CTA-STGT: A Cortical Topology-Aware Spatiotemporal Graph Transformer for EEG-Based eVTOL Pilot Cognitive Workload Estimation

Chenglin Liu[†]
Thrust of Intelligent Transportation
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
cliu549@connect.hkust-gz.edu.cn

Wenjun Zhang[†]
Southwest Jiaotong University
Chengdu, China
zwj726@my.swjtu.edu.cn

Zhijie Yi
Thrust of Intelligent Transportation
The Hong Kong University of Science and Technology (Guandzhou)
Guangzhou, China
yzj@mail.bnu.edu.cn

Jiyao Wang
McGill University
Montreal, Quebec, Canada
jiyao.wang@mcgill.ca

Xiang Chang
Cornell Tech
Cornell University
New York, New York, USA
xc529@cornell.edu

Dengbo He[*]
Thrust of Intelligent Transportation
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
Division of Emerging Interdisciplinary Areas
The Hong Kong University of Science and Technology
Hong Kong, China
dengbohe@hkust-gz.edu.cn

Abstract

As urban air mobility (UAM) advances toward operational deployment, electric vertical takeoff and landing (eVTOL) pilots may face heightened cognitive demands from dense low-altitude traffic, strict energy constraints, and safety-critical flight phases with complex control-mode transitions. Real-time pilot state monitoring is therefore essential for safe operations. Prior electroencephalography (EEG)-based workload-estimation models for pilots predominantly focus on temporal dynamics, while insufficiently modeling inter-channel dependencies, which are crucial for effective spatial aggregation across EEG electrodes. In this study, we propose the Cortical Topology-Aware Spatiotemporal Graph Transformer (CTA-STGT), which explicitly encodes structured inter-channel relationships via a topology-aware graph convolutional network constrained by a cortical-topology-derived adjacency matrix. Temporal patch embeddings and spatial graph embeddings are projected into a unified latent space and then jointly fused by a Transformer encoder to capture spatiotemporal dependencies. Evaluated on EEG data from a 6-degree-of-freedom eVTOL simulator experiment with three workload levels based on NASA Task Load Index (NASA-TLX) scores, CTA-STGT achieved an accuracy of 0.89 and an F1-score of 0.89, outperforming the baseline models. These results provide a technical basis for adaptive workload-mitigation support in high-workload eVTOL scenarios.

CCS Concepts

• **Networks** → *Layering*; *Deep packet inspection*; • **Computing methodologies** → **Supervised learning**.

Keywords

electroencephalography (EEG), cognitive workload estimation, cortical topology, graph Transformer, eVTOL

ACM Reference Format:
Chenglin Liu, Wenjun Zhang, Zhijie Yi, Jiyao Wang, Xiang Chang, and Dengbo He. 2026. CTA-STGT: A Cortical Topology-Aware Spatiotemporal Graph Transformer for EEG-Based eVTOL Pilot Cognitive Workload Estimation. In *18th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI Adjunct '26), September 20–23, 2026, Gothenburg, Sweden.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3828158.3838270

1 Introduction

As urban air mobility (UAM) advances toward real-world deployment, electric vertical takeoff and landing (eVTOL) cockpits impose substantial cognitive demands on pilots. Dense low-altitude traffic, micro-meteorological uncertainty, and multiple flight modes, including takeoff/climb, transition, cruise, and landing, compete for pilots' attentional resources [2]. The transition phase is particularly challenging because propulsion reconfiguration can abruptly change control dynamics, increasing workload and error risk. Therefore, real-time cognitive state monitoring is essential for future intelligent cockpits of eVTOLs.

[*]Corresponding author.
[†]Chenglin Liu and Wenjun Zhang contributed equally to this work.



This work is licensed under a Creative Commons Attribution 4.0 International License.
AutomotiveUI Adjunct '26, Gothenburg, Sweden
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2815-0/26/09
https://doi.org/10.1145/3828158.3838270

Electroencephalography (EEG) is regarded as a reliable neurophysiological measure of mental workload [11]. In aviation scenarios, EEG sensors can be integrated into pilot headsets to support relatively non-invasive and unobtrusive signal acquisition [10]. However, existing deep learning methods mainly focus on temporal EEG dynamics while insufficiently modeling spatial topological relationships among electrodes.

Stage-wise modeling, in which temporal and spatial features are learned in separate modules and then fused [4, 5, 8], may help reveal interactions between temporal and spatial features. However, such designs may defer direct interactions between temporal and spatial representations until the fusion stage. In late-fusion dual-branch architectures, spatial topology and temporal dynamics are modeled separately and concatenated only before classification.

To address these limitations, we propose CTA-STGT. The model explicitly captures structured inter-channel relationships using a lightweight graph convolutional network (GCN) branch based on electrode spatial topology, while temporal patch embeddings and spatial GCN embeddings are projected into a unified token space. A joint-fusion Transformer is then employed to model cross-token interactions and long-range spatiotemporal dependencies, enabling effective representation of pilots' cognitive workload.

The main contributions of this study are as follows:
(1) A topology-aware spatial modeling strategy based on graph convolution is proposed to explicitly model structured inter-channel relationships in EEG signals;
(2) A joint-fusion mechanism is developed to facilitate enhanced spatiotemporal interaction at the token level, enabling a more comprehensive representation of EEG characteristics;
(3) A 6-DOF flight-simulator-based eVTOL experiment was designed to elicit distinct cognitive workload levels and collect 32-channel pilot EEG data. The resulting dataset was used to train and evaluate CTA-STGT, which achieved an accuracy of 88.80% in multilevel workload classification using five-fold cross-validation with within-subject data partitioning.

2 Methodology

2.1 Problem Formulation

Let a preprocessed EEG segment be $X \in \mathbb{R}^{B\times C\times T}$, where B is the batch size, C represents the number of EEG channels, and T denotes the number of samples within a sliding time window. Let $A \in \mathbb{R}^{C\times C}$ be the adjacency matrix that characterizes the spatial topological relationships among electrodes. The objective is to construct a parameterized mapping function $f_\theta(X, A) \to y$, which learns the transformation from the spatiotemporal EEG representations to workload levels, where y denotes the discrete cognitive state labels.

2.2 Electrode Spatial Topology Graph Construction

Each EEG channel is modeled as a graph node, and the graph $G = (V, E)$ is constructed using an adjacency matrix A defined by the spatial topology of the electrodes. Let the initial node features be $H(0) \in \mathbb{R}^{C\times d}$, where d denotes the feature dimension. The graph convolution operation is defined as

$$H^{(l+1)} = \sigma\left(AH^{(l)}W^{(l)}\right), \tag{1}$$

where $W^{(l)}$ is the learnable weight matrix of the l-th graph convolution layer, $\sigma(\cdot)$ denotes a nonlinear activation function, and A is the normalized adjacency matrix. Through the message-passing mechanism of graph convolution, the resulting representations integrate local spatial topological information, enabling topology-aware modeling of spatial dependencies in EEG signals.

2.3 Temporal and Spatial Token Embedding

Two lightweight branches extract local features before fusion:

Temporal Patch Embedding: The temporal branch models dynamic temporal dependencies by splitting the input sequence into non-overlapping patches of length P. A combination of pointwise and multiscale depthwise 1D convolutions is employed to capture local patterns at different temporal resolutions. After feature fusion and downsampling, each patch is embedded into a D-dimensional token with temporal positional encoding.

Spatial GCN Embedding: The spatial branch captures inter-channel dependencies by first compressing temporal information via 1D convolution. A graph convolutional network (GCN) [3, 7] is employed to propagate information over the graph defined by the adjacency matrix, enabling feature aggregation among neighboring channels. Each channel representation is then projected into a D-dimensional token with spatial positional encoding.

2.4 Joint Spatiotemporal Transformer with Fusion Mechanism

After concatenation, the temporal and spatial tokens form a unified sequence that is fed into a multi-head self-attention module to perform joint spatiotemporal fusion. Self-attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{2}$$

Specifically, both temporal and spatial tokens are jointly projected into queries (Q), keys (K), and values (V), and attention weights are computed over the entire concatenated sequence. Since Q, K, and V are all derived from the unified spatiotemporal token sequence, each temporal token can attend to all spatial tokens, and vice versa. Following the multi-head self-attention, each token is processed by a position-wise feed-forward block implemented as a two-layer multilayer perceptron (MLP) with Gaussian error linear unit (GELU) activation and dropout. Both the attention block and the feed-forward block are wrapped with residual connections and employ pre-layer normalization, with dropout applied to the sub-layer outputs to improve training stability. This architecture enables cross-dimensional information exchange during feature modeling, allowing spatiotemporal dependencies to be fully integrated at the representation-learning level.

2.5 State Classification Head

After the joint-fusion Transformer, global average pooling is applied along the token dimension to obtain a unified spatiotemporal representation. This representation is then fed into a three-layer MLP classifier to enhance discriminative capability. The model outputs

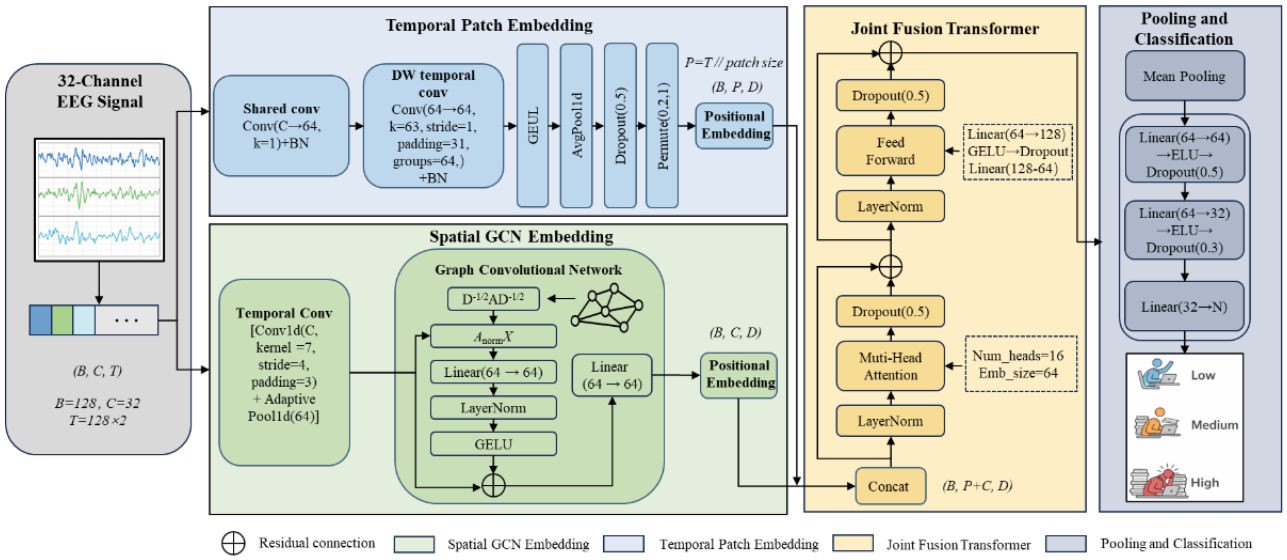


Figure 1: Overall architecture of the proposed CTA-STGT with joint spatiotemporal fusion.

Table 1: Summary of CTA-STGT modules.

Module Name	Details	Core Value
Input Layer	32-channel EEG (10–20 system)	Provides multi-channel electrophysiological signals representing distributed cortical activity.
Graph Construction	Adjacency matrix based on electrode spatial topology	Encodes structured inter-channel relationships and models non-Euclidean spatial dependencies.
Temporal Embedding	Patch embedding via 1D convolution and pooling	Extracts localized temporal patterns and reduces sequence length for efficient modeling.
Spatial Embedding	Topology-aware graph convolutional network (GCN) with residual connection and linear projection	Captures inter-channel spatial interactions and enhances topology-constrained representation learning.
Fusion Layer	Joint fusion of spatial and temporal embeddings	Enables unified spatiotemporal representation and facilitates cross-domain interaction.
Encoder	Joint Transformer with multi-head self-attention	Models long-range dependencies and structured spatiotemporal correlations across unified tokens.

class logits and is trained with cross-entropy loss:

$$\mathcal{L} = -\sum_{c=1}^{C_s} y_c \log(\hat{y}_c), \tag{3}$$

where C_s is the number of cognitive states, y_c is the one-hot label indicator for class c, and $\hat{y}_c$ is the predicted probability for class c.

2.6 Overall Network Structure

The CTA-STGT architecture is shown in Fig. 1. Multi-channel EEG signals are processed by parallel temporal and spatial branches, and the resulting tokens are fed into a joint-fusion Transformer to learn a unified spatiotemporal representation. Global average pooling and an MLP classifier are then used for cognitive state prediction. Module details are provided in Table 1.

3 Experiments and Results

3.1 Experimental Setup

Participants: A total of 33 participants without any flight experience including simulators, drones, or real flights were recruited for this study.To ensure baseline operational competence, all participants successfully completed the comprehensive tiltrotor eVTOL training curriculum proposed by Liu et al. prior to the formal experiment [9]. This ensured that participants possessed the requisite skills to execute complex missions safely.

Experimental Setup: The experimental platform featured a 6-degree-of-freedom (6-DOF) flight simulator. The visual system comprised three 43-inch displays providing a 150° horizontal field of view, supplemented by two 10.4-inch monitors for flight instruments (see Fig. 2(a)). Participants wore a Neuroelectrics Enobio 32-channel dry-electrode EEG system to record continuous EEG



Figure 2: (a) 6-degree-of-freedom (6-DOF) flight simulator and 32-channel EEG system; (b) Joby S4 eVTOL aircraft.

data at 500 Hz while piloting a simulated Joby S4 eVTOL aircraft (see Fig. 2(b)) using a standard throttle and joystick configuration.

Task Scenarios: Participants completed a highly realistic simulated intercity Regional Air Mobility (RAM) task (see Fig. 3). We designed the flight phases to evoke different levels of pilot workload, a strategy widely used in civil aviation [10].

Data Collection: EEG was recorded continuously throughout the entire flight, covering all three flight phases. Subjective workload data were collected separately for the three flight phases. Immediately after each phase, the experiment was paused, and participants completed the NASA Task Load Index (NASA-TLX) [14] based on their experience during the previous phase. The next phase began after the assessment was completed. This procedure was used to capture participants' immediate workload perceptions and reduce potential recall bias.

3.2 Data Processing

Raw 32-channel EEG data from each flight phase were first imported, and abnormal recordings were excluded. The signals were then processed using band-pass filtering (1–45 Hz) and notch filtering to remove DC drift, electromyographic (EMG) noise, high-frequency noise, and power-line interference. Artifacts were further reduced through bad-segment rejection, bad-channel identification, interpolation, and independent component analysis (ICA)-based removal of ocular and muscle components. Channels showing persistently poor signal quality due to inadequate electrode–scalp contact were identified as bad channels and subsequently interpolated. For each participant, approximately 2 min of resting-state EEG data were recorded before the experiment. EEG signals were then segmented into non-overlapping 2-s windows at 250 Hz and normalized to z scores using the pre-experiment resting-state EEG as the baseline. After preprocessing, a total of 42,312 EEG samples were obtained.

3.3 Statistical Results and Analysis

To examine whether the predefined flight phases corresponded to distinct subjective workload levels, NASA-TLX scores were first compared across the three flight phases. The descriptive statistics are presented in Table 2.

Table 2: Descriptive statistics of NASA-TLX scores across flight phases.

Flight phase	N	Mean	SD	Median	IQR
Takeoff/climb	33	50.21	16.20	51.83	19.67
Cruise	33	52.64	16.39	52.33	21.67
Descent/landing	33	63.63	18.47	68.67	26.17

Note: SD denotes standard deviation, and IQR denotes interquartile range.

Given the bounded NASA-TLX scores and the within-participant repeated-measures design, the Friedman test was selected as a conservative distribution-free omnibus test to evaluate overall phase-related differences. The results showed a significant difference in NASA-TLX scores across the three flight phases, $\chi^2(2) = 33.152$, $p < 0.01$, with Kendall's $W = 0.502$, indicating a moderate effect of flight phase on subjective workload. Consistent with this non-parametric omnibus analysis, post-hoc paired comparisons were performed using Wilcoxon signed-rank tests with Holm correction for multiple comparisons. The results are summarized in Table 3.

The pairwise comparison results indicated that NASA-TLX scores did not differ significantly between the takeoff/climb and cruise phases after Holm correction. This suggests that, in the present simulated eVTOL flight task, the takeoff/climb and cruise phases did not exhibit the expected distinction between medium and low workload levels commonly assumed in conventional civil aviation [15]. In contrast, the descent/landing phase showed significantly higher NASA-TLX scores than both the takeoff/climb phase and the cruise phase, indicating that descent/landing imposed the highest workload among the three phases.

3.4 Workload Labeling

Based on the statistical results, this study did not directly use the conventional flight-phase definitions as workload labels. Instead, a data-driven workload labeling strategy was constructed based on NASA-TLX scores. Specifically, hierarchical agglomerative clustering was performed on the phase-specific weighted NASA-TLX scores using Euclidean distance and Ward's linkage criterion [14], yielding three clusters representing low, medium, and high workload levels (see Fig. 4(a)). The distribution across the three workload

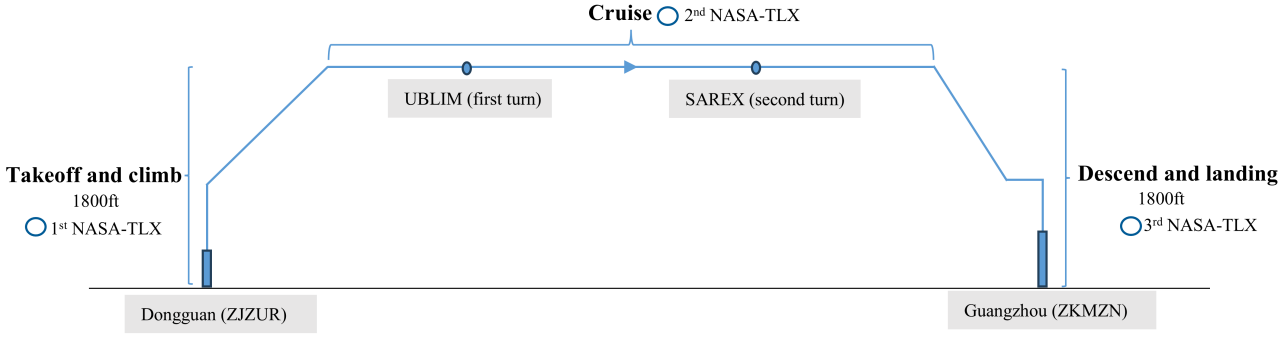


Figure 3: Diagram showing flight phase segmentation, mission task elements, and data collection points. Note that "ZJZUR", "UBLIM", "SAREX", and "ZKMZN" are location codes.

Table 3: Pairwise comparisons of NASA-TLX scores across flight phases.

Comparison	Mean difference (descriptive)	Wilcoxon signed-rank W	Holm-adjusted p
Cruise vs. takeoff/climb	+2.42	227.0	0.35
Descent/landing vs. takeoff/climb	+13.42	21.0	< 0.001
Descent/landing vs. cruise	+11.00	12.0	< 0.001

levels was as follows: 9,822 samples for low workload, 17,392 samples for medium workload, and 15,098 samples for high workload. The Wilcoxon signed-rank tests showed significant differences in mean NASA-TLX scores among the three clusters ($p < 0.05$) (see Fig. 4(b)). Thus, these clusters were subsequently used as the ground-truth labels for model training and evaluation.

3.5 Classification Performance Comparison

To comprehensively evaluate the proposed model, six models for EEG signal classification were selected as baseline models, including EEGNet [6], IFNet [13], MSCFormer [17], ATCNet [1], CTNet [16], and EEGWaveNet [12]. Five-fold cross-validation with within-subject data partitioning was employed to evaluate model performance. In addition, five widely adopted evaluation metrics were used to provide a comprehensive assessment of model performance [15].

Table 4 demonstrates that CTA-STGT outperforms all competing methods across all evaluation metrics. It achieves an accuracy of 0.89, surpassing IFNet's accuracy of 0.85 and improving upon EEGNet by 14.00 percentage points. In addition, CTA-STGT attains the highest precision, sensitivity, and specificity, reaching 0.89, 0.89, and 0.94, respectively, indicating strong discriminative capability and robust classification performance. Its F1-score of 0.89 further confirms a favorable balance between precision and recall.

3.6 Ablation Study

To assess the contribution of topology-aware spatial modeling and joint fusion, we ablated branch composition and fusion strategy, as shown in Table 5. "Spatial-only (without GCN)" denotes the spatial branch with the GCN module removed, whereas "Spatial-only (with GCN)" retains the topology-aware GCN module. Let $Z_T \in \mathbb{R}^{B\times N_T\times D}$ and $Z_S \in \mathbb{R}^{B\times N_S\times D}$ denote the temporal and spatial tokens, respectively. After projection to the same embedding dimension, joint fusion concatenates them along the token dimension, yielding $Z_0 = [Z_T; Z_S] \in \mathbb{R}^{B\times(N_T+N_S)\times D}$. The combined sequence is processed by a shared Transformer encoder, followed by global average pooling and the MLP classifier. Late fusion processes the two branches separately and concatenates their pooled features before classification. Decision fusion uses branch-specific classifiers and averages their logits.

As a result, the joint model achieved an accuracy of 88.80% and an F1-score of 89.00%, exceeding the strongest non-joint variant, late fusion, by 3.55 and 3.47 percentage points, respectively. Adding the GCN to the spatial-only branch yielded descriptive gains of 1.19 points in accuracy and 1.10 points in F1-score. These results are consistent with the expected effects of topology-aware spatial modeling and joint token processing under the evaluation protocol with within-subject data partitioning; however, the reported differences should not be interpreted as statistically significant because no inferential comparison across repeated runs was conducted.

4 Conclusions

4.1 Limitations and Future Work

Several limitations should be noted. First, as the model was evaluated using five-fold cross-validation with within-subject data partitioning, the current results should not be interpreted as evidence of generalization to unseen participants. Future directions include self-supervised pretraining to reduce labeling costs and improve model generalization to unseen participants. Second, our study was conducted in a flight simulator, which may bias participants' behavior relative to real-world flight conditions. Future research should validate the model in operational eVTOLs before it can be deployed in eVTOL cockpits.

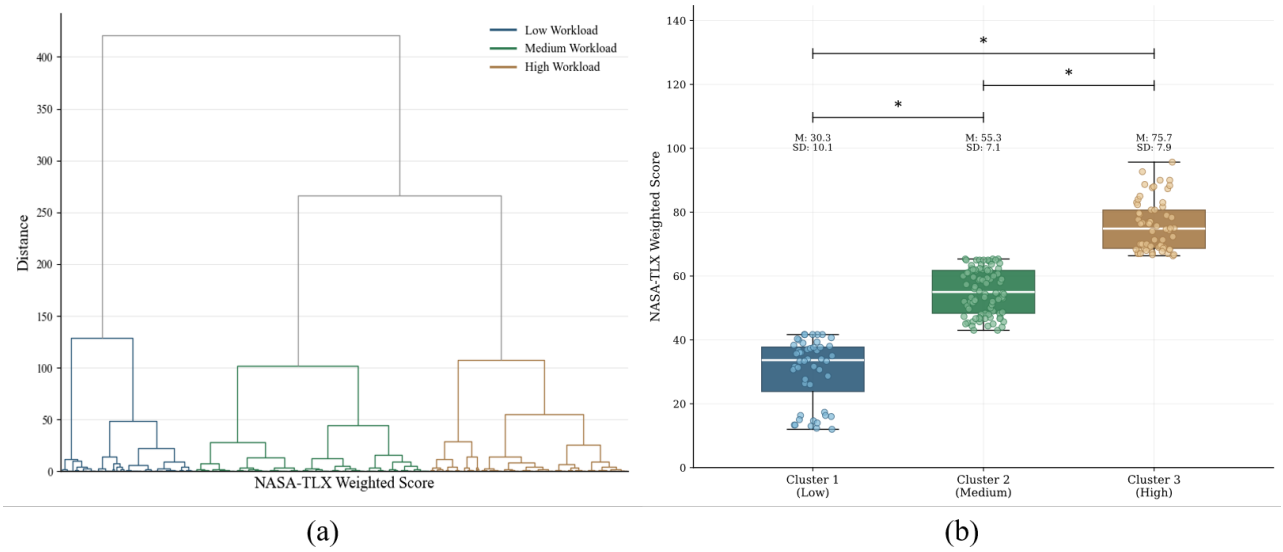


Figure 4: (a) Hierarchical clustering dendrogram of weighted NASA-TLX scores; (b) comparison of weighted NASA-TLX scores among clusters, with asterisks indicating significant differences ($p < 0.05$).

Table 4: Model performance on three-level cognitive workload classification.

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-score (%)
EEGNet	74.80	77.99	74.27	85.42	75.63
IFNet	85.25	86.81	85.37	92.06	85.86
MSConformer	79.95	81.55	80.01	88.81	80.53
ATCNet	82.91	85.40	82.31	90.70	83.44
CTNet	81.99	83.02	82.30	90.46	82.50
EEGWaveNet	82.53	83.00	82.79	90.39	82.83
CTA-STGT	**88.80**	**89.26**	**88.77**	**94.08**	**89.00**

4.2 Summary

This study proposes CTA-STGT, a joint-fusion spatiotemporal graph Transformer that incorporates cortical topology for real-time pilot cognitive state classification. The model achieves an accuracy of 0.89 and an F1-score of 0.89 in simulated eVTOL tasks. It outperforms baseline EEG processing models and provides interpretable attention patterns that are important for cockpit safety. Beyond workload classification, CTA-STGT may provide a basis for adaptive cockpit interaction. Its workload estimates could inform simplified interface presentation and trigger workload-mitigation support when sustained high workload is detected, providing a foundation for future workload-aware cockpit design.

References

[1] Hamdi Altaheri, Ghulam Muhammad, and Mansour Alsulaiman. 2023. Physics-Informed Attention Temporal Convolutional Network for EEG-Based Motor Imagery Classification. *IEEE Transactions on Industrial Informatics* 19, 2 (2023), 2249–2258. doi:10.1109/TII.2022.3197419

[2] Marcin Piotr Biernacki and Rafał Lewkowicz. 2024. The role of visual conditions and aircraft type on different aspects of pilot workload. *Applied Ergonomics* 118 (2024), 104268.

[3] Arezoo Hamidi and Kourosh Kiani. 2025. Motor Imagery EEG signals classification using a Transformer-GCN approach. *Applied Soft Computing* 170 (2025), 112686.

[4] Fo Hu, Kailun He, Can Wang, Qinxu Zheng, Bin Zhou, Gang Li, and Yu Sun. 2025. STRFLNet: Spatio-Temporal representation fusion learning network for EEG-Based emotion recognition. *IEEE Transactions on Affective Computing* (2025).

[5] Xueyu Jia, Yonghao Song, Lie Yang, and Longhan Xie. 2022. Joint spatial and temporal features extraction for multi-classification of motor imagery EEG. *Biomedical signal processing and control* 71 (2022), 103247.

[6] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. 2018. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of neural engineering* 15, 5 (2018), 056013.

[7] Guangqiang Li, Ning Chen, and Jing Jin. 2022. Semi-supervised EEG emotion recognition model based on enhanced graph fusion and GCN. *Journal of Neural Engineering* 19, 2 (2022), 026039.

[8] Benke Liu, Yongxiong Wang, Zhe Wang, Xin Wan, and Chenguang Li. 2025. A temporal-spatial feature fusion network for emotion recognition with individual differences reduction. *Neuroscience* 569 (2025), 195–209.

[9] Chenglin Liu, Yuxin Du, Dengbo He, Haoyuan Jiang, Jiyao Wang, Zhijie Yi, Yufan Chen, Chaozhe Jiang, and Hai Yang. [n. d.]. Training Ab Initio eVTOL Pilots for Regional Air Mobility Missions: A Comparison of Curricula. *Available at SSRN 5971712* ([n. d.]).

[10] Chenglin Liu, Chenyang Zhang, Luohao Sun, Kun Liu, Haiyue Liu, Wenbing Zhu, and Chaozhe Jiang. 2023. Detection of pilot's mental workload using a wireless EEG headset in airfield traffic pattern tasks. *Entropy* 25, 7 (2023), 1035.

[11] Siwei Ma, Xuedong Yan, Jac Billington, Natasha Merat, and Gustav Markkula. 2024. Cognitive load during driving: EEG microstate metrics are sensitive to task difficulty and predict safety outcomes. *Accident Analysis & Prevention* 207 (2024), 107769.

[12] Punnawish Thuwajit, Phurin Rangpong, Phattarapong Sawangjai, Phairot Autthasan, Rattanaphon Chaisaen, Nannapas Banluesombatkul, Puttaranun Boonchit, Nattasate Tatsaringkansakul, Thapanun Sudhawiyangkul, and Theerawit Wilaiprasitporn. 2021. EEGWaveNet: Multiscale CNN-based spatiotemporal feature extraction for EEG seizure detection. *IEEE transactions on industrial informatics* 18, 8 (2021), 5547–5557.

[13] Jiaheng Wang, Lin Yao, and Yueming Wang. 2023. IFNet: An interactive frequency convolutional neural network for enhancing motor imagery decoding from EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 31 (2023), 1900–1911.

[14] Nusrat Zerin Zenia. 2025. Measuring NASA-TLX Factors of the Cognitive Workload: An Objective Approach. (2025).

[15] Chenyang Zhang, Chaozhe Jiang, Yuanyi Xie, Shi Cao, Jiajun Yuan, Chuang Liu, Weiwei Cao, and Yaohua Li. 2024. Assessing pilot workload during takeoff and climb under different weather conditions: a fNIRS-based modeling using deep learning algorithms. *IEEE Trans. Aerospace Electron. Systems* 61, 2 (2024), 1705–1724.

[16] Wei Zhao, Xiaolu Jiang, Baocan Zhang, Shixiao Xiao, and Sujun Weng. 2024. CTNet: a convolutional transformer network for EEG-based motor imagery classification. *Scientific reports* 14, 1 (2024), 20237.

[17] Wei Zhao, Baocan Zhang, Haifeng Zhou, Dezhi Wei, Chenxi Huang, and Quan Lan. 2025. Multi-scale convolutional transformer network for motor imagery brain-computer interface. *Scientific reports* 15, 1 (2025), 12935.

Table 5: Ablation study of fusion strategies and branches.

Model Variants						Evaluation Metrics				
Branch			Fusion Strategy							
Temporal	Spatial-only (without GCN)	Spatial-only (with GCN)	Late	Decision	Joint (ours)	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-score (%)
✓	–	–	–	–	–	77.05	80.17	76.85	87.55	77.89
–	✓	–	–	–	–	81.48	81.89	80.98	89.62	81.71
–	–	✓	–	–	–	82.67	84.44	81.76	90.24	82.81
✓	–	✓	✓	–	–	85.25	86.35	84.90	92.61	85.53
✓	–	✓	–	✓	–	81.53	83.65	81.14	90.03	82.12
✓	–	✓	–	–	✓	**88.80**	**89.26**	**88.77**	**94.08**	**89.00**