.044 ± 0.487	0.369	0.867	−0.198
Alpha/beta ratio	1.247 ± 0.284	1.345 ± 0.323	0.412	0.867	0.182
Relative theta power	0.167 ± 0.025	0.174 ± 0.057	0.456	0.867	−0.165
Central theta power	1967.354 ± 4216.440	2104.038 ± 3107.310	0.456	0.867	0.165
Frontal theta asymmetry (F4-F3)	0.226 ± 1.239	0.028 ± 0.454	0.480	0.867	−0.157

Figure 10 summarizes the direction and relative magnitude of Cliff's delta values for the selected EEG features. Because the corresponding statistical tests did not remain significant after FDR correction, this figure is intended only to visualize descriptive effect-size patterns. These results suggest that EEG features may still contribute useful information within the multimodal model, but the present record-level analysis does not support using any single EEG feature as a robust discriminator between Alert and Fatigue states.

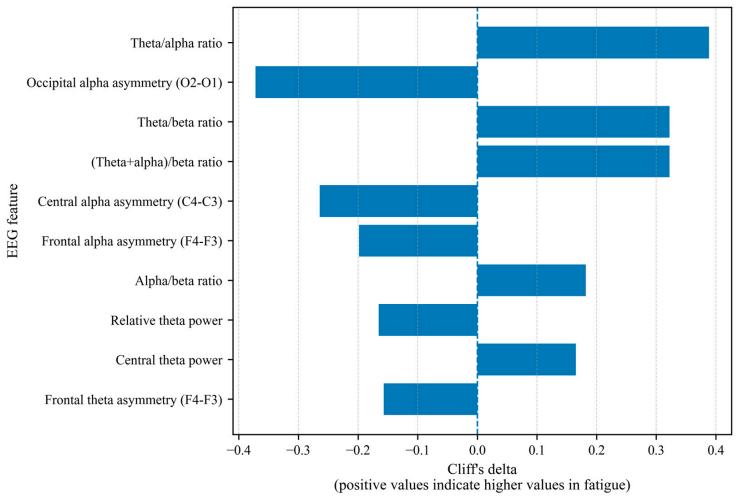

Figure 10. Effect sizes of representative EEG features between Alert and Fatigue states. Positive Cliff's delta values indicate higher feature values in the Fatigue state, whereas negative values indicate higher feature values in the Alert state.

5. Discussion

5.1. Effectiveness and Interpretation of EEG–Eye-Tracking Fusion

The results indicate that EEG–eye-tracking fusion is beneficial for window-level fatigue detection in remote tower controllers. Under the subject-dependent setting, GEEF-Net outperformed the EEG-only and Eye-only models in most overall metrics, suggesting that the two modalities provide complementary information. This finding is consistent with previous EEG–eye-tracking fatigue detection research, which has shown that a single modality may not fully capture fatigue-related state changes [7].

From a human-factor perspective, EEG and eye-tracking signals characterize different aspects of fatigue. EEG features mainly reflect changes in vigilance, spectral activity, and neural regulation. Recent EEG-based fatigue detection studies have also moved from traditional frequency-domain features toward deep neural networks, spatio-temporal modeling, and multidimensional feature fusion [3,18,36]. In contrast, eye-tracking features capture overt visual behavior, including fixations, saccades, pupil dynamics, and gaze patterns, which are closely related to visual fatigue and attentional changes during sustained monitoring tasks [2,5]. Remote tower operations require continuous visual scanning, multisource information integration, and sustained situation awareness. Therefore, fatigue in this context may involve both internal neurophysiological regulation and observable changes in visual monitoring behavior.

The comparison among the four models further highlights the contribution of modality gating. The EEG-only and Eye-only models were limited by modality-specific noise and incomplete fatigue representation, whereas Early Fusion improved performance by directly combining the two feature sets. However, direct concatenation assigns fixed importance to both modalities. In contrast, GEEF-Net adaptively weighted EEG and eye-tracking representations for each window. Compared with Early Fusion, GEEF-Net achieved higher Accuracy, Specificity, F1-score, and ROC-AUC, although Early Fusion showed higher Recall. This suggests that the main contribution of modality gating lies in improving overall discrimination and false-alarm control rather than simply increasing fatigue-window sensitivity.

The learned modality-gating weights further support this interpretation. GEEF-Net assigned relatively higher weights to eye-tracking features, especially in Fatigue windows. This pattern is consistent with the screen-mediated nature of remote tower operations, where fatigue-related changes may be more directly reflected in visual search, fixation

allocation, pupil variation, and gaze transition behavior. However, the higher eye-tracking weight does not imply that EEG is unimportant. Rather, EEG provides complementary internal physiological information, while eye-tracking features may offer more direct behavioral evidence in certain time windows.

Traffic-condition analysis also suggests that task context may influence fatigue detectability. GEEF-Net showed slightly better performance under the high-traffic scenario, particularly in Accuracy, Balanced Accuracy, and ROC-AUC. However, this result should be interpreted cautiously because the high-traffic scenario involved not only more aircraft, but also longer task duration and greater operational complexity. Therefore, the observed difference may reflect the combined effects of aircraft number, time-on-task, visual scanning demand, conflict monitoring, and coordination complexity. Future studies should further control task duration and independently manipulate traffic volume and task difficulty.

5.2. Lightweight Gating Versus Complex Interaction Fusion

This study adopted a lightweight modality-gating strategy rather than a more complex cross-modal interaction structure. Deep multimodal fusion methods, including attention-based and Transformer-based interaction models, can enhance representation learning but often require larger datasets, higher computational resources, and more stable training conditions [20,21]. These requirements are particularly relevant in human-factors experiments, where EEG and eye-tracking data are often limited in sample size, affected by signal noise, and characterized by strong individual differences.

The ablation results support the use of a compact gating structure. Modality-specific encoders improved feature representation compared with direct feature concatenation, while the gating mechanism allowed the model to adjust the relative contribution of EEG and eye-tracking features at the sample level. Similar gating strategies have been used in multimodal learning to dynamically regulate the contribution of different modalities according to sample characteristics [37]. In the present study, more complex residual or explicit interaction structures did not provide consistent additional gains, suggesting that sample-level adaptive weighting was sufficient to capture the main complementary information between EEG and eye-tracking features under the current data conditions.

From a deployment perspective, lightweight gating also has practical advantages. Remote tower fatigue monitoring requires continuous inference, low computational burden, and interpretable model behavior. The gating weights provide model-level cues regarding modality reliance under different fatigue states and traffic conditions. These weights should not be treated as direct physiological causal evidence, but they improve model transparency compared with black-box fusion structures and may support subsequent human-centered system design.

5.3. Cross-Subject Generalization and Subject-Specific Calibration

The cross-subject results show that individual differences remain a major challenge for remote tower fatigue detection. When target-subject samples were completely excluded from training, all models showed substantial performance degradation. This finding is consistent with EEG transfer learning studies, which have shown that EEG data are affected by non-stationarity, inter-individual variability, and cross-session differences [10,11]. In remote tower tasks, differences in physiological baselines, fatigue sensitivity, visual search strategies, task experience, and response behavior may further affect the feature distribution of each controller.

The paired-subject sensitivity analysis further showed that retaining only participants with both Alert and Fatigue windows did not fully resolve the generalization problem. This suggests that the performance decline cannot be attributed only to subject–label imbalance,

but also reflects deeper inter-individual differences in physiological and behavioral fatigue expression. Therefore, subject-dependent performance alone is insufficient for estimating the deployability of fatigue detection models in operational settings.

Few-shot subject-specific calibration provides a more practical direction. The calibration results showed that using a small proportion of target-subject samples improved the overall performance of GEEF-Net, especially in Accuracy, Balanced Accuracy, and ROC-AUC. This suggests that limited individual data can help adjust subject-specific decision boundaries. However, F1-score remained relatively low and varied substantially across subjects, indicating that stable detection of Fatigue windows remains difficult under class imbalance and heterogeneous fatigue responses. In practical remote tower applications, a fully generic cross-subject model may therefore be less suitable than a model combined with lightweight individual calibration, subject-specific threshold adjustment, or cost-sensitive learning. Future work should further investigate more robust cross-subject adaptation strategies, such as domain adaptation, transfer learning, subject-invariant representation learning, and subject-specific threshold calibration, to reduce the dependence on individual calibration samples. These findings suggest that practical deployment may require a hybrid strategy that combines a general population-level model with lightweight subject-specific calibration. Such calibration could help adjust individual decision boundaries and reduce the influence of physiological baseline differences and visual-search habits in previously unseen controllers.

5.4. Limitations and Future Work

Several limitations should be acknowledged. The data were collected in a remote tower simulation environment. Although the task procedures, traffic scenarios, and visual monitoring demands were designed to approximate operational requirements, simulated tasks cannot fully reproduce real operational risk pressure, shift schedules, organizational factors, and long-term fatigue accumulation. Therefore, the current findings should be interpreted as evidence of feasibility under controlled simulation conditions rather than direct proof of operational deployment readiness. Further validation in more realistic operational or semi-operational remote tower scenarios is still required before the proposed approach can be considered applicable to practical fatigue monitoring. This issue is particularly relevant because controller fatigue is closely related to fatigue risk management and operational safety at the organizational level [12].

The sample size and participant source may also limit the generalizability of the findings. Although the experiment included air traffic control instructors and trainees, the retained synchronized EEG–eye-tracking dataset may not fully represent the diversity of operational controllers, traffic environments, and fatigue levels. Future studies should collect larger datasets from more diverse operational backgrounds and evaluate the model across different days, traffic scenarios, fatigue accumulation stages, and real or semi-operational remote tower environments.

Another limitation is that this study did not include a conventional tower control condition. Therefore, the present results cannot directly determine whether fatigue-related EEG and eye-tracking patterns in remote tower operations differ from those in complex conventional tower environments with multiple runways. Such conventional tower environments may impose workload levels comparable to remote tower scenarios in terms of traffic density, conflict monitoring, communication load, runway configuration, and ground-movement coordination. However, remote towers introduce additional interface-mediated visual monitoring, camera-based information acquisition, and possible display-switching demands. Future studies should include parallel conventional-tower and remote-

tower scenarios to examine whether controller workload and fatigue-related physiological–behavioral patterns differ between the two operational modes.

This study also did not quantify workload-time, controller-occupancy, or airport-capacity indicators. Therefore, it cannot determine whether overload thresholds used in conventional tower capacity analysis, such as time-based active workload criteria, are directly applicable to remote tower operations. Although the present model estimates fatigue state rather than airport capacity, fatigue monitoring may provide complementary information for workload and capacity management because excessive controller workload can constrain the number of aircraft movements that can be safely handled within a given period. In remote tower centres, this relationship may further depend on the number of aerodromes being monitored, traffic density, runway configuration, communication load, conflict events, controller intervention frequency, and interface design. Future studies should integrate physiological fatigue indicators with operational workload and capacity metrics, including communication time, control-task occupancy, aircraft movements, runway crossings, conflict events, and controller intervention frequency, to evaluate workload thresholds, throughput, and safety margins in remote tower environments.

Fatigue labeling also requires further refinement. This study used SP-7-based binary labels with an extreme-group strategy, which helped reduce ambiguity in borderline samples but may still simplify the continuous and multidimensional nature of fatigue. Although SP-7 provides a practical subjective reference for fatigue-state labeling, fatigue cannot be fully characterized by subjective ratings alone. Future studies should combine subjective scales with objective task-performance indicators, such as response time, conflict-detection accuracy, operational errors, communication workload, PERCLOS, EEG vigilance indices, and eye-movement-based alertness measures, to construct more robust fatigue labels or continuous fatigue-risk scores.

Practical deployment will also require attention to sensor usability and human–machine interaction. Wearing comfort, signal quality variation, missing data handling, online calibration, and alarm-threshold design may all influence system acceptance and stability. Future research should therefore evaluate not only classification performance, but also whether model outputs can support workload management, situation awareness, and safety decision-making in remote tower operations.

6. Conclusions

This study proposed a Gated EEG–Eye Fusion Network (GEEF-Net) for fatigue detection in remote tower controllers by integrating EEG and eye-tracking features. The model was evaluated through subject-dependent validation, strict cross-subject validation, few-shot subject-specific calibration, and record-level EEG feature analysis. Under the main subject-dependent setting, GEEF-Net achieved an Accuracy of 0.883, an F1-score of 0.788, and a ROC-AUC of 0.944, outperforming single-modality models and showing a more balanced overall performance than Early Fusion. The modality-gating analysis suggested that GEEF-Net could adaptively adjust the relative contributions of EEG and eye-tracking features, supporting the value of multimodal fusion in sustained visual monitoring tasks. However, strict cross-subject validation indicated that individual differences remain a major challenge, while few-shot subject-specific calibration improved target-subject adaptation. Record-level EEG analysis did not identify statistically significant feature differences after FDR correction, indicating that single EEG indicators alone were insufficient for stable fatigue discrimination in the present dataset. Overall, these findings provide preliminary simulation-based evidence for the feasibility of EEG–eye-tracking fusion and lightweight modality gating in remote tower controller fatigue detection. Future work should validate the approach using larger, multi-session, and operationally realistic datasets, incorporate

more objective fatigue measures, and develop more robust cross-subject adaptation and online calibration strategies before practical deployment.

Author Contributions: Conceptualization, D.S. and W.P.; methodology, D.S.; validation, H.G., Z.Y. and B.H.; formal analysis, W.P.; investigation, H.G. and B.H.; resources, W.P.; data curation, Z.Y.; writing—original draft preparation, D.S.; writing—review and editing, D.S.; visualization, D.S.; supervision, W.P.; funding acquisition, W.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (U2333209), Sichuan Science and Technology Program (2024ZDZX0046), Key Laboratory of Flight Techniques and Flight Safety, CAAC (Grant Number: F2024KF11C), Sichuan Provincial Civil Aviation Flight Technology and Flight Safety Engineering Research Center (No. GY2024-66, No. GY2025-40D and No. GY2025-38D), and the Fundamental Research Funds for the Central Universities (26CAFUC03045).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the Civil Aviation Flight University of China.

Informed Consent Statement: Informed consent was obtained from all participants involved in the study.

Data Availability Statement: The data generated and analyzed in this study are not publicly available due to confidentiality agreements and privacy protection concerns for the participants.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ATCO	Air Traffic Controller
EEG	Electroencephalography
GEEF-Net	Gated EEG–Eye Fusion Network
SP-7	Samn–Perelli 7-point Fatigue Scale
LOSO	Leave-One-Subject-Out
PSD	Power Spectral Density
MLP	Multilayer Perceptron
ReLU	RectLeave-One-Subject-Out
PERCLOS	Percentage of Eyelid Closure
ROC-AUC	Area Under the Receiver Operating Characteristic Curve
FDR	False Discovery Rate

References

1. Yu, X.; Chen, C.H.; Yang, H. Air traffic controllers' mental fatigue recognition: A multi-sensor information fusion-based deep learning approach. *Adv. Eng. Inform.* **2023**, *57*, 102123. [CrossRef]
2. Sun, W.; Wang, Y.; Hu, B.; Wang, Q. Exploration of eye fatigue detection features and algorithm based on eye-tracking signal. *Electronics* **2024**, *13*, 1798. [CrossRef]
3. Othmani, A.; Sabri, A.Q.M.; Aslan, S.; Chaieb, F.; Rameh, H.; Alfred, R.; Cohen, D. EEG-based neural networks approaches for fatigue and drowsiness detection: A survey. *Neurocomputing* **2023**, *557*, 126709. [CrossRef]
4. Hamann, A.; Carstengerdes, N. Assessing the development of mental fatigue during simulated flights with concurrent EEG-fNIRS measurement. *Sci. Rep.* **2023**, *13*, 4738. [CrossRef]
5. Hu, Y.; Shen, H.; Pan, H.; Wei, W. Fatigue detection of air traffic controllers through their eye movements. *Aerospace* **2024**, *11*, 981. [CrossRef]
6. Yin, Z.; Pan, W.; Ni, S.; Huang, Y.; Wang, X.; Zhao, X.; Liu, K. RTFnet: A multi-modal physiological data fusion approach for fatigue detection in remote tower controllers. *Adv. Eng. Inform.* **2025**, *67*, 103506. [CrossRef]
7. Lian, Z.; Xu, T.; Yuan, Z.; Li, J.; Thakor, N.; Wang, H. Driving fatigue detection based on hybrid electroencephalography and eye tracking. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 6568–6580. [CrossRef]

8. Virk, J.S.; Singh, M.; Panjwani, U.; Ray, K. A multimodal feature fusion framework for sleep-deprived fatigue detection to prevent accidents. *Sensors* **2023**, *23*, 4129. [CrossRef]
9. Vortmann, L.M.; Ceh, S.; Putze, F. Multimodal EEG and eye tracking feature fusion approaches for attention classification in hybrid BCIs. *Front. Comput. Sci.* **2022**, *4*, 780580. [CrossRef]
10. Wan, Z.; Yang, R.; Huang, M.; Zeng, N.; Liu, X. A review on transfer learning in EEG signal analysis. *Neurocomputing* **2021**, *421*, 1–14. [CrossRef]
11. Wu, D.; Xu, Y.; Lu, B.L. Transfer learning for EEG-based brain–computer interfaces: A review of progress made since 2016. *IEEE Trans. Cogn. Dev. Syst.* **2020**, *14*, 4–19. [CrossRef]
12. European Union Aviation Safety Agency. *Study on the Analysis, Prevention and Management of Air Traffic Controller Fatigue*; European Union Aviation Safety Agency: Cologne, Germany, 2024. Available online: https://www.easa.europa.eu/ (accessed on 6 May 2026).
13. Di Mascio, P.; Carrara, R.; Frasacco, L.; Luciano, E.; Ponziani, A.; Moretti, L. How the tower air traffic controller workload influences the capacity in a complex three-runway airport. *Int. J. Environ. Res. Public Health* **2021**, *18*, 2807. [CrossRef]
14. Josefsson, B.; Jakobi, J.; Papenfuss, A.; Polishchuk, T.; Schmidt, C.; Sedov, L. Identification of complexity factors for remote towers. In Proceedings of the SESAR Innovation Days 2018, Salzburg, Austria, 3–7 December 2018.
15. Kearney, P.; Li, W.C.; Zhang, J.; Braithwaite, G.; Wang, L. Human performance assessment of a single air traffic controller conducting multiple remote tower operations. *Hum. Factors Ergon. Manuf. Serv. Ind.* **2020**, *30*, 114–123. [CrossRef]
16. Zhong, L.; Luo, P.; Hu, R.; Zhong, Q.; Zuo, Q.; Li, Y.; Ai, Y.; Pan, W. EEG-Based Fatigue Detection for Remote Tower Air Traffic Controllers Using a Spatio-Temporal Graph with Center Loss Network. *Aerospace* **2025**, *12*, 786. [CrossRef]
17. Hussein, R.M.; Miften, F.S.; George, L.E. Driver drowsiness detection methods using EEG signals: A systematic review. *Comput. Methods Biomech. Biomed. Eng.* **2023**, *26*, 1237–1249. [CrossRef] [PubMed]
18. Gao, D.; Wang, K.; Wang, M.; Zhou, J.; Zhang, Y. SFT-Net: A network for detecting fatigue from EEG signals by combining 4D feature flow and attention mechanism. *IEEE J. Biomed. Health Inform.* **2023**, *28*, 4444–4455. [CrossRef]
19. Guo, H.; Chen, S.; Zhou, Y.; Xu, T.; Zhang, Y.; Ding, H. A hybrid critical channels and optimal feature subset selection framework for EEG fatigue recognition. *Sci. Rep.* **2025**, *15*, 2139. [CrossRef]
20. Zhao, F.; Zhang, C.; Geng, B. Deep multimodal data fusion. *ACM Comput. Surv.* **2024**, *56*, 1–36. [CrossRef]
21. Xu, P.; Zhu, X.; Clifton, D.A. Multimodal learning with transformers: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12113–12132. [CrossRef]
22. Fu, B.; Gu, C.; Fu, M.; Xia, Y.; Liu, Y. A novel feature fusion network for multimodal emotion recognition from EEG and eye movement signals. *Front. Neurosci.* **2023**, *17*, 1234162. [CrossRef] [PubMed]
23. Samn, S.W.; Perelli, L.P. Estimating Aircrew Fatigue: A Technique with Application to Airlift Operations *Technical Report No. SAM-TR-82-21*. USAF School of Aerospace Medicine: Dayton, OH, USA, December 1982.
24. Borghini, G.; Astolfi, L.; Vecchiato, G.; Mattia, D.; Babiloni, F. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neurosci. Biobehav. Rev.* **2014**, *44*, 58–75. [CrossRef]
25. Borghini, G.; Vecchiato, G.; Toppi, J.; Astolfi, L.; Maglione, A.; Isabella, R.; Caltagirone, C.; Kong, W.; Wei, D.; Zhou, Z.; et al. Assessment of mental fatigue during car driving by using high resolution EEG activity and neurophysiologic indices. In *Proceedings of the 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*; IEEE: New York, NY, USA, 2012; pp. 6442–6445. [CrossRef]
26. Welch, P. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Trans. Audio Electroacoust.* **1967**, *15*, 70–73. [CrossRef]
27. Jap, B.T.; Lal, S.; Fischer, P.; Bekiaris, E. Using EEG spectral components to assess algorithms for detecting fatigue. *Expert Syst. Appl.* **2009**, *36*, 2352–2359. [CrossRef]
28. Zhang, C.; Wang, H.; Fu, R. Automated detection of driver fatigue based on entropy and complexity measures. *IEEE Trans. Intell. Transp. Syst.* **2013**, *15*, 168–177. [CrossRef]
29. Beatty, J. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychol. Bull.* **1982**, *91*, 276. [CrossRef] [PubMed]
30. Reimer, J.; McGinley, M.J.; Liu, Y.; Rodenkirch, C.; Wang, Q.; A McCormick, D.; Tolias, A.S. Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. *Nat. Commun.* **2016**, *7*, 13289. [CrossRef]
31. Joshi, S.; Gold, J.I. Pupil size as a window on neural substrates of cognition. *Trends Cogn. Sci.* **2020**, *24*, 466–480. [CrossRef] [PubMed]
32. Di Stasi, L.L.; Catena, A.; Cañas, J.J.; Macknik, S.L.; Martinez-Conde, S. Saccadic velocity as an arousal index in naturalistic tasks. *Neurosci. Biobehav. Rev.* **2013**, *37*, 968–975. [CrossRef]
33. Cazzoli, D.; Antoniades, C.A.; Kennard, C.; Nyffeler, T.; Bassetti, C.L.; Müri, R.M. Eye movements discriminate fatigue due to chronotypical factors and time spent on task–a double dissociation. *PLoS ONE* **2014**, *9*, e87146. [CrossRef] [PubMed]

34. Kaufman, S.; Rosset, S.; Perlich, C.; Stitelman, O. Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data (TKDD)* **2012**, *6*, 1–21. [CrossRef]
35. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: New York, NY, USA, 2013.
36. Gao, D.; Li, P.; Wang, M.; Liang, Y.; Liu, S.; Zhou, J.; Wang, L.; Zhang, Y. CSF-GTNet: A novel multi-dimensional feature fusion network based on Convnext-GeLU-BiLSTM for EEG-signals-enabled fatigue driving detection. *IEEE J. Biomed. Health Inform.* **2023**, *28*, 2558–2568. [CrossRef]
37. Arevalo, J.; Solorio, T.; Montes-y-Gómez, M.; González, F.A. Gated multimodal units for information fusion. *arXiv* **2017**, arXiv:1702.01992. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.