aerospace

MDPI

Article

MGRF-Net: Situation Awareness Prediction for Remote Tower Controllers Based on Multimodal Physiological Data

Qinghai Zuo [1], Ruihan Liang [1,*], Weijun Pan [2,*] and Zirui Yin [3]

[1] College of Air Traffic Management, Civil Aviation Flight University of China, Guanghan 618307, China; zuoqh@cafuc.edu.cn
[2] Flight Technology and Flight Safety Research Base of the Civil Aviation Administration of China, 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
* Correspondence: ruihanliang2750@163.com (R.L.); wjpan@cafuc.edu.cn (W.P.)

Abstract

The remote tower operation mode has changed how controllers acquire, integrate, and interpret operational information, making Situation Awareness (SA) prediction more challenging because of the coupling of multiple heterogeneous information sources. To address the limitations of existing physiological-data-based studies in modeling cross-modal relationships and deep multimodal interactions, this study proposes MGRF-Net, a multimodal physiological data-driven model for predicting remote tower controllers' SA. The model first encodes eye-tracking, electroencephalography, electrocardiography, and electrodermal activity signals independently to obtain high-level temporal representations. A graph attention-enhanced relational learning module is then introduced to capture interactive dependencies among modalities, followed by a dual-branch gated fusion mechanism to adaptively integrate multimodal information and improve prediction stability. Using multimodal physiological data collected from a remote tower simulation experiment and evaluated with 12-fold cross-validation, MGRF-Net achieved 0.0658 RMSE, 0.0461 MAE, 0.8579 R^2, and 0.9308 PCC, outperforming LightGBM, MLP, PatchTST, iTransformer, and TimeMixer. Ablation experiments and SHAP analysis further confirmed the effectiveness and interpretability of the proposed model. The results indicate that MGRF-Net can effectively capture cross-modal coupling patterns in the formation of controllers' SA and provides a promising approach for complex cognitive state monitoring and intelligent assistance in remote tower operations.

Keywords: situation awareness prediction; multimodal physiological data; graph attention; relational learning; gated fusion

Check for updates

Academic Editor: Weili Zeng

Received: 9 April 2026
Revised: 2 May 2026
Accepted: 8 May 2026
Published: 10 May 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

As an important direction in the digitalization and intelligent development of air traffic management, the remote tower is gradually transforming the traditional tower operation mode that relies on direct out-the-window observation. Compared with conventional towers, remote towers acquire airport surface operational information through multiple sensing devices, such as cameras, radar, and ADS-B, and present the traffic situation to controllers via centralized display terminals [1]. This operating mode offers clear advantages in reducing the construction and operating costs of small and medium-sized airports and in improving the efficiency of air traffic management resource allocation [2]. However,

when information presentation shifts from natural continuous visual scenes to digital multi-window and multi-channel interfaces, controllers must continuously switch visual attention and integrate information across different views and sources. As a result, their cognitive processing patterns and task load characteristics also change accordingly [3]. In this context, accurately capturing the dynamic changes in controllers' SA has become a critical issue in remote tower operational safety research.

SA is generally defined as the perception of elements in the environment, the comprehension of their meaning, and the projection of their future status, and it constitutes a key cognitive basis for operational judgment and decision-making in air traffic control tasks [4,5]. Previous studies have shown that variations in SA directly affect operators' ability to extract, integrate, and interpret dynamic information, making SA particularly important in highly dynamic and safety-critical task environments such as air traffic control [6]. However, traditional SA measurement methods, such as the freeze-probe technique, post-task subjective rating scales, and real-time probe methods, have always involved clear trade-offs among directness, continuity, and task intrusiveness [7,8]. For remote tower scenarios characterized by continuous information flow and a fast task rhythm, intermittent measurements alone are insufficient to meet the demands of real-time monitoring and continuous prediction.

With the development of wearable sensors and physiological computing technologies, non-intrusive SA sensing based on physiological data has gradually become an important research direction. A systematic review by Zhang et al. indicated that physiological indicators such as eye movements, electroencephalography, and heart rate variability show considerable potential for reflecting visual attention allocation, cognitive workload, and task-state changes [9]. Among them, eye-tracking (ET) indicators currently provide the strongest body of evidence, with the largest number of related studies and generally promising predictive potential. For example, Li et al. demonstrated that design improvements based on fixation behavior analysis could enhance pilots' decision-making efficiency and SA during task-critical moments by evaluating fixation counts and fixation duration [10]. In addition, electroencephalography (EEG) has become another mainstream approach for monitoring SA changes by recording brain activity [11]. For instance, Yeo et al. used features such as α-, θ-, and β-band power, entropy, and power ratios as inputs to machine learning models for predicting air traffic controllers' SA and showed that EEG-based support vector regression achieved favorable predictive performance [12]. Furthermore, electrocardiography (ECG) and electrodermal activity (EDA) can provide complementary representations of cognitive-state changes from the perspective of autonomic nervous system regulation [13]. For example, Wei et al. found in a flight simulation study comparing different cockpit display interfaces that conditions associated with higher SA corresponded to higher heart rates, suggesting that heart rate may be sensitive to SA variation, while Wang et al. reported correlations between EDA and SA in cockpit display research, indicating that electrodermal activity can reflect SA-related changes in arousal [14,15]. In remote tower scenarios, previous studies have constructed SA prediction models using ET and heart rate variability features and have demonstrated the feasibility of using non-intrusive physiological indicators for SA prediction [16]. Overall, however, existing remote tower SA studies still mainly rely on limited modalities and shallow feature-learning frameworks and remain insufficient in characterizing more complex multimodal coupling patterns and their dynamic relationships.

Meanwhile, the combination of multimodal physiological data and deep learning methods is becoming an important direction for modeling complex cognitive states. Existing studies have shown that, compared with unimodal approaches, multimodal physiological signals usually provide better robustness and generalization in tasks such as

fatigue detection, workload assessment, and state recognition. For example, Yang et al. proposed a neural network for binary SA recognition in autonomous driving and achieved a maximum accuracy of 90.6% using 11 EEG and ET features [17]. Yu et al. used multisensor physiological signals and deep learning models to recognize air traffic controllers' mental fatigue, demonstrating the effectiveness of multimodal fusion for monitoring complex human factor states [18]. Du et al. improved fatigue detection performance by fusing heart rate and percentage of eyelid closure over the pupil over time (PERCLOS) information, further showing that cross-modal feature complementarity can significantly enhance model representation capability [19]. In remote tower-related studies, multimodal physiological data have also been used for visual fatigue modeling and state recognition, and the results suggest that ET, EEG, ECG, and EDA can reflect operator states from different dimensions, including visual behavior, cortical activity, autonomic regulation, and arousal level [20]. These studies suggest that multimodal physiological fusion has become an important approach for modeling complex human factor states. However, for continuous SA prediction, existing studies still focus mainly on feature selection and performance comparison, with limited attention to the explicit modeling of cross-modal dynamic relationships and deep fusion mechanisms. Traditional multimodal methods usually rely on feature concatenation, weighted averaging, or shallow ensemble strategies, which are easy to implement but often insufficient to capture the dynamically changing dependencies among modalities under different task states.

In recent years, graph neural networks and graph attention mechanisms have provided new ideas for modeling complex multivariate relationships [21,22]. Compared with traditional concatenation-based fusion, graph structures can more naturally represent non-Euclidean dependencies among different variables, while graph attention mechanisms can highlight critical connections and suppress irrelevant information through adaptive weight allocation, thereby enhancing information propagation across features and modalities [23,24]. Song et al. proposed a dynamical graph convolutional neural network for EEG emotion recognition, which learns inter-channel relationships through a trainable adjacency matrix and improves classification performance [25]. In remote tower-related human factors research, graph structure learning and graph attention mechanisms have already been applied to fatigue detection and have been shown to improve the fusion of multimodal physiological data [26]. However, this line of research has not yet been fully extended to the problem of SA prediction for remote tower air traffic control, and there is still a lack of an integrated framework tailored to SA prediction that combines graph-enhanced relational modeling with gated fusion.

To address these issues, this study proposes MGRF-Net, a prediction framework for remote tower controllers' SA based on multimodal physiological data. Specifically, multimodal physiological signals, including EEG, ET, ECG, and EDA, are first encoded independently to extract high-level representations for each modality. A graph attention-enhanced relational learning module is then introduced to characterize the interactive dependencies among different modalities in a data-driven manner. On this basis, a gating mechanism is further incorporated to achieve adaptive multimodal information fusion, thereby improving the accuracy and stability of SA prediction. Experimental results show that the proposed method outperforms several representative baseline models across multiple evaluation metrics and demonstrates strong predictive performance.

The main contributions of this paper are as follows:

- A multimodal physiological data-based prediction framework for controllers' SA is developed for remote tower operation scenarios, providing a new research perspective for the data-driven modeling of complex cognitive states;

- A method combining graph attention-enhanced relational learning and a gating mechanism is proposed to explicitly characterize cross-modal dependencies and achieve adaptive fusion of heterogeneous physiological information, thereby improving multimodal feature representation capability;
- Systematic validation is conducted based on remote tower simulation data, and the results demonstrate the effectiveness of the proposed method. In addition, ablation experiments and SHAP-based interpretability analysis further verify the rationality of the model design and its explanatory value.

2. Data Acquisition

2.1. Participants and Task Scenarios

A total of 36 participants were recruited in this study, all from the School of Air Traffic Management of the Civil Aviation Flight University of China, including 6 controller instructors with extensive practical experience (mean control experience: approximately 9000 $\pm$ 500 h) and 30 controller trainees who had completed professional training and passed the required assessments (mean control experience: approximately 500 $\pm$ 100 h). All participants met the Class I medical certification standards specified in the Rules for the Management of Medical Certificates for Civil Aviation Personnel issued by the Civil Aviation Administration of China. Before the experiment, all participants were informed of the research purpose, experimental procedures, and potential risks, and written informed consent was obtained. During the experiment, all participant information was anonymized to ensure privacy and data security.

It should be noted that the participants included both highly experienced controller instructors and trained controller trainees. This mixed composition was retained in the main modeling process for two reasons. First, the primary objective of this study was to develop a general multimodal physiological prediction framework for SA rather than to compare experience-related cognitive mechanisms. Second, remote tower operations in practical deployment may involve personnel with different levels of operational experience, and a mixed-subject modeling strategy can provide a preliminary approximation of such heterogeneous user conditions. Nevertheless, because only six highly experienced instructors were included, the influence of experience level on model generalization should be interpreted cautiously. Therefore, an exploratory experience-level analysis was further conducted in Section 4.4.

The experimental tasks were conducted using the Tower Client remote tower simulation system. This system integrates airport scene rendering, flight operation logic simulation, and interactive dispatching functions, and is capable of realistically reproducing the apron control environment in remote tower operations. During the experiment, participants completed typical control tasks through the controller interface, including issuing voice instructions, managing electronic flight progress strips, and planning taxi routes. Considering both the operational characteristics of remote tower environments and the objectives of this study, three representative experimental scenarios were designed: a low-traffic routine scenario, a high-traffic scenario, and an emergency aircraft blockage scenario. The specific settings are shown in Table 1.

To minimize the influence of external factors on physiological signals, participants were instructed to avoid consuming stimulating beverages such as coffee and strong tea within 24 h before the experiment and to ensure no less than 7 h of sleep on the night before testing. Before the formal experiment, each participant first completed a brief practice session to become familiar with the experimental platform and operating procedures. During the formal experiment, participants maintained a standard controller seated posture and simultaneously obtained external virtual tower scene information

through three display screens. Each participant completed two sets of experimental tasks, comprising a total of 10 scenarios, including four Scenario 1 trials, four Scenario 2 trials, and two Scenario 3 trials. To reduce order effects, the flight sequence and flight numbers within the same type of scenario were randomized. A rest period of approximately 5 min was arranged between adjacent scenarios, and the two task sets were completed on separate days with a one-day interval.

Table 1. Experimental task scenarios.

Scenario	Arrivals	Departures	Design Purpose
1	3	4	Low traffic density and low ground traffic pressure; represents routine remote tower operations.
2	7	7	High traffic density and high ground traffic pressure; represents increased operational complexity.
3	3	4	Taxiway blockage caused by an aircraft malfunction; represents an abnormal conflict resolution scenario.

2.2. Data Acquisition and Situation Awareness Labeling

During the experiment, four types of physiological data were synchronously collected, including ET, EEG, ECG and EDA. In addition, the operation logs of the Tower Client simulator, together with the responses and reaction times to the Situation Present Assessment Method (SPAM) probe questions, were also recorded. Specifically, ET data were collected using Tobii Pro Glasses 3 (Tobii AB, Stockholm, Sweden) at a sampling rate of 100 Hz. ECG signals were acquired using the ErgoLAB Biosensing wearable device (Kingfar, Guangzhou, China) at 512 Hz. EEG data were recorded using the ErgoLAB Portable EEG system (32 channels, 512 Hz; Kingfar, China), while EDA signals were collected using the ErgoLAB Wireless EDA sensor (bilateral hand electrodes, 512 Hz; Kingfar, China). The remote tower experimental scenario and data acquisition devices are shown in Figure 1. Specifically, Figure 1a illustrates the remote tower experimental interface and its area-of-interest (AOI) partitioning, which was used to support the subsequent extraction of visual attention allocation features, whereas Figure 1b presents the multimodal physiological data acquisition devices used in the experiment.

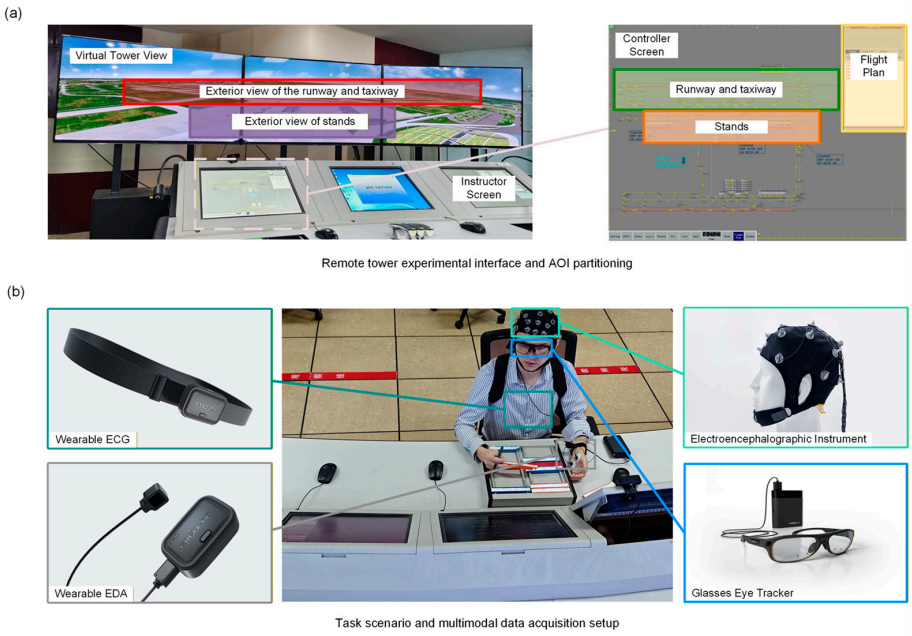


Figure 1. (**a**) Remote tower experimental interface and AOI partitioning; (**b**) Task scenario and multimodal data acquisition setup.

Situation Awareness labeling was performed using SPAM. SPAM is a real-time probe-based SA measurement method that evaluates participants' perception, comprehension, and projection abilities by inserting task-relevant questions during task execution. To make the probe questions more compatible with the remote tower context, the apron control workflow was first analyzed using hierarchical task analysis, and key task stages were then annotated according to Endsley's three-level SA model. Based on this process, SPAM probe questions matching the experimental scenarios were designed.

In the experiment, two SPAM probes were embedded in each scenario at predefined task-critical moments rather than at fixed clock-time intervals. This event-based strategy was adopted because the cognitive relevance of remote tower tasks is more closely associated with operational events, such as aircraft taxiing, runway crossing, departure release, and abnormal blockage handling [27]. Typical questions required participants to identify aircraft movement and taxiway traffic status, judge congestion, safe separation, and parking stand occupancy, and predict potential taxiway conflicts or the expected departure readiness time of a specific flight. Each participant completed ten scenarios, resulting in 20 SPAM probe responses per participant and 720 valid probe-centered samples across all 36 participants.

To align the physiological signals with the discrete SPAM responses, a fixed 30 s physiological window immediately preceding each probe onset was used for feature extraction. This pre-probe window was selected to ensure that the extracted physiological features reflected the cognitive state before the SA response rather than the response process itself. The 30 s length was adopted as a compromise between temporal specificity and feature stability: it was sufficient for extracting relatively stable features, while short enough to reduce the inclusion of unrelated task states.

In this study, the SPAM score was calculated as a weighted combination of response accuracy and reaction time, as follows:

$$\mathrm{SPAM}_{\mathrm{score}} = \mathrm{w}_{\mathrm{A}} \cdot \mathrm{Accuracy} + \mathrm{w}_{\mathrm{RT}} \cdot (1 - \mathrm{RT}_{\mathrm{norm}}) \quad (1)$$

where Accuracy denotes the response accuracy to the probe questions, and $\mathrm{RT}_{\mathrm{norm}}$ denotes the normalized reaction time. Considering that SA judgment primarily depends on whether the participant can provide a correct response, whereas reaction time mainly reflects the efficiency of information retrieval, the weights were set to $\mathrm{w}_{\mathrm{A}} = 0.7$ and $\mathrm{w}_{\mathrm{RT}} = 0.3$, respectively. This setting was based on the characteristics of safety-critical remote tower tasks and domain expert judgment. In remote tower operations, the correctness of perceiving, comprehending, and projecting traffic situations is more directly related to SA than response speed, whereas reaction time mainly reflects information retrieval efficiency under the same or similar level of response accuracy. Therefore, accuracy was assigned a higher weight, while reaction time was retained as a supplementary component.

To further examine whether the constructed SA labels were overly dependent on this fixed weighting scheme, a sensitivity analysis was additionally designed. Several alternative combinations of the accuracy weight and reaction time weight were considered. For each weighting scheme, the SPAM-derived SA labels were reconstructed while keeping the physiological input windows, model architecture, and cross-validation procedure unchanged. The corresponding prediction results are reported in Section 4.3. This analysis was used to evaluate the robustness of the proposed model and the rationality of assigning a higher weight to response accuracy in remote tower tasks.

2.3. Data Preprocessing and Feature Construction

The raw eye-tracking data were first imported into the ErgoLAB platform for preprocessing, including sliding median filtering, interpolation-based signal repair, and event

detection. The processed data were then aligned with task timestamps, from which fixation-, saccade-, blink-, pupil-, and AOI-related features were extracted. To quantify the dispersion of visual attention across different functional regions, Stationary Gaze Entropy (SGE) was further calculated based on the fixation probability distribution over AOIs [28].

The EEG, ECG, and EDA signals were processed in MATLAB R2024a with relevant toolboxes. EEG signals were preprocessed using the EEGLAB toolbox through re-referencing, band-pass filtering, independent component analysis, and artifact removal, followed by the extraction of band power, spectral entropy, and typical band ratio features from different brain regions [29,30]. ECG and EDA signals were processed using adaptive band-pass filtering, wavelet denoising, and corresponding toolbox-based procedures, from which RR-interval features, time and frequency domain HRV features, and indicators related to skin conductance level (SCL) and skin conductance response (SCR) were extracted, respectively [31,32].

To ensure that the heterogeneous physiological signals are comparable, all signals were normalized before being input into the graph neural network. Specifically, EEG and ECG signals were Z-score normalized across their respective channels to account for individual variability. Eye-tracking features were min–max normalized to scale them between 0 and 1. EDA signals were also Z-score normalized to standardize their range. This preprocessing step ensures that all signals are dimensionless and comparable for the model input.

During feature construction, a candidate set of physiological features was established from EEG, ET, ECG, and EDA modalities by integrating representative indicators that have been widely reported in previous studies as effective for characterizing cognitive state, alertness, and workload variations, while also considering the combined demands of visual scanning, cognitive processing, and autonomic response in remote tower tasks [16,33–37]. The selected features not only cover the core representational dimensions of different physiological channels but also facilitate subsequent multimodal modeling within a unified framework. The composition of the candidate feature set is presented in Table 2.

Table 2. Candidate physiological feature set.

Type	Feature	Brief Description
EEG	Frontal Theta Power	Frontal theta-band relative power
	Frontal Delta Power	Frontal delta-band relative power
	Occipital Alpha Power	Occipital alpha-band relative power
	Parietal Beta Power	Parietal beta-band relative power
ET	Pupil Diameter	Mean pupil diameter
	Blink Count	Number of blinks
	Saccade Amplitude SD	Standard deviation of saccade amplitude
	Mean Fixation Duration	Mean fixation duration
	Stationary Gaze Entropy (SGE)	Entropy of fixation distribution across AOIs
ECG	LF/HF Ratio	Ratio of low- to high-frequency power
	Mean RR Interval	Mean interval between adjacent heartbeats
	RR Interval Standard Deviation (SDNN)	Overall variability of RR intervals
EDA	Skin Conductance Level (SCL)	Mean skin conductance level
	Skin Conductance Response Count (SCR Count)	Number of skin conductance responses
	Skin Conductance Response Amplitude (SCR Amplitude)	Amplitude of skin conductance responses

The final dataset consisted of 720 valid probe-centered samples collected from 36 participants. Each sample corresponded to a 30 s physiological window immediately preceding a valid SPAM probe. Within each 30 s window, modality-specific features were extracted over successive time steps and organized into temporal feature sequences for ET, EEG, ECG, and EDA, which were then used as model inputs.

3. Methodology

3.1. Problem Definition and Overall Framework of the Proposed Model

In the remote tower operational environment, controllers' SA is a continuous cognitive state that evolves dynamically over time, and its formation depends on the integration and processing of multisource information. To effectively characterize this cognitive state, this study develops a SA prediction model based on multimodal physiological signals. Let the multimodal physiological signals collected within a time window of length N_T be denoted as:

$$X = \{X^{eeg}, X^{et}, X^{ecg}, X^{eda}\}, X^{m} \in R^{N_T \times d_m}, \quad (2)$$

where X^{eeg}, X^{et}, X^{ecg} and X^{eda} denote the feature sequences of EEG, ET, ECG, and EDA, respectively. The objective of the model is to learn a mapping function:

$$f : X \rightarrow \hat{y}, \quad (3)$$

where $\hat{y} \in R$ denotes the predicted SA score. This problem can be formulated as a multimodal time series regression task, whose main challenges lie in the heterogeneity of different modalities in both feature space and temporal scale, the complex and dynamically changing interactions among modalities, and the uncertainty of modality contributions across samples.

To address these challenges, this study proposes a Multimodal Graph Relational Fusion Network (MGRF-Net). The model follows a unified framework of intra-modal temporal modeling, inter-modal relational learning, and adaptive fusion prediction for effective multimodal physiological signal modeling. Specifically, each modality-specific time series is first encoded independently to extract its intrinsic dynamic features. The resulting high-level modality representations are then regarded as graph nodes, and a learnable relational function is used to construct the cross-modal connectivity structure, followed by graph attention-guided information propagation within the relational learning module. Finally, a gated fusion mechanism is introduced to adaptively integrate multimodal information for each sample, generating a unified representation for SA prediction. The overall architecture of the proposed model is shown in Figure 2.

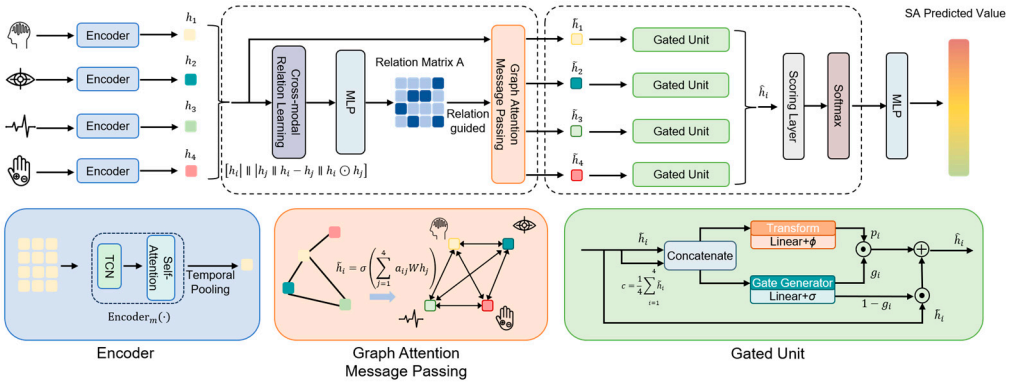


Figure 2. Overall architecture of the proposed MGRF-Net model.

3.2. Unimodal Temporal Representation Learning

Given the substantial differences among physiological signals in sampling mechanisms, noise characteristics, and physiological meanings, directly performing early fusion on multimodal data may introduce unnecessary interference. Therefore, each modality was first encoded independently to obtain high-level representations in a unified semantic space.

For any modality $m \in \{eeg, et, ecg, eda\}$, its input sequence X^m is mapped by the encoder as:

$$H^m = \text{Encoder}_m(X^m), H^m \in R^{N_T \times d}, \quad (4)$$

where Encoder_m consists of a Temporal Convolutional Network (TCN) and a self-attention mechanism. The TCN captures local temporal dependencies through a dilated convolution structure, while the self-attention mechanism further models long-range dependencies among different time steps in the sequence, thereby yielding a more comprehensive representation of temporal dynamics. This combined architecture provides strong representation capability while maintaining good training stability, making it suitable for physiological-signal scenarios with relatively limited sample sizes.

To obtain a global representation for each modality, the temporal dimension was further aggregated. In this study, average pooling was applied to compress the temporal features into a fixed-dimensional vector:

$$h_m = \frac{1}{N_T} \sum_{t=1}^{N_T} H_t^m, \ h_m \in R^d, \quad (5)$$

thus yielding the multimodal representation set:

$$H = \{h_{eeg}, h_{et}, h_{ecg}, h_{eda}\}. \quad (6)$$

This process maps the raw time series into a high-level semantic space, enabling the model to capture key dynamic features related to cognitive state while providing a unified input for subsequent cross-modal relational modeling.

3.3. Cross-Modal Relational Modeling Enhanced by Graph Attention

After obtaining high-level representations for each modality, effectively characterizing the interactions among different modalities becomes a key factor affecting model performance. The formation of SA essentially depends on the integrated processing of multisource information. Physiological signals from different modalities do not function independently; instead, they jointly reflect an individual's cognitive state through dynamic interactions. However, most existing multimodal methods rely on feature concatenation or attention-based weighting for fusion, which often lack explicit modeling of inter-modal structural relationships and therefore struggle to accurately capture the collaborative mechanisms among modalities.

To address this issue, the multimodal representations were modeled as a graph structure, in which the relationship strength among modalities was learned dynamically in a data-driven manner. Specifically, each high-level modality representation h_m was treated as a node in the graph, forming the node set:

$$V = \{h_1, h_2, h_3, h_4\}, \quad (7)$$

where the four nodes correspond to EEG, ET, ECG, and EDA, respectively. Unlike graph models that rely on predefined priors, this study does not preset the connectivity among nodes. Instead, the adjacency matrix is adaptively constructed at the sample level through a learnable function.

For any pair of modality nodes h_i and h_j, their relational feature representation is first constructed as:

$$r_{ij} = [h_i \| h_j \| |h_i - h_j| \| h_i \odot h_j], \quad (8)$$

where the concatenation terms preserve the original modality information, the difference term $|h_i - h_j|$ characterizes the discrepancy between modalities, and the element-wise

product term $h_i \odot h_j$ reflects their interaction patterns. By integrating these three types of information, the relational feature can describe the potential association between modalities from multiple perspectives.

The relational feature is then mapped by a multilayer perceptron to obtain the relational score between nodes:

$$e_{ij} = \mathrm{MLP}(r_{ij}). \tag{9}$$

A softmax operation is subsequently applied along the neighborhood dimension of each node i to normalize the scores:

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_j \exp(e_{ij})}, \tag{10}$$

thus yielding the cross-modal relational matrix $A \in R^{4\times4}$. This matrix is generated dynamically for each sample, enabling it to reflect the relative importance and interaction strength among different modalities under the current state.

Based on the learned relational matrix, a graph attention mechanism is further employed to enhance the modality representations through relationally guided information propagation. Compared with simple weighted summation or concatenation, the graph structure explicitly models the dependencies among modalities and thus guides information flow under structural constraints. Specifically, taking the multimodal representation set $H = \{h_1, h_2, h_3, h_4\}$ as graph nodes and the relational matrix A as the connection weights, the enhanced representation of node:

$$\tilde{h}_i = \sigma\left(\sum_{j=1}^{4} a_{ij} W h_j\right), \tag{11}$$

where $W \in R^{d\times d}$ is a learnable linear transformation matrix, a_{ij} denotes the influence weight of node j on node i, and $\sigma(\cdot)$ is a nonlinear activation function such as ReLU. Through this mechanism, each modality representation preserves its own information while incorporating features from other related modalities.

Different from conventional graph neural networks with fixed adjacency structures, the relational matrix A in this study is dynamically generated based on the current sample. Therefore, the graph attention process can adaptively adjust the inter-modal interaction pattern across different samples. For example, in some cases EEG signals may exhibit stronger associations with ET features, whereas in other cases autonomic nervous system indicators, such as ECG and EDA, may play a more important role. Through this dynamic relational modeling strategy, the model can flexibly adjust the information fusion path under different situations, thereby improving feature representation capability.

After graph-based relational enhancement, the updated modality representation set is obtained as:

$$\tilde{H} = \tilde{h}_1, \tilde{h}_2, \tilde{h}_3, \tilde{h}_4, \tilde{h}_i \in R^d. \tag{12}$$

This representation preserves modality-specific characteristics while incorporating cross-modal contextual information, thus providing a more informative semantic basis for the subsequent fusion stage.

3.4. Gated Fusion and Prediction

Although the preceding relational modeling and graph attention modules have enabled inter-modal information interaction, the quality and usefulness of different modalities may still vary significantly across samples. Affected by individual differences, sensor noise, and task-state variations, the reliability of each modality is not always consistent over time. For example, under some conditions EEG signals may better reflect changes in cognitive

workload, whereas in other situations eye-tracking or autonomic indicators may provide more stable information. Therefore, assigning fixed weights to all modalities or simply averaging them during fusion is unlikely to yield optimal performance.

To address this issue, a dual-branch gated fusion mechanism is introduced on top of the cross-modal relational enhancement module to enable adaptive selection and weighted fusion of multimodal information.

To incorporate global contextual information, all modality representations are first aggregated to construct a global context vector:

$$c = \frac{1}{4}\sum_{i=1}^{4} \tilde{h}_i, \tag{13}$$

then, each modality representation is concatenated with the global context to form a context-enhanced representation:

$$u_i = [\tilde{h}_i || c]. \tag{14}$$

Based on this representation, a dual-branch gating structure is designed to perform feature selection and reconstruction.

On the one hand, in the candidate mapping branch, the input is transformed through a linear mapping followed by a nonlinear activation function:

$$p_i = \phi(W_p u_i + b_p), \tag{15}$$

where $p_i \in R^d$ denotes the candidate feature representation generated under the current contextual constraint, and $\phi(\cdot)$ is the nonlinear activation function ReLU.

On the other hand, in the gate generation branch, a Sigmoid function is used to produce element-wise gating weights:

$$g_i = \sigma(W_g u_i + b_g), \tag{16}$$

where $g_i \in R^d$ is used to characterize the importance of different feature dimensions.

Subsequently, the original feature and the candidate feature are fused in an element-wise manner to obtain the gated enhanced representation:

$$\hat{h}_i = g_i \odot p_i + (1 - g_i) \odot \tilde{h}_i. \tag{17}$$

This structure introduces a residual path, allowing the model to adaptively incorporate transformed features according to the gating weights while preserving the original information, thereby preventing important information from being excessively suppressed during fusion.

After obtaining the gated enhanced representation $\hat{h}_i$ for each modality, a normalized weight allocation mechanism is further introduced at the modality level. First, a fusion score is computed for each modality:

$$s_i = W_s \hat{h}_i + b_s, \tag{18}$$

then, the modality weights are normalized using the Softmax function:

$$\alpha_i = \frac{\exp(s_i)}{\sum_{j=1}^{4} \exp(s_j)}. \tag{19}$$

Finally, the multimodal fused representation is obtained as:

$$z = \sum_{i=1}^{4} \alpha_i \hat{h}_i. \tag{20}$$

Through this design, the proposed fusion strategy achieves two-level adaptive regulation, from feature dimension selection to modality-level weight allocation. Compared with conventional fusion methods, it enables fine-grained feature selection through element-wise gating, incorporates global contextual information to better reflect the overall multimodal state, and preserves useful original information through the residual fusion structure, thereby alleviating information loss during fusion.

Based on the fused representation z, a regression prediction module is further employed to estimate the SA level. Specifically, MLP is used as the prediction head:

$$\hat{y} = f_{reg}(z). \tag{21}$$

In implementation, the regression module consists of a two-layer fully connected network:

$$\hat{y} = W_2 \phi(W_1 z + b_1) + b_2, \tag{22}$$

where W_1, W_2, b_1 and b_2 are learnable parameters, and $\phi(\cdot)$ denotes the ReLU activation function. The final output $\hat{y} \in R$ represents the predicted SA score.

Since SA is treated as a continuous variable, a regression loss function is adopted for model optimization. To balance predictive accuracy and robustness, a weighted combination of mean squared error (MSE) and mean absolute error (MAE) is used:

$$L = \lambda_1 L_{MSE} + \lambda_2 L_{MAE}, \tag{23}$$

where

$$L_{MSE} = \frac{1}{N} \sum_{n=1}^{N} (y_n - \hat{y}_n)^2, \tag{24}$$

$$L_{MAE} = \frac{1}{N} \sum_{n=1}^{N} \left| y_n - \hat{y}_n \right|, \tag{25}$$

here, λ_1 and λ_2 are used to balance the contributions of the two loss terms. MSE imposes a larger penalty on large prediction errors, thereby improving the fitting capability of the model, whereas MAE is more robust to outliers and helps enhance overall prediction stability. Their combination allows the model to achieve a better trade-off between accuracy and robustness.

3.5. Implementation Details

To ensure training stability and the reproducibility of experimental results, the proposed model and all comparison models were trained and evaluated under a unified experimental setting. All models were implemented in Python3.8, and the deep learning models were developed using the PyTorch2.0.1 framework. During training, the Adam optimizer was adopted with an initial learning rate of 0.001, a batch size of 32, and a maximum of 100 training epochs. To improve convergence stability, a learning rate decay strategy was applied, and early stopping was used when the validation performance no longer improved. To reduce the risk of overfitting, a dropout mechanism with a dropout rate of 0.2 was introduced in the deep models.

For the proposed MGRF-Net, the output dimension of each unimodal encoder was set to 64 by considering both representation capability and the relatively limited sample size. The TCN consisted of two dilated convolutional layers with a kernel size of 3 and

dilation factors of 1 and 2, respectively. This configuration was adopted to capture local temporal dependencies under different receptive fields while avoiding excessive model complexity. The self-attention module employed four attention heads to model long-range dependencies within each modality-specific temporal sequence. The number of attention heads was selected as a compromise between representation capacity and training stability. A larger number of attention heads may increase model complexity and overfitting risk under limited samples, whereas fewer heads may restrict the ability to capture diverse temporal dependency patterns.

In the cross-modal relational learning module, the dynamic relational matrix was learned by a multilayer perceptron. For each pair of modality representations, the relational input was constructed by combining the two modality vectors, their difference, and their element-wise interaction. This input was then passed through an MLP consisting of one hidden layer with 128 units, ReLU activation, and an output layer that generated the relational score between the two modalities. The relational scores were subsequently normalized to obtain the sample-specific cross-modal relational matrix. In the graph attention-enhanced relational learning module, a single graph attention layer was used to propagate and enhance inter-modal features. This single-layer design was chosen to balance relational modeling capability and training stability, and to avoid over-smoothing or overfitting that may occur when multiple graph propagation layers are used with a small number of modality nodes. Finally, the regression prediction module consisted of a two-layer fully connected network with a hidden dimension of 64 and an output dimension of 1 to generate the predicted SA value.

The main hyperparameters were determined through preliminary validation experiments rather than an exhaustive grid search. Candidate settings were compared for key components, including the encoder output dimension, the number of attention heads, the hidden dimension of the relational MLP, the number of graph attention layers, the dropout rate, and the learning rate. The final configuration was selected according to validation performance, convergence stability, and the risk of overfitting. In general, more complex settings, such as larger hidden dimensions or multiple graph attention layers, did not provide stable performance improvement and increased the risk of overfitting. Therefore, a compact model configuration was adopted in the final experiments.

All experiments were evaluated using 12-fold cross-validation, and the final results are reported as the mean and standard deviation. In each fold, the test fold was held out and used only for final performance evaluation. The remaining data were further divided into a training subset and a validation subset. The validation subset was independent of the training subset and was used only for early stopping and hyperparameter selection; it was not involved in model parameter updating. The test fold was not used during training, validation, or hyperparameter tuning. This procedure was adopted to reduce the risk of data leakage and to ensure a fair evaluation of model performance. To ensure fairness in the comparative experiments, all baseline models were trained and tested using the same input data, the same data partitioning strategy, and the same evaluation metrics.

4. Results

4.1. Overall Prediction Performance

To comprehensively evaluate the effectiveness of the proposed MGRF-Net for predicting remote tower controllers' SA, both a traditional machine learning model and several representative deep time series models were selected for comparison, including LightGBM, MLP, PatchTST, iTransformer, and TimeMixer. Among them, LightGBM, as an ensemble learning model based on gradient boosting decision trees, is capable of modeling nonlinear feature relationships and is widely used for regression tasks with tabular features. MLP, as

a classical feedforward neural network, was employed as a basic deep regression model. PatchTST enhances local temporal pattern modeling by dividing time series into local patches for Transformer encoding; iTransformer takes the variable dimension rather than the temporal dimension as the basic unit of attention and thus emphasizes correlations among variables; and TimeMixer decomposes and integrates sequence variation patterns across multiple temporal scales through a multi-scale mixing mechanism. These models represent traditional regression methods, baseline deep learning approaches, and recent mainstream time series modeling methods, thereby providing a comprehensive basis for validating the performance advantage of the proposed model.

All models were trained under the same input features, identical data partitioning, and a unified training strategy, and were evaluated using 12-fold cross-validation. Multiple continuous prediction metrics were used to comprehensively assess model performance, including root mean square error (RMSE), MAE, coefficient of determination (R^2), and Pearson correlation coefficient (PCC). Among them, RMSE is more sensitive to large prediction deviations and reflects the overall prediction accuracy of the model. MAE measures the average absolute deviation between predicted and true values, providing intuitive and stable error estimation. R^2 characterizes the model's ability to fit the variation trend of SA, with values closer to 1 indicating stronger explanatory power for the target variable. PCC is used to measure the linear correlation between the predicted and true values, with values closer to 1 indicating higher consistency in characterizing the trend of SA variation. These metrics are defined as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{26}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \tag{27}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{28}$$

$$\mathrm{PCC} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}}, \tag{29}$$

where N denotes the number of test samples, y_i denotes the true SA value of the *i*-th sample, and $\hat{y}_i$ denotes the corresponding predicted value. $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ and $\bar{\hat{y}} = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_i$ represent the mean of the true values and the mean of the predicted values in the test set, respectively. The overall experimental results of all models are presented in Table 3, and the corresponding performance trends are shown in Figure 3.

Table 3. Evaluation metric results of each model on the situation awareness prediction task.

Model	RMSE	MAE	R^2	PCC
LightGBM	0.0968	0.0735	0.6884	0.8347
MLP	0.0904	0.0662	0.7278	0.8589
PatchTST	0.0789	0.0586	0.7913	0.8939
iTransformer	0.0751	0.0539	0.8115	0.9064
TimeMixer	0.0728	0.0517	0.8256	0.9148
MGRF-Net	0.0658	0.0461	0.8579	0.9308

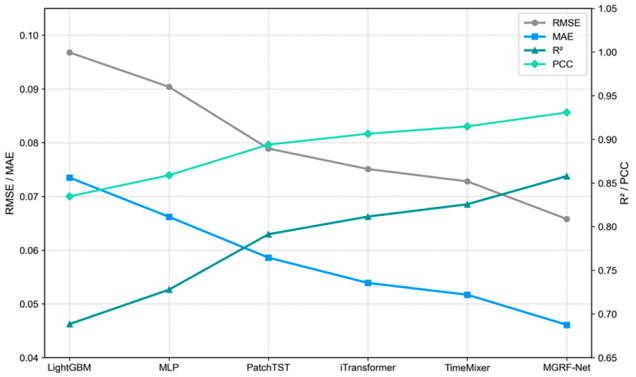


Figure 3. Performance comparison of different models on the SA prediction task.

As shown in Table 3 and Figure 3, the proposed MGRF-Net achieved the best performance across all four evaluation metrics. Compared with the baseline models, MGRF-Net yielded lower prediction errors in terms of RMSE and MAE, while achieving higher R^2 and PCC values, indicating superior overall performance in the continuous SA prediction task.

A further comparison of the different models shows that LightGBM and MLP exhibited relatively weaker overall performance, whereas modern deep time series models such as PatchTST, iTransformer, and TimeMixer achieved better predictive performance than these two baseline methods, suggesting that stronger temporal modeling capability contributes to improved SA prediction. Nevertheless, MGRF-Net still outperformed all comparison models across the overall metrics, demonstrating better error control and fitting capability.

4.2. Ablation Study

To further evaluate the effects of the key components of the proposed model and the contributions of different input modalities to the SA prediction task, ablation experiments were conducted from two perspectives: model architecture and modality composition. Specifically, module ablation was designed to examine the contributions of cross-modal relational learning, graph attention enhancement, and the gated fusion mechanism to the overall performance, whereas modality ablation was used to compare the predictive roles of different physiological modalities and to verify the effectiveness of multimodal input fusion.

4.2.1. Module Ablation

To further verify the contribution of each key component in the proposed model to SA prediction performance, a progressive module ablation design was adopted. Under the same input data, training strategy, and evaluation criteria, a stepwise enhanced framework was constructed, evolving from basic relational modeling to the complete model. It should be noted that, in the full model, the relational learning process is embedded within the graph attention-enhanced cross-modal relational modeling module. In this section, however, to separately examine the contribution of relational modeling itself, it was extracted in its basic form and used for prediction through relation-guided fusion. The specific settings are as follows.

- Relation: Only the cross-modal relational learning and relation-guided fusion processes were retained. Specifically, the relational learning module was first used to construct a sample-level cross-modal relational matrix to characterize the association strength among different modalities. The relational matrix was then aggregated to obtain modality-level relational weights, which were further used to perform weighted fusion of modality representations for final regression prediction. This setting was

used to evaluate the independent contribution of cross-modal relational modeling to the prediction task;

- Relation + Graph Attention: Based on the Relation model, a graph attention enhancement module was further introduced to enable inter-modal information propagation and feature enhancement under the guidance of the learned relational matrix;
- MGRF-Net: Based on Relation + GAT, the gated fusion mechanism was further incorporated to form the complete model.

All ablation models were trained and tested using the same input data, training strategy, and evaluation criteria. The experimental results are presented in Table 4.

Table 4. Performance comparison of module ablation models.

No.	Relation	Graph Attention	Gate	RMSE	MAE	R^2	PCC
1	√			0.0812	0.0601	0.7814	0.8892
2	√	√		0.0709	0.0516	0.8347	0.9184
3	√	√	√	0.0658	0.0461	0.8579	0.9308

As shown in Table 4, with the progressive introduction of relational learning, graph attention enhancement, and gated fusion, the model exhibited a consistent improvement across RMSE, MAE, R^2, and PCC. When only the Relation module was used, the model was already able to exploit inter-modal relational information to some extent for SA prediction, although its overall performance remained slightly lower than that of some general time series models. After introducing the graph attention enhancement module, all four evaluation metrics improved markedly, and the resulting model outperformed the general time series baselines. With the further inclusion of the gated fusion mechanism, the complete model achieved the best results, indicating that relational learning, graph attention enhancement, and gated fusion contribute to performance improvement in a progressive manner.

These results indicate that cross-modal relational learning provides an effective structural basis for the model, while graph attention enhancement and gated fusion further improve feature representation capability and predictive performance on top of this foundation. Overall, the progressive module ablation experiment verifies the effectiveness of the proposed architecture, from relational modeling to feature enhancement and then to adaptive fusion.

4.2.2. Modality Ablation

To analyze the influence of different physiological modalities on model performance, comparative experiments with different modality combinations were conducted while keeping the overall model architecture unchanged. Specifically, taking the full input of EEG, ET, ECG, and EDA as the reference condition, one modality was removed at a time to construct multiple input combinations for performance comparison. The specific modality combinations and the corresponding experimental results are presented in Table 5.

Table 5. Performance comparison of modality ablation models.

No.	EEG	ET	ECG	EDA	RMSE	MAE	R^2	PCC
1		√	√	√	0.0846	0.0628	0.7623	0.8781
2	√		√	√	0.0819	0.0604	0.7776	0.8873
3	√	√		√	0.0718	0.0526	0.8297	0.9142
4	√	√	√		0.0731	0.0537	0.8231	0.9101
5	√	√	√	√	0.0658	0.0461	0.8579	0.9308

As shown in Table 5, the complete multimodal input yielded the best performance, outperforming all single-modality removal conditions in terms of RMSE, MAE, R^2, and PCC. This result indicates that EEG, ET, ECG, and EDA provide complementary information for the SA prediction task, and that multimodal fusion contributes to improving the overall predictive performance of the model.

A further comparison of the different modality removal conditions shows that the degree of performance degradation varies depending on which modality is excluded, suggesting that the contributions of different modalities are not identical. Overall, the full model achieves lower prediction error and stronger fitting ability by jointly exploiting multimodal information. These modality ablation results further validate the effectiveness of multimodal fusion modeling for the task addressed in this study.

4.3. Supplementary Robustness Analyses

To further examine the robustness and applicability of the proposed model, two supplementary analyses were conducted. First, considering the unequal distribution of participants with different experience levels, an exploratory subgroup analysis was performed to examine whether the predictive performance of MGRF-Net differed between controller trainees and experienced instructors. Second, because the SPAM-derived SA score was calculated using fixed weights for response accuracy and reaction time, a sensitivity analysis was conducted to evaluate whether the prediction results were overly dependent on this weighting scheme.

4.3.1. Exploratory Experience-Level Subgroup Analysis

As described in Section 2.1, the participants included both experienced controller instructors and controller trainees. Although the main model was trained and evaluated using the pooled dataset to reflect the heterogeneous personnel composition that may occur in practical remote tower deployment, the imbalance between the two groups may affect the generalization of the model. Therefore, an exploratory subgroup analysis was conducted based on participant experience level.

Specifically, the test samples obtained from the 12-fold cross-validation procedure were divided into trainee and instructor groups, and the prediction metrics were calculated separately for each group. This analysis was intended to provide preliminary evidence on whether the proposed model could maintain predictive capability across participants with different experience levels, rather than to draw definitive conclusions about experience-related cognitive mechanisms.

As shown in Table 6, MGRF-Net achieved stable prediction performance in both experience groups. The trainee group showed slightly better performance, which may be partly related to its larger sample size and greater influence on the overall training distribution. The instructor group still achieved acceptable predictive performance, although its RMSE and MAE were slightly higher and its R^2 and PCC were slightly lower than those of the trainee group. These results suggest that the proposed model retained preliminary cross-experience generalization ability. However, because only six experienced instructors were included, this subgroup analysis should be interpreted as exploratory rather than conclusive.

Table 6. Prediction performance across different experience groups.

Group	Participants	Samples	RMSE	MAE	R^2	PCC
Trainees	30	600	0.0649	0.0455	0.8618	0.9324
Instructors	6	120	0.0702	0.0493	0.8326	0.9178

4.3.2. Sensitivity Analysis of SPAM Weight Settings

To evaluate the influence of the SPAM weight setting on the prediction results, MGRF-Net was retrained and tested under different combinations of response accuracy and reaction time weights. The original 0.7/0.3 setting was compared with four alternative settings. All other experimental conditions, including the 30 s pre-probe physiological window, input features, model architecture, and 12-fold cross-validation strategy, were kept unchanged.

As shown in Table 7, the prediction performance varied only slightly across different weighting schemes. The original 0.7/0.3 setting achieved the best overall performance, with the lowest RMSE and MAE and the highest R^2 and PCC. Nevertheless, the performance under the neighboring settings of 0.6/0.4 and 0.8/0.2 remained very close to that of the original setting, and no substantial performance degradation was observed under the other alternative settings. These results indicate that the proposed model is not overly dependent on a single subjective weighting configuration. Meanwhile, the relatively favorable performance of settings with higher accuracy weights supports the reasonableness of assigning greater importance to response accuracy, because accuracy more directly reflects the correctness of perception, comprehension, and projection in safety-critical remote tower tasks.

Table 7. Sensitivity analysis results of SPAM weight settings.

Accuracy Weight	Reaction Time Weight	RMSE	MAE	R^2	PCC
0.5	0.5	0.0706	0.0495	0.8364	0.9187
0.6	0.4	0.0674	0.0472	0.8509	0.9271
0.7	0.3	0.0658	0.0461	0.8579	0.9308
0.8	0.2	0.0666	0.0468	0.8541	0.9289
0.9	0.1	0.0693	0.0487	0.8423	0.9224

4.4. Model Interpretability Analysis

To further reveal the decision basis of the proposed model, SHAP was employed to perform a global interpretability analysis of the prediction results of MGRF-Net. Figure 4 presents the SHAP summary beeswarm plot of the model, in which the horizontal axis indicates the magnitude of each feature's impact on the model output, with a larger absolute value representing higher importance. The color represents the feature value, where red denotes a high value and blue denotes a low value.

As shown in Figure 4, the high-contribution features are distributed across multiple physiological modalities, including ET, EEG, EDA, and ECG, indicating that the prediction of remote tower controllers' SA is not dominated by a single channel but instead results from the joint effect of multisource physiological information. Overall, Frontal Theta Power, Pupil Diameter, SGE, Parietal Beta Power, Blink Count, Mean RR Interval, and SCL rank among the most important features, suggesting that indicators related to visual attention allocation, brain functional activity, and autonomic nervous responses all play important roles in the model output, with ET and EEG features contributing particularly prominently.

A closer inspection of the color distribution on both sides of the SHAP axis further shows that the effect directions of different features are not consistent. For features such as Pupil Diameter, Parietal Beta Power, SCR Count, and Saccade Amplitude SD, higher feature values are more frequently distributed in the positive SHAP region, indicating that increases in these indicators generally make positive contributions to the predicted SA level. In contrast, higher values of Frontal Theta Power, SGE, Mean Fixation Duration, and SCR Amplitude are more concentrated in the negative SHAP region, suggesting that

increases in these features are more likely to be associated with negative contributions. This finding indicates that the model has learned not a simple and uniform linear physiological response pattern, but rather a differentiated joint influence mechanism with heterogeneous effect directions across modalities during SA formation.

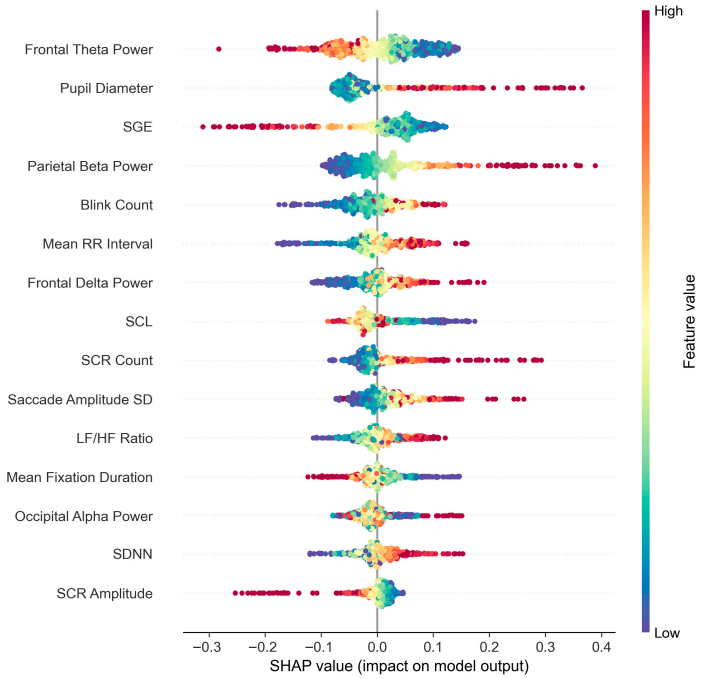


Figure 4. SHAP summary beeswarm plot of MGRF-Net for the SA prediction task.

The SHAP results demonstrate that the proposed model is capable of identifying key multimodal physiological variables with high contributions to SA prediction, and that the relationships between these features and the predicted outcome do not exhibit a simple linear structure but instead reflect a certain degree of nonlinearity and heterogeneity. This finding is consistent with the conclusions drawn from the modality ablation experiments and further confirms the necessity of multimodal fusion modeling. From an interpretability perspective, it also supports the rationality of the proposed model design.

5. Discussion

5.1. Main Findings

In this study, we proposed MGRF-Net, a multimodal graph relational fusion network for predicting remote tower controllers' SA from physiological data. Compared with conventional approaches that mainly rely on single-modal features or simple fusion strategies, the proposed model establishes a unified framework consisting of unimodal temporal representation learning, cross-modal relational modeling enhanced by graph attention, and gated fusion-based prediction. The overall results showed that MGRF-Net outperformed all comparison models in RMSE, MAE, R^2, and PCC, indicating that it can more effectively capture the continuous variation in SA. These findings suggest that, for SA prediction in remote tower operations, stronger general temporal modeling alone is not sufficient, and that effective modeling of cross-modal interactions is critical for further performance improvement.

An important reason for the effectiveness of the proposed model is that it does not treat multimodal inputs as simple parallel feature sets but explicitly incorporates the latent associations among modalities into the modeling process. After unimodal temporal encoding, the model constructs a sample-specific cross-modal relational matrix and further

performs information propagation and feature enhancement through graph attention. The module ablation results showed a progressive improvement from Relation to Relation + GAT and finally to the complete MGRF-Net, indicating that relational modeling, graph-based enhancement, and gated fusion contribute jointly to the final prediction performance. This also suggests that the advantage of the proposed model comes not only from using multiple modalities, but from organizing and exploiting the relationships among them more effectively.

To more intuitively understand this synergistic mechanism, the cross-modal relational matrix and the graph attention weight distribution within the model were further visualized, as shown in Figure 5.

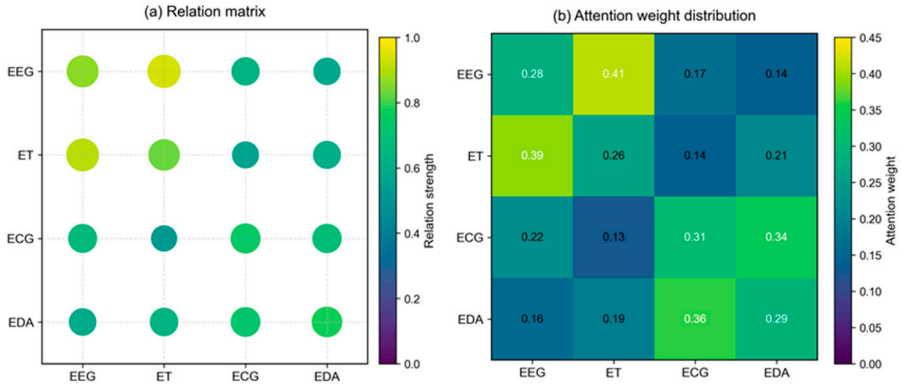


Figure 5. Cross-modal relational matrix and graph attention weight distribution.

The visualization results provide intuitive evidence that inter-modal interactions are not uniformly distributed but instead exhibit clear selectivity. As shown in Figure 5a, EEG and ET display relatively stronger relational strength, indicating tighter coupling between brain activity and eye movement behavior in SA modeling. In contrast, ECG and EDA form a comparatively stable auxiliary interaction pattern, suggesting that autonomic physiological responses provide complementary support. This relational pattern is consistent with the modality ablation results, which showed that removing EEG or ET led to more evident performance degradation, whereas ECG and EDA mainly played supportive roles.

Figure 5b further illustrates the graph attention weight distribution. It can be seen that attention is not evenly assigned across all modality connections but is more concentrated on modality pairs that are more relevant under the current state. In particular, the connections involving EEG and ET receive relatively higher attention weights, while ECG and EDA also form a more concentrated propagation pathway. This suggests that the graph attention mechanism does not simply aggregate all modality information in a fully connected manner; rather, it selectively emphasizes more task-relevant interactions under the structural guidance of the relational matrix, thereby reducing the influence of irrelevant or weakly related information on the final fused representation. This helps explain why the complete model achieves better performance than both the basic Relation variant and the version without gated fusion.

At the modality level, the results indicate that EEG, ET, ECG, and EDA all contribute to SA prediction, although their roles are not identical. The complete multimodal input achieved the best performance, confirming the complementarity among modalities. Further comparison showed that removing EEG or ET caused larger performance degradation, suggesting that brain activity and eye movement behavior are more directly related to SA variation in remote tower tasks. This result is consistent with the human factor characteristics of remote tower operations. Compared with conventional tower control, remote tower controllers obtain out-the-window information mainly through video-based displays and

must integrate aircraft movement, taxiway status, runway use, parking stand occupancy, and flight plan information across multiple interface areas. Therefore, SA formation in remote towers depends not only on visual detection, but also on visual attention allocation and cognitive integration of information from different display regions. This may explain why EEG and ET showed more prominent contributions and stronger cross-modal relationships in the proposed model. By contrast, ECG and EDA provide complementary information related to autonomic nervous activity and physiological arousal.

The SHAP-based interpretability analysis further supports these findings at the feature level. Important features were distributed across ET, EEG, ECG, and EDA rather than concentrated in a single modality, indicating that the model extracts discriminative information jointly from multiple channels. Among them, SGE, pupil diameter, blink count, and frontal theta power are particularly relevant to the remote tower context. SGE reflects gaze distribution across AOIs and corresponds to visual scanning and information sampling across multiple functional regions of the interface. Pupil diameter and frontal theta power are related to cognitive effort and information-processing demand, suggesting that maintaining SA under video-based remote monitoring requires continuous integration of dispersed visual and operational information. Blink count and fixation-related features may further reflect visual workload and monitoring stability during prolonged screen-based observation. These feature-level results indicate that SA prediction in remote towers is closely associated with visual search, attention allocation, and cognitive integration. In addition, the asymmetric distribution of SHAP values suggests that physiological features affect SA in nonlinear and heterogeneous ways across modalities. Combined with the visualization results in Figure 5, this nonlinear cross-modal coupling is precisely the mechanism that the proposed model is designed to capture.

Overall, the results suggest that remote tower controllers' SA has a clear multimodal physiological basis involving cognitive activity, visual attention allocation, and autonomic nervous responses. More importantly, the stronger roles of EEG and ET, together with the contribution of gaze distribution and cognitive-load-related features, indicate that SA prediction in remote tower operations is strongly shaped by the screen-based and multi-source information environment.

5.2. Limitations and Future Work

Although the proposed MGRF-Net showed promising performance in SA prediction, several limitations should be acknowledged. First, the SA labels used in this study were derived from SPAM, which remains an intermittent probe-based measurement method. Therefore, the constructed labels should be regarded as probe-centered approximations of SA rather than direct moment-to-moment ground truth measurements. Although a 30 s pre-probe physiological window was used to improve the temporal correspondence between physiological features and SPAM responses, temporal misalignment and label noise may still exist. Future studies may combine SPAM with operational logs, behavioral indicators, dynamic temporal alignment, label smoothing, and semi-supervised label refinement to improve the temporal precision and reliability of SA labeling.

Second, the participant composition was relatively limited and imbalanced in terms of experience level. The dataset included six experienced controller instructors and thirty controller trainees but did not include active frontline controllers. Although the mixed-subject modeling strategy reflects, to some extent, the heterogeneous personnel composition that may occur in practical remote tower deployment, the learned physiological SA mapping may still be influenced more strongly by trainee data. The exploratory experience-level analysis provided preliminary evidence of cross-experience applicability, but it is not sufficient to fully characterize the effect of operational experience on SA prediction. Future work

should collect more data from active controllers and expert controllers, and may introduce mixed-effect models, experience-aware modeling, or domain adaptation techniques to reduce potential experience-related bias.

Finally, future work may further explore the practical integration of the proposed model into safety-critical remote tower systems. MGRF-Net should not be regarded as an autonomous decision-making module or a replacement for certified ATC procedures. Instead, it may serve as a predictive engine for controller state monitoring within a human–machine collaborative assistance framework. In future applications, the model output could be integrated into the remote tower workstation as a non-intrusive SA risk indicator. When the predicted SA score falls below a predefined safety threshold, the system could trigger graded support measures, such as highlighting critical aircraft or taxiway conflicts, prioritizing relevant information windows, prompting confirmation of key clearances, or notifying a supervisor for additional monitoring. Such interventions should be designed as advisory and explainable support functions, and their thresholds, timing, and human–machine interaction logic must be validated against existing ATC safety requirements and operational standards before deployment.

6. Conclusions

This study proposed MGRF-Net, a prediction model based on multimodal physiological data for remote tower controllers' SA. To address the limitation of existing methods in insufficiently characterizing cross-modal interactions during heterogeneous physiological information fusion, a unified framework integrating unimodal temporal representation learning, graph attention-enhanced relational learning, and gated fusion-based prediction was developed. By jointly modeling multimodal physiological signals, including ET, EEG, ECG, and EDA, the proposed model can more effectively capture the coupling relationships among different modalities, thereby improving SA prediction performance.

Experimental results showed that the proposed model outperformed LightGBM, MLP, PatchTST, iTransformer, and TimeMixer in terms of RMSE, MAE, R^2, and PCC, demonstrating superior prediction accuracy and fitting capability. Further ablation experiments confirmed the effectiveness of the graph attention-enhanced relational learning and gated fusion mechanisms in improving model performance, while the modality ablation results indicated that multimodal input provides more comprehensive complementary information, with EEG and ET contributing more prominently. Meanwhile, the SHAP-based interpretability analysis further demonstrated that the model can identify key multimodal physiological variables closely related to SA, thereby supporting the rationality of the proposed model design.

The findings of this study indicate that combining multimodal physiological data with graph attention-enhanced relational learning and a gated fusion mechanism can effectively characterize the cross-modal interaction patterns underlying the formation of SA in remote tower controllers. This provides an effective data-driven approach for complex cognitive state prediction and also offers a new perspective for human-state monitoring and intelligent assistance in remote tower operations. Further validation with larger and more diverse operational datasets will help strengthen the generalizability and practical applicability of the proposed framework.

Author Contributions: Conceptualization, Q.Z. and R.L.; methodology, Q.Z.; validation, R.L. and Z.Y.; formal analysis, W.P.; investigation, R.L. and Z.Y.; resources, Q.Z. and W.P.; data curation, Z.Y.; writing—original draft preparation, R.L.; writing—review and editing, Q.Z.; visualization, Q.Z.; 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), the Open Fund Project of the Key Laboratory of Flight Techniques and Flight Safety CAAC (F2024KF11C), the Fundamental Research Funds for the Central Universities (25CAFUC03127), the Civil Aviation Education Talent Program Projects (MHJY2025002, MHJY2025003).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the Flight Technology and Flight Safety Research Base, 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. For requests to access the dataset, please contact the corresponding author at ruihanliang2750@163.com.

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

Abbreviations

The following abbreviations are used in this manuscript:

AOI	Area of Interest
ET	eye-tracking
EEG	electroencephalography
ECG	electrocardiography
EDA	electrodermal activity
HRV	Heart Rate Variability
LF/HF	Low Frequency/High Frequency
MAE	Mean Absolute Error
MGRF-Net	Multimodal Graph Relational Fusion Network
MLP	Multilayer Perceptron
MSE	Mean Squared Error
PatchTST	Patch Time Series Transformer
PCC	Pearson Correlation Coefficient
PERCLOS	Percentage of eyelid closure over the pupil over time
RMSE	Root Mean Square Error
R^2	Coefficient of Determination
SA	Situation Awareness
SART	Situation Awareness Rating Technique
SCL	Skin Conductance Level
SCR	Skin Conductance Response
SDNN	Standard Deviation of Normal-to-Normal intervals
SHAP	SHapley Additive exPlanations
SPAM	Situation Present Assessment Method
TCN	Temporal Convolutional Network

References

1. Schier-Morgenthal, S. Multiple Remote Tower Simulation Environment. In *Virtual and Remote Control Tower: Research, Design, Development, Validation, and Implementation*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 239–264. [CrossRef]
2. Kearney, P.; Li, W.-C. Multiple Remote Tower for Single European Sky: The Evolution from Initial Operational Concept to Regulatory Approved Implementation. *Transp. Res. Part Policy Pract.* **2018**, *116*, 15–30. [CrossRef]
3. Zhang, J.; Tian, X.; Pan, J.; Chen, Z.; Zou, X. A Field Study on Safety Performance of Apron Controllers at a Large-Scale Airport Based on Digital Tower. *Int. J. Environ. Res. Public Health* **2022**, *19*, 1623. [CrossRef]
4. Endsley, M.R. Measurement of Situation Awareness in Dynamic Systems. *Hum. Factors* **1995**, *37*, 65–84. [CrossRef]
5. Endsley, M.R. Situation Awareness in Aviation Systems. In *Handbook of Aviation Human Factors*; Lawrence Erlbaum Associates Publishers: Mahwah, NJ, USA, 1999; Volume 11, pp. 257–276.

6. Hauss, Y.; Eyferth, K. Securing Future ATM-Concepts' Safety by Measuring Situation Awareness in ATC. *Aerosp. Sci. Technol.* **2003**, *7*, 417–427. [CrossRef]
7. Taylor, R.M. Situational Awareness Rating Technique (SART): The Development of a Tool for Aircrew Systems Design. In *Situational Awareness*; Routledge: Oxfordshire, UK, 2017; pp. 111–128. [CrossRef]
8. Salmon, P.M.; Stanton, N.A.; Walker, G.H.; Jenkins, D.; Ladva, D.; Rafferty, L.; Young, M. Measuring Situation Awareness in Complex Systems: Comparison of Measures Study. *Int. J. Ind. Ergon.* **2009**, *39*, 490–500. [CrossRef]
9. Zhang, T.; Yang, J.; Liang, N.; Pitts, B.J.; Prakah-Asante, K.; Curry, R.; Duerstock, B.; Wachs, J.P.; Yu, D. Physiological Measurements of Situation Awareness: A Systematic Review. *Hum. Factors J. Hum. Factors Ergon. Soc.* **2023**, *65*, 737–758. [CrossRef] [PubMed]
10. Li, W.-C.; Zhang, J.; Le Minh, T.; Cao, J.; Wang, L. Visual Scan Patterns Reflect to Human-Computer Interactions on Processing Different Types of Messages in the Flight Deck. *Int. J. Ind. Ergon.* **2019**, *72*, 54–60. [CrossRef]
11. Kästle, J.L.; Anvari, B.; Krol, J.; Wurdemann, H.A. Correlation between Situational Awareness and EEG Signals. *Neurocomputing* **2021**, *432*, 70–79. [CrossRef]
12. Yeo, L.G.; Sun, H.; Liu, Y.; Trapsilawati, F.; Sourina, O.; Chen, C.-H.; Mueller-Wittig, W.; Ang, W.T. Mobile EEG-Based Situation Awareness Recognition for Air Traffic Controllers. In *Proceedings of the 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*; IEEE: New York, NY, USA, 2017; pp. 3030–3035. [CrossRef]
13. Shaffer, F.; Ginsberg, J.P. An Overview of Heart Rate Variability Metrics and Norms. *Front. Public Health* **2017**, *5*, 258. [CrossRef]
14. Wei, H.; Zhuang, D.; Wanyan, X.; Wang, Q. An Experimental Analysis of Situation Awareness for Cockpit Display Interface Evaluation Based on Flight Simulation. *Chin. J. Aeronaut.* **2013**, *26*, 884–889. [CrossRef]
15. Wang, Q.; Zhuang, D.; Wei, H.; Wanyan, X. Evaluation Method of Cockpit Display Interface Based on Situation Awareness. In *Proceedings of the 2012 5th International Conference on Biomedical Engineering and Informatics*; IEEE: New York, NY, USA, 2012; pp. 528–531. [CrossRef]
16. Pan, W.; Liang, R.; Wang, Y.; Song, D.; Yin, Z. Situational Awareness Prediction for Remote Tower Controllers Based on Eye-Tracking and Heart Rate Variability Data. *Sensors* **2025**, *25*, 2052. [CrossRef] [PubMed]
17. Yang, J.; Liang, N.; Pitts, B.J.; Prakah-Asante, K.O.; Curry, R.; Blommer, M.; Swaminathan, R.; Yu, D. Multimodal Sensing and Computational Intelligence for Situation Awareness Classification in Autonomous Driving. *IEEE Trans. Hum.-Mach. Syst.* **2023**, *53*, 270–281. [CrossRef]
18. 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]
19. Du, G.; Zhang, L.; Su, K.; Wang, X.; Teng, S.; Liu, P.X. A Multimodal Fusion Fatigue Driving Detection Method Based on Heart Rate and PERCLOS. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 21810–21820. [CrossRef]
20. Liang, R.; Pan, W.; Zuo, Q.; Zhang, C.; Chen, S.; Chen, S.; Deng, L. Modeling Visual Fatigue in Remote Tower Air Traffic Controllers: A Multimodal Physiological Data-Based Approach. *Aerospace* **2025**, *12*, 474. [CrossRef]
21. Yang, S.; Li, M.; Wang, J. Fusing sEMG and EEG to Increase the Robustness of Hand Motion Recognition Using Functional Connectivity and GCN. *IEEE Sens. J.* **2022**, *22*, 24309–24319. [CrossRef]
22. Zhao, H.; Wang, Y.; Duan, J.; Huang, C.; Cao, D.; Tong, Y.; Xu, B.; Bai, J.; Tong, J.; Zhang, Q. Multivariate Time-Series Anomaly Detection via Graph Attention Network. In *Proceedings of the 2020 IEEE International Conference on Data Mining (ICDM)*; IEEE: New York, NY, USA, 2020; pp. 841–850. [CrossRef]
23. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P.S. A Comprehensive Survey on Graph Neural Networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4–24. [CrossRef]
24. Jin, M.; Koh, H.Y.; Wen, Q.; Zambon, D.; Alippi, C.; Webb, G.I.; King, I.; Pan, S. A Survey on Graph Neural Networks for Time Series: Forecasting, Classification, Imputation, and Anomaly Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 10466–10485. [CrossRef]
25. Song, T.; Zheng, W.; Song, P.; Cui, Z. EEG Emotion Recognition Using Dynamical Graph Convolutional Neural Networks. *IEEE Trans. Affect. Comput.* **2018**, *11*, 532–541. [CrossRef]
26. 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]
27. Loft, S.; Morrell, D.B.; Huf, S. Using the Situation Present Assessment Method to Measure Situation Awareness in Simulated Submarine Track Management. *Int. J. Hum. Factors Ergon.* **2013**, *2*, 33–48. [CrossRef]
28. Krejtz, K.; Szmidt, T.; Duchowski, A.T.; Krejtz, I. Entropy-Based Statistical Analysis of Eye Movement Transitions. In *Proceedings of the Symposium on Eye Tracking Research and Applications*; Association for Computing Machinery: New York, NY, USA, 2014; pp. 159–166. [CrossRef]
29. Delorme, A.; Makeig, S. EEGLAB: An Open Source Toolbox for Analysis of Single-Trial EEG Dynamics Including Independent Component Analysis. *J. Neurosci. Methods* **2004**, *134*, 9–21. [CrossRef]
30. Henry, J.C. Electroencephalography: Basic Principles, Clinical Applications, and Related Fields. *Neurology* **2006**, *67*, 2092. [CrossRef]

31. Perakakis, P. *HEPLAB: A Matlab Graphical Interface for the Preprocessing of the Heartbeat-Evoked Potential*, v1.0.1; Zenodo: Geneva, Switzerland, 2019.
32. Benedek, M.; Kaernbach, C. A Continuous Measure of Phasic Electrodermal Activity. *J. Neurosci. Methods* **2010**, *190*, 80–91. [CrossRef] [PubMed]
33. Hui, L.; Pei, Z.; Quan, S.; Ke, X.; Zhe, S. Cognitive Workload Detection of Air Traffic Controllers Based on mRMR and Fewer EEG Channels. *Brain Sci.* **2024**, *14*, 811. [CrossRef]
34. Bhavsar, P.; Srinivasan, B.; Srinivasan, R. Quantifying Situation Awareness of Control Room Operators Using Eye-Gaze Behavior. *Comput. Chem. Eng.* **2017**, *106*, 191–201. [CrossRef]
35. Navia, J.A.; Ibáñez-Gijón, J.; Travieso, D.; Montes, A.; de Frutos, P.L. Eye-Tracking Metrics for Estimating Workload and Characterizing Errors in Conflict Detection and Resolution during Simulated En Route Air-Traffic Control. *Front. Psychol.* **2025**, *16*, 1644721. [CrossRef] [PubMed]
36. Koskelo, J.; Lehmusaho, A.; Laitinen, T.P.; Hartikainen, J.E.; Lahtinen, T.M.; Leino, T.K.; Huttunen, K. Cardiac Autonomic Responses in Relation to Cognitive Workload during Simulated Military Flight. *Appl. Ergon.* **2024**, *121*, 104370. [CrossRef] [PubMed]
37. Lutnyk, L.; Rudi, D.; Schinazi, V.R.; Kiefer, P.; Raubal, M. The Effect of Flight Phase on Electrodermal Activity and Gaze Behavior: A Simulator Study. *Appl. Ergon.* **2023**, *109*, 103989. [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.