Keeping pilots in the loop: An explainable spatiotemporal EEG-driven deep learning framework for adaptive automation in cruising flight phase

Cho Yin Yiu, Kam K.H. Ng, *Member, IEEE*, Qinbiao Li, and Xin Yuan

***Abstract*—Automation has been extensively used in flight operations, so pilots are less involved in actual flight control. With the long idle time during cruising, pilots may have their vigilance level reduced and eventually become out-of-the-loop. This research proposes a two-stage explainable adaptive automation approach to keep pilots in the loop based on Convolutional Neural Networks, Long Short-Term Memory, and EEG data collected from 24 participants in a one-hour simulator-based flight task in each level of automation. Our proposed spatiotemporal model yields test accuracy of 0.9918 and 0.9907 in the first and second stages, respectively, outperforming other benchmarking models by 30.79% and 10.73%, respectively. Furthermore, the Shapley additive explanations are adopted to strengthen the model interpretability and trustworthiness for safety-critical applications. Our model successfully identified that high delta and theta waves with low beta and gamma waves contribute positively to the out-of-the-loop state. It indicates that the classification aligns with the theoretical background and is trustworthy. The trustworthy adaptive deep learning model supports the dynamical automation configuration for improving human-automation collaboration in cruising flights.**

***Index Terms*—Adaptive automation; CNN-LSTM; EEG; human-automation teaming; spatiotemporal deep learning**

I. INTRODUCTION

A. Automation and human performance in flight operations

MANY aviation accidents are attributed to the poor interaction between the operators and automation [1], including the notable Asiana Airlines Flight 214 and Air France Flight 447. In Asiana Airlines Flight 214, pilots relied heavily on autopilot (A/P) and auto-thrust (A/THR). The reduced vigilance caused inadequate monitoring of the aircraft flight profile, thus leading to improper speed management. Indeed, as automation works well during normal scenarios [2], modern aircraft are designed with a high level of automation (LOA) to relieve the pilot's workload during flying [3]. As these automation techniques are generally considered reliable [4], pilots usually place a high level of trust in automation [5]. However, it might result in automation misuse, such as complacency [6].

Automation complacency refers to non-vigilance due to an unjustified assumption of a satisfactory system state and self-satisfaction [7]. In flight operations, it is usually characterised by the pilot's overreliance on automation to fly, as automation is designed to assist pilots instead of replacing the roles of pilots [8, 9], while pilots reduce their vigilance and leave automation as the main operator in most real-life cases. When pilots over-rely on automation, they may not be able to identify the potential faults of automation. For instance, A/P and/or A/THR may be disengaged for various reasons, like airspeed discrepancies in Air France Flight 447. To take over from automation and continue manual flying, pilots shall understand the situation well. However, pilots are considered "out-of-the-loop" (OOTL) due to the long idling time during cruising flights [10]. As with the life-threatening consequences of human OOTL accidents, automation shall be carefully designed to ensure flight safety by minimising the possibility of human operators becoming OOTL.

Knowing that a high LOA cannot guarantee error-free operations, especially during non-normal situations [11], human intelligence from pilots is necessary for taking over from automation in non-normal operations to mitigate the likelihood of catastrophic errors [12]. Nevertheless, the trust and reliance on automation may impede pilots from understanding the situation correctly and result in erroneous decisions. Indeed, the problem of mind wandering and human operators becoming OOTL has been well-recognised in the literature. Bainbridge [13] first proposed the ironies of automation and claimed that automation designers can never automate certain tasks. While the highly automated design of aircraft turns automation into the main 'controller' of the aircraft, pilots remain necessary to monitor the work done by automation. As the last resort to provide appropriate control inputs when automation fails, the monitoring task cannot be done by automation. Therefore, the automation designer still

Manuscript received 25 February 2024. This research is supported by Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University, Hong Kong SAR. Our gratitude is also extended to the Research Committee of the Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University for support of the project (RLPA, CE1G, BDWV, RJX2). Cho Yin Yiu is a recipient of the Hong Kong PhD Fellowship (Reference number: PF21-62058).

This study has been granted human ethics approval from the PolyU Institutional Review Board of the Hong Kong Polytechnic University (IRB Reference Number: HSEARS20210318002). *(Corresponding author: Kam K.H. Ng).*

Cho Yin Yiu, Kam K.H. Ng, Qinbiao Li, and Xin Yuan are with Human Factors and Ergonomics Laboratory, Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong SAR, China (e-mail: james-cho-yin.yiu@connect.polyu.hk, kam.kh.ng@polyu.edu.hk, qinbiao-leo.li@polyu.edu.hk, xin-olivia.yuan@connect.polyu.hk).

Data would be made available on request.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org

leaves room for human errors, even though flight operations are highly automated. Indeed, monitoring automation is monotonous and mentally demanding, causing a reduction in vigilance and human performance [14, 15]. With reduced vigilance, pilots may perform irrelevant tasks and feel drowsy [16, 17]. Eventually, such design may impede pilots from determining whether automation acts correctly, understanding how automation works, or even what automation is working on for making appropriate decisions [18].

B. Motivations

Various accidents illustrated in the previous section demonstrate the significance of human-automation interaction (HAI) and collaborative decisions to achieve safer flights. Indeed, Rothfuß, et al. [19] revealed that the collaborative decisions between humans and automation outperform that of either humans or automation. Hence, it is imperative to enhance human-automation teaming for safer flights, especially with the surging air traffic flow [20]. Berberian, et al. [10] first suggested adopting neuroergonomics like electroencephalography (EEG) to address human OOTL with a better understanding of the underlying cognitive mechanisms [21]. Neuroergonomics has a wide range of applications in enhancing human performance in aviation [22-24]. However, most research does not focus on the problem brought about by pilot-autopilot disconnect. Instead, most human OOTL research is theoretical and demands empirical evidence for practical insights [25, 26]. There is a lack of rigorous studies and neuroergonomics-driven automation designs that enhance HAI in flight operations. Therefore, novel technologies shall be designed based on neuroergonomics to promote and support teaming between humans and automation.

Neuroergonomics serve as an objective evaluation method to assess the human operator's performance in real-time. Most of the neuroergonomic data, such as EEG, can be streamed to various deep learning models to identify diverse human mental states, such as situational awareness (SA), drowsiness, and mental workload [27]. With a proper tuning of its network structure and hyperparameters, deep learning models learn from complex neuroergonomic data to achieve a high accuracy in classifying human mental states, which may hardly be done by conventional models. Among various deep learning approaches, Convolutional Neural Networks (CNN) can capture the rich spatial information in EEG across the brain, where different convolution modules may also affect the prediction accuracy. Meanwhile, Long Short-Term Memory (LSTM) can capture the temporal relationship of human OOTL in time-series EEG recordings for a more accurate and trustworthy prediction. Therefore, deep learning with neuroergonomics can possibly address human OOTL through timely identification from its EEG patterns and subsequent remedial actions, which deserves further investigation.

C. Objectives and contributions

The current research aims to bridge the pilots and automation and keep the pilots in the control loop through neuroergonomics. By mitigating the monotonous and mind-wandering episodes, pilots should be able to maintain their situational awareness and quickly take over from automation in non-normal situations. It can be achieved by tuning the LOA to an optimal level that balances human OOTL and task performance. However, a cognitive workload trade-off is expected as pilots shall take more control. Therefore, in some cases, task performance may not be optimised with the reduction in LOA. To yield the best overall task performance, this paper proposes an adaptive automation approach based on EEG spectral features and an explainable hybrid architecture with CNN and LSTM to keep pilots in the loop.

TABLE I

EXPERIMENT SCENARIOS

Scenario	Autopilot (A/P)	Auto-thrust (A/THR)
FAF	Available	Available
PAF	Available	Disengaged since cruising
MF	Disengaged since cruising	Disengaged since cruising

This research presents an advancement of HAI in flight operations through deep learning techniques considering the neurophysiological perspective of pilots. In terms of theoretical contribution, this research captures the underlying cognitive mechanisms with EEG to reflect human performance objectively. Based on the classified mental state of pilots, both human and task performance can be optimised dynamically by tuning the LOA. This approach advances HAI by addressing the limitations of a static LOA throughout the entire cruising flight. Moreover, an explainable artificial intelligence (XAI) approach facilitates the understanding of brain regions activated during different LOAs, which strengthens the trustworthiness of the model. With an explainable model, this study is original in providing a trustworthy adaptive automation solution in cruising operations rather than merely a black-box solution.

D. Organisation of the paper

The rest of the paper is organised as follows: **Section II** reviews the related works. **Section III** describes the dataset, and **Section IV** illustrates the proposed deep learning-based adaptive automation model. **Section V** presents and discusses the results. Finally, the conclusions, limitations, and future works are described in **Section VI.**

II. Literature Review

A. Human-automation interaction

Automation is not designed to replace humans but to interact for efficient operations in complex systems like aircraft [8]. Indeed, automation and artificial intelligence have many limitations yet to be overcome [28], so human operators remain essential and take responsibility for potential accidents [29]. However, many automation designs hinder human operators from properly overseeing and interacting with it for operational safety [30]. These designs might result in human OOTL that operators fail to respond to events appropriately, automation confusion that operators fail to understand automation, and automation interaction problems that operators cannot direct automation timely. Among them, human OOTL accounts for 58% of aircraft accidents related to automation. Indeed, human operators under a fully automated system merely monitor the work of automation, which is generally considered monotonous. Their vigilance reduces with time and may work on mission-irrelevant tasks [17], which impede them from responding to events timely [10]. The consequences of human OOTL evidenced that automation shall be adequately designed to facilitate its interaction with human operators to ensure system safety.

Different automation designs consist of different function allocations between operators and automation, resulting in different LOA: the degree to which the operators and automation interact. Therefore, it is critical to search for the optimal LOA for HAI. Past research shows that an increase in LOA reduces situational awareness (SA) [31-33] and task performance [34], which suggests full automation might not be the optimal solution. Onnasch, et al. [11] further suggested that the optimal LOA is a cost-benefit trade-off problem: When the LOA exceeds a critical boundary, the negative consequences of automation result. Therefore, developing flexible and dynamic LOA control strategies is essential to optimise the overall performance. In this regard, adaptive and adaptable automation is proposed to enhance the performance of the human-automation team [35]. The LOA of the previous is changed by the system based on certain metrics, while that of the latter is changed by human operators.

Sauer, et al. [36] highlighted the advantages of flexible automation through task allocation compared to static automation that balances mental workload between human and automation. Indeed, EEG engagement index was used in an adaptive automation system by comparing the experimental and baseline recording [37]. However, as EEG contains various spatial and temporal information, the complex EEG patterns shall be further captured using advanced deep learning or transfer learning methods to classify different mental or operation state [38]. In current research, these advanced methods shall improve adaptive automation decisions so humans can interact with automation more seamlessly.

B. Deep learning with EEG in transportation

Neuroergonomics have been widely applied to improve human performance in transport operations [22, 39]. Among various neuroergonomic tools, EEG has been the most adopted neuroimaging tool given its ease of implementation, high temporal resolution, spatial information retrieval, and the ability for time-frequency analysis [40]. Its spectral data provide information on the human mental state of drowsiness, attention, concentration, etc. [41]. With the emergence of deep learning and enhancement in computational performance [42], various EEG features have been utilised as the input features of deep learning classification models [43], such as event-related potentials (ERP) [44] and band power [45]. These models aim to diagnose and predict the mental states that may result in potentially unsafe conditions for preventive measures.

In flight operations, Wu, et al. [46] proposed a deep contractive sparse auto-encoder network to identify the fatigue status of pilots using EEG signals. They further improved their fatigue classification accuracy with an adversarial deep Bayesian neural network (BNN) [47]. Other than flight operations, several studies utilise EEG to identify the occurrence of various human factor issues in transport operations. Fatigue is one of the most studied constructs using EEG and deep learning [48]. Han, et al. [27] constructed a deep CNN with EEG data to classify human operators into four mental states: distraction, workload, fatigue, and normal. CNNs were also deployed to detect the microsleep of drivers with nearly 90% testing accuracy [49, 50]. Other than mental state detection, Angkan, et al. [51] designed a CNN-driven brain-computer interface (BCI) for driver cognitive load assessment. Furthermore, the temporal information brought by time-series EEG data are also of great significance. The review of Mi, et al. [52] mentioned that applying LSTM on driver EEG data can preserve their behavioural features for a longer time period. The LSTM layer can also be appended after convolutional modules to capture both spatial and temporal information from EEG data, including classification of abnormal states of pilots [53] and mental fatigue recognition [54]. These studies demonstrate the possibility of using deep learning and EEG for mental state assessment. Nonetheless, there is a dearth of studies that employ image-based EEG data to capture its spatial information.

C. Model interpretability and its significance in safety-critical domains

While deep learning has showcased impressive performance with neuroergonomic features, its black-box nature is yet to be resolved. The decisions made by deep learning are purely data-driven and usually non-transparent, which may hardly be understood by human beings [55]. A wrong prediction can result in a wrong recommendation, which may have irreversible consequences or even result in casualties in safety-critical domains like aviation [56]. Therefore, accountability and reliability are essential before deep learning models are considered safe to apply in real life. To achieve this, the predictions made by deep learning models shall be associated with explanations to establish their trustworthiness [57]. In this regard, XAI is an emerging research field that warrants attention for real-life applications of data-driven models. Among different XAI techniques, SHAP is a game theoretic approach that utilises Shapley values to explain the reasoning behind a model's prediction for a specific instance and the contributions of its predictors [58].

D. Research gap

In the era of automated flight, optimising the interaction between pilots and automation is indispensable to achieve safer flight operations. While there are initiatives for adaptive and adaptable automation, its integration with complex physiological patterns using deep learning needs further investigation. Although numerous studies employ deep learning and EEG data for mental state assessment, little

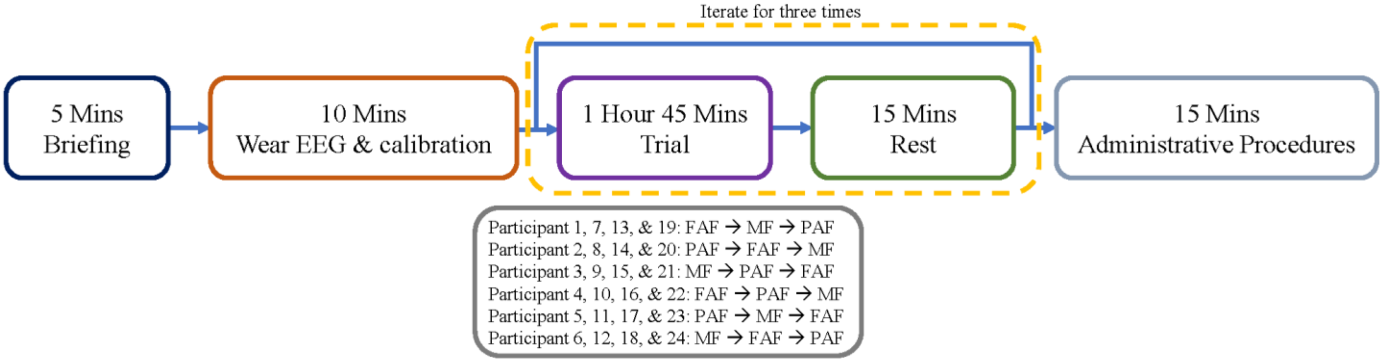

Iterate for three times
5 Mins
Briefing
10 Mins
Wear EEG & calibration
1 Hour 45 Mins
Trial
15 Mins
Rest
15 Mins
Administrative Procedures
Participant 1, 7, 13, & 19: FAF → MF → PAF
Participant 2, 8, 14, & 20: PAF → FAF → MF
Participant 3, 9, 15, & 21: MF → PAF → FAF
Participant 4, 10, 16, & 22: FAF → PAF → MF
Participant 5, 11, 17, & 23: PAF → MF → FAF
Participant 6, 12, 18, & 24: MF → FAF → PAF

Fig. 1. Timeline of the experiment. (Readers may refer to TABLE I for details of each flight condition, FAF, PAF, and MF.)

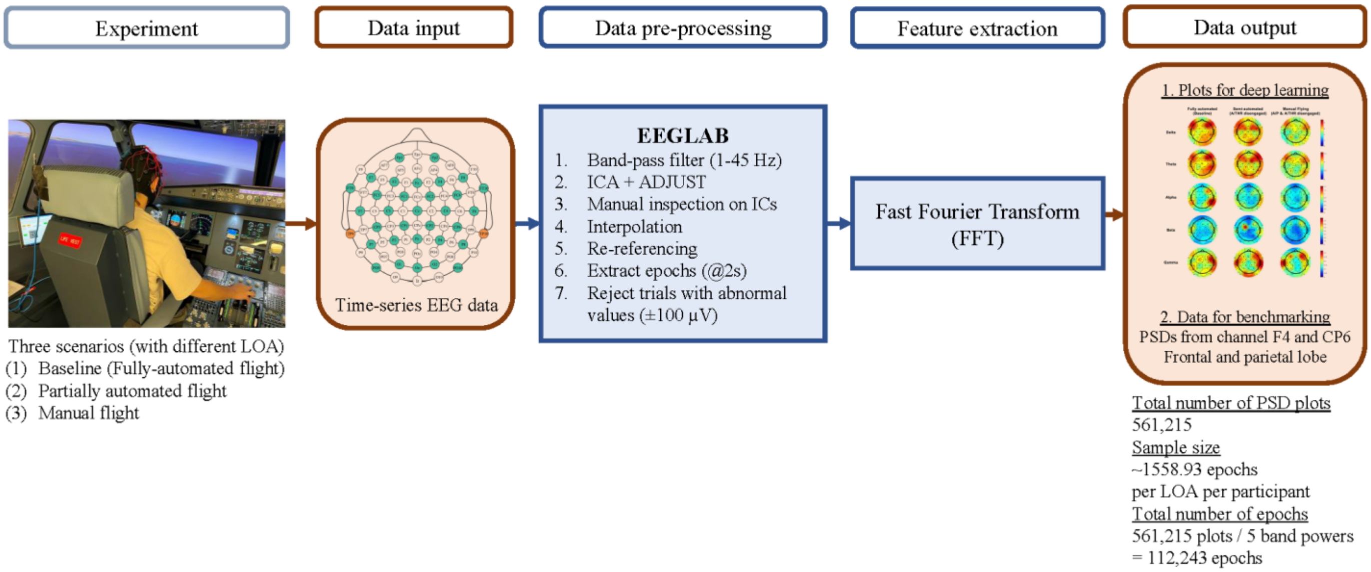

Experiment
Data input
Data pre-processing
Feature extraction
Data output
Three scenarios (with different LOA)
(1) Baseline (Fully-automated flight)
(2) Partially automated flight
(3) Manual flight
Time-series EEG data
EEGLAB
1. Band-pass filter (1-45 Hz)
2. ICA + ADJUST
3. Manual inspection on ICs
4. Interpolation
5. Re-referencing
6. Extract epochs (@2s)
7. Reject trials with abnormal values (±100 μV)
Fast Fourier Transform (FFT)
1. Plots for deep learning
2. Data for benchmarking
PSDs from channel F4 and CP6
Frontal and parietal lobe
Total number of PSD plots
561,215
Sample size
~1558.93 epochs
per LOA per participant
Total number of epochs
561,215 plots / 5 band powers
= 112,243 epochs

Fig. 2. Graphical illustration of data processing. (The arrows represent the data flow from raw data to processed data.)

research has focused on addressing the consequence of automation and employing image-based EEG data as the model input. The interpretability of the model also remains a critical hindrance for these 'black-box' models to be applied in real-life scenarios, as models shall be trustworthy to maintain aviation safety. Therefore, this paper is particularly interested in developing an EEG-driven explainable deep learning model to facilitate adaptive automation in flight operations for mitigating human OOTL.

III. Dataset

A. Data collection

The dataset adopted in this study was collected through an experiment conducted on an Airbus A320 flight simulator in the *Human Factors and Ergonomics Laboratory, Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University*. The platform is constructed based on [59], with the display of flight simulation changed to a 220-degree curved projection screen to enhance the fidelity.

1) Experiment setting: Participants are asked to operate a flight between Hong Kong International Airport (ICAO: VHHH) and Fukuoka Airport (ICAO: RJFF). Each experiment consists of three scenarios with different LOAs. These scenarios include baseline – fully automated flight (FAF), partially automated flight (PAF), and manual flight (MF).

The cruising phase of the flight starts at waypoint CONGA (around 40 minutes after the flight starts from the gate; for inducing human OOTL for pilots during a long idling time). Upon reaching cruising altitude, A/P and/or A/THR are disengaged in PAF and MF scenarios. Finally, the scenario terminates at waypoint MOLKA, so the cruising time is around 60 minutes. TABLE I describes the three experiment scenarios. The scenario sequence ($3! = 6$ in total) is iterated with participant number to mitigate the effects of task sequence and is single blinded before cruising. Fig. 1 shows the timeline of the experiment.

2) Participants and measures: Twenty-four participants were recruited from the student pilots enrolled in the Cadet Pilot program of a local airline. They have 213.9 simulator hours on average. Pre-experiment ethical approval was granted by the PolyU Institutional Review Board (Reference number: HSEARS20210318002). Written consent was obtained before the experiment, and participants received coupons as research incentives.

During the experiment, participants are required to wear a saline-driven EEG headset EMOTIV EPOC Flex. This EEG headset has 32 channels arranged according to the International 10-10 system. It captures most brain regions,

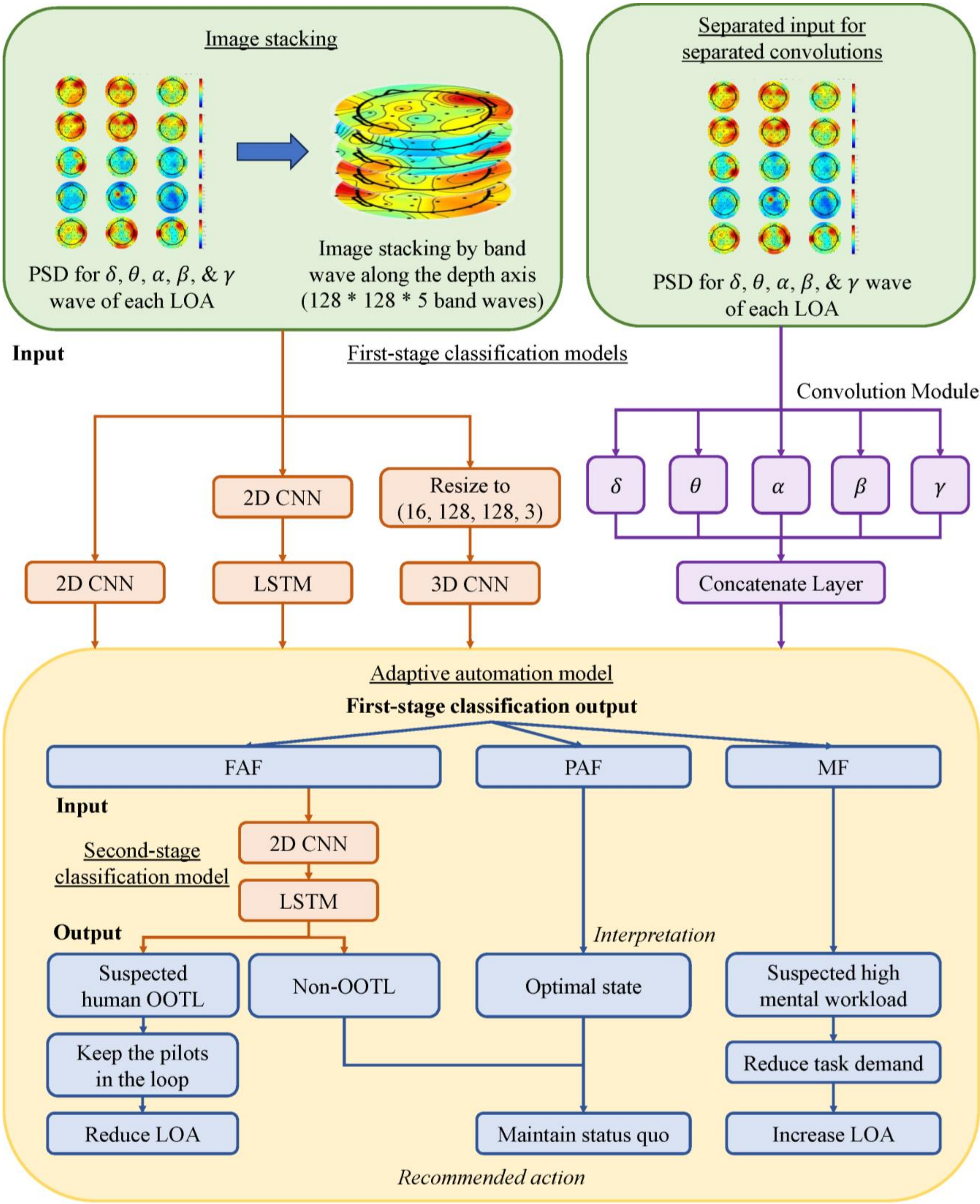


Image stacking
PSD for δ, θ, α, β, & γ wave of each LOA
Image stacking by band wave along the depth axis (128 * 128 * 5 band waves)
Separated input for separated convolutions
PSD for δ, θ, α, β, & γ wave of each LOA
Input
First-stage classification models
Convolution Modules
δ
θ
α
β
γ
2D CNN
Resize to (16, 128, 128, 3)
2D CNN
LSTM
3D CNN
Concatenate Layer
Adaptive automation model
First-stage classification output
FAF
PAF
MF
Input
Second-stage classification model
2D CNN
LSTM
Output
Interpretation
Suspected human OOTL
Non-OOTL
Optimal state
Suspected high mental workload
Keep the pilots in the loop
Reduce task demand
Reduce LOA
Maintain status quo
Increase LOA
Recommended action

Fig. 3. Two-stage adaptive automation model design

including the prefrontal cortex, frontal lobe, temporal lobe, parietal lobe, and occipital lobe. The recording starts before each scenario begins to avoid unnecessary interruption to the participants, while it terminates when the scenario ends at waypoint MOLKA. However, only the cruising phase is analysed as this study focuses on the OOTL problem during cruising. Therefore, EEG markers are placed by the researchers on a computer recording EEG when the flight reaches waypoint CONGA for post-recording analysis. In total, we have 24 60-minute EEG recordings for each LOA.

A key aspect to assess whether pilots are "in the loop" is whether they can response accurately and timely to certain stimuli. Therefore, we measured the reaction time of the participants with a Microsoft Surface Pro 7. The screen was initially black and will turn red with a calvary charge sound periodically. Participants were requested to touch the monitor to terminate the stimulus, which the time elapsed to stimulus termination was measured as the reaction time. Furthermore, after each experiment scenario, participants were asked to complete a NASA Task Load Index (TLX) questionnaire with ratings of six dimensions, including mental demand, temporal demand, physical demand, performance, effort, and frustration. A pairwise comparison was performed on each dimension pair for a weighted NASA-TLX score. These measures were used to cross-validate with the LOA-based labels, which will be discussed in **Section IV(A)**.

B. Data processing

Raw EEG data are usually pre-processed for eye, muscle, and head movement artefacts removal and processed in the time or frequency domain for interpretation. EEGLAB 2023.1 is used [60]. Fig. 2 shows the data processing workflow.

In this study, we pre-process our data based on [61] using Python 3.7.9 with GNU Octave 6.4.0. To begin, EEG markers are used to select data collected from the cruising flight phase. Then, we apply a bandpass filter of 1-45 Hz to filter only meaningful frequencies. Next, we conducted a 32-component Independent Component Analysis (ICA) to identify EEG artefact components with ADJUST 1.1.1 for automatic artefact rejection. Manually inspection is conducted to ensure that bad components are removed from the data. Then, each recording is re-referenced to the average and separated into two-second epochs, which is considered appropriate given its ability to reflect the pilot’s response. We further identify abnormal values in EEG recording by rejecting outlier epochs with values exceeding ±100 μV [60]. Finally, there is an average of 1,558.93 epochs per recording after artefact removal. We utilise the Fast Fourier Transform (FFT) to calculate the power spectral densities (PSD, unit: $\mu V^2/Hz$) of each epoch in the frequency domain for spectral analysis [62]. While there are many other feature extraction methods like wavelet transform and deep learning, FFT serves as a robust and straightforward approach to extract relevant frequency-domain features from EEG signals. Wavelet transform requires a careful selection of mother wavelets and parameters, while deep learning feature extraction cannot guarantee all features are correctly labelled due to possible classification error, which may impact the development of subsequent adaptive automation models. Most importantly, the implications of each FFT-based features are well-established in neuroscience literature and facilitate the interpretation of pilot’s mental state. Hence, FFT is used for feature extraction to compute the PSDs for each band power.

PSDs are then plotted topographically for each band wave and epoch in MATLAB R2021b using EEGLAB 2023.1. Each plot has a width and depth of 128 pixels with three channels (RGB). To normalise the plots across different participants, the scale of the colour bars is set as the same for each band wave. The minimum value of the colour bar scale (in blue) is set as 0, as there shall be no negative PSDs. The maximum value (in red) is set to its maximum value across all participants as it differs across band waves (Delta: 8.7709 $\mu V^2/Hz$, Theta: 1.4963 $\mu V^2/Hz$, Alpha: 1.3491 $\mu V^2/Hz$, Beta: 0.7489 $\mu V^2/Hz$, Gamma: 0.3742 $\mu V^2/Hz$). There are 112,243 epochs in total for 24 participants. It is equivalent to 561,215 plots, as there are five band waves (delta, theta, alpha, beta, gamma). Each epoch is labelled with the scenario (LOA) in which the data is collected. The label is used for classification into different mental states based on the LOA. It supports adaptive automation decisions as the classified label allows reducing LOA when pilots start mind wandering or increasing LOA when pilots are overloaded during manual flying. Further discussions of how the dataset supports the adaptive automation model are denoted in **Section IV(A)**. Furthermore, we also selected data from two channels for benchmarking, including CP6 from the parietal lobe and F4 from the frontal lobe. Details of the performance comparison using non-image classification models are provided in **Section IV(E)**.

TABLE II

RESULTS OF ONE-WAY REPEATED MEASURES ANOVA ON REACTION TIME AND NASA-TLX

Item	F		p	η_p^2
Reaction Time*	(1.58,36.45) = 10.26		0.001	0.308
Scenario	M (s)	Scenario	M (s)	p
FAF	1007.8 (213.78)	PAF	858.3 (122.47)	< 0.001
FAF	1007.8 (213.78)	MF	991.1 (134.96)	1.000
PAF	858.3 (122.47)	MF	991.1 (134.96)	< 0.001
Item	F		p	η_p^2
NASA-TLX^	(1.45, 33.41) = 45.92		< 0.001	0.666
Scenario	M (s)	Scenario	M (s)	p
FAF	24.0 (19.40)	PAF	30.2 (12.31)	0.177
FAF	24.0 (19.40)	MF	59.1 (18.36)	< 0.001
PAF	30.2 (12.31)	MF	59.1 (18.36)	< 0.001

Note – *: Huynh-Feldt corrected, ^: Greenhouse-Geisser corrected.

IV. EEG-DRIVEN ADAPTIVE AUTOMATION MODEL

This section presents the proposed EEG-driven adaptive automation model, including its design to achieve adaptive automation, configuration of spatiotemporal deep learning, and interpretation techniques.

A. Adaptive automation design

This research aims to develop an adaptive automation model based on the pilot’s EEG spectral features. Such an adaptive design can suit the pilot’s mental state and need dynamically to yield the optimal overall performance. The problem is formulated as a two-stage classification task, where the first stage is a three-class classification problem, and the second stage is a binary classification problem. Fig. 3 shows the adaptive automation model design using a conceptual flowchart.

1) First-stage model: In the first stage, we first classify the pilot's mental state based on EEG PSD plots from three different LOAs with different influences on operator's performance. The operator's performance is assessed based on the reaction time and the NASA-TLX using one-way repeated measures ANOVA. The results are listed in TABLE II.

From TABLE II, the reaction time of pilots (in seconds) at different LOA was significantly different ($p = 0.001$). The reaction time in the PAF was significantly shorter than that of the FAF and the MF (both $p < 0.001$), while the difference in reaction time between the FAF and the MF was statistically nonsignificant ($p = 1.000$). On the other hand, the NASA-TLX score was significantly different between LOA ($p < 0.001$). Compared to MF, a significant lower NASA-TLX is observed in the FAF and the PAF (both $p < 0.001$), while the difference between FAF and PAF are statistically nonsignificant ($p = 0.177$).

Thus, the results indicated that FAF has a longer reaction time with a generally lower cognitive workload, while MF has significantly higher cognitive workload but a longer reaction time. The PAF yielded a relatively short reaction time with a relatively low cognitive workload, which is an optimal state for operations. Coupling with [19]'s finding that the best performance is observed in human-automation collaboration, our model assumes PAF as the optimal overall (task and human) performance state. Conversely, MF is known to be demanding to pilots and will increase the reaction time [63]. Hence, when the pilot's mental state is classified as MF, the LOA can be increased adaptively to reduce the task demand of the human operators. However, pilots may experience mind wandering episodes and become OOTL in FAF, but pilots may not necessarily become OOTL in FAF. Therefore, we utilise a second-stage model to further classify the FAF mental states with OOTL and non-OOTL.

2) Second-stage model: In the second stage, we further classify whether the pilot being classified in FAF state is active or OOTL. We utilise the data collected from FAF as the training data. The data were clustered using a two-component Gaussian Mixture Model (GMM) based on its associated reaction time for an objective labelling of the OOTL and the non-OOTL state. The trials in the cluster of shorter reaction time are labelled as non-OOTL, while the others with longer reaction time are labelled as OOTL.

Hence, if the pilot is being classified as FAF in the first stage, the second stage model further classifies whether the pilot is OOTL. In the OOTL case, the LOA can be reduced to keep the pilots in the loop. Meanwhile, the active (non-OOTL) cases remain unchanged like in PAF.

B. Convolutional neural networks

As the dataset is topographical plots of PSD recorded from EEG, we propose to use a CNN as the backbone (classifier) of this proposed model. CNN is a feedforward neural network that extracts spatial features from data using several convolutional layers. Each convolutional layer performs convolution operations, enabling CNN to automatically learn and extract meaningful features from the input data [64]. Filters are used to slide over the input image to perform convolutions to capture spatial patterns and local relationships between neighbouring pixels for generating feature maps. Then, pooling layers down-sample the feature maps to reduce dimensionality and extract the most salient features [65]. Finally, based on the learned representations from the previous layers, fully connected layers integrate the extracted features and make predictions [66].

TABLE III

GRID SEARCH MATRIX FOR NETWORK ARCHITECTURE

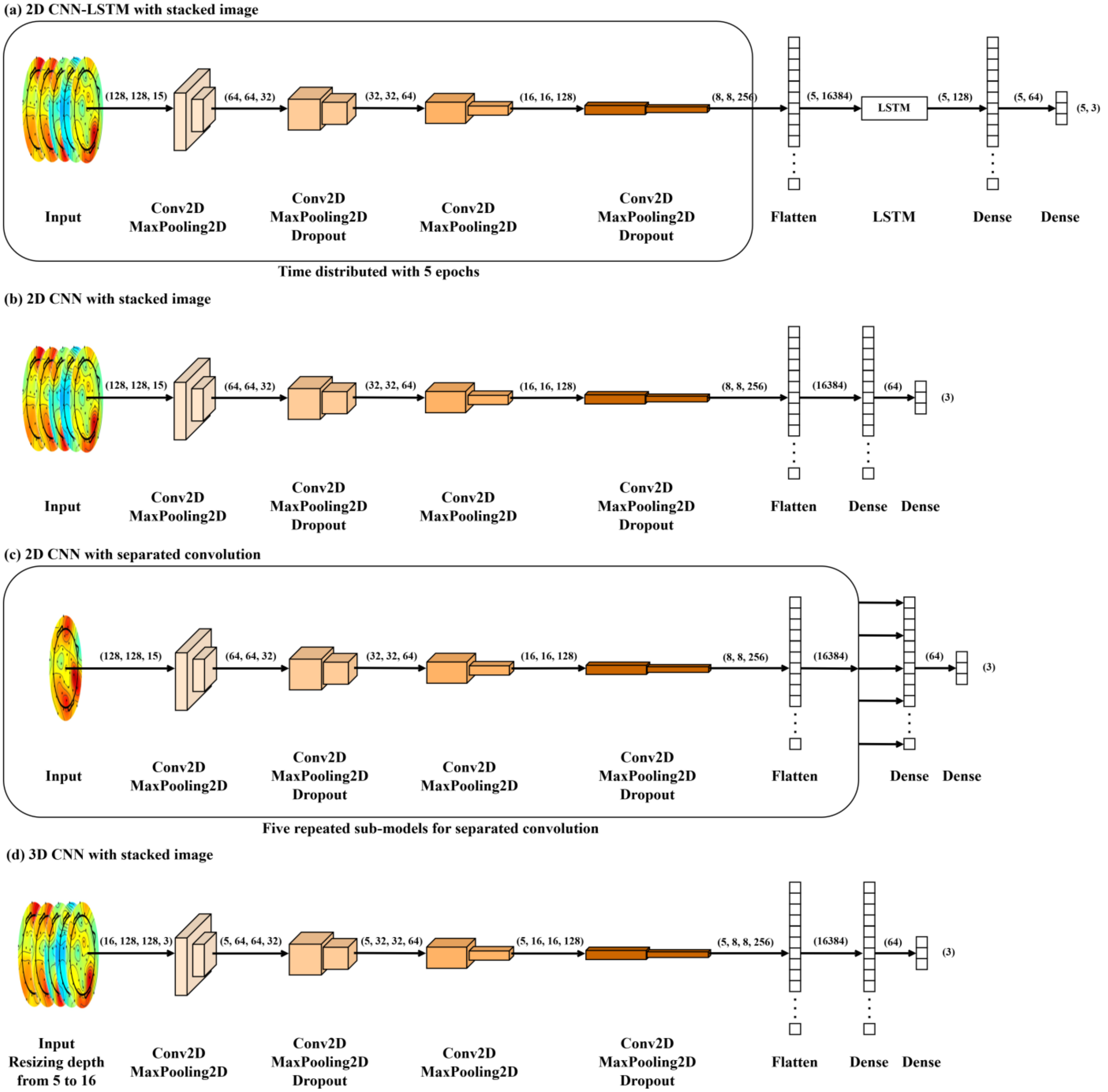
Parameter	Searching range
Objective	Maximise the validation accuracy
Early stopping condition	No loss reduction for ten consecutive epochs
Learning rate	0.001, 0.0001
Batch size	32, 64, 128
Number of layers/filters (2D CNN/CNN-LSTM)	(32, 64, 128), (64, 128, 256), (32, 64, 128, 256), (64, 128, 256, 512)
Number of layers/filters (3D CNN)	(64), (128), (64, 128), (128, 256)
Number of units (LSTM)	64, 128, 256

Note – The number of layers/filters are presented in (Filters in layer 1, filters in layer 2, …).

The main application areas of CNN are visual data processing and analysis, image classification, object detection, and image segmentation. As most BCIs are developed based on various neuroimaging techniques like EEG, its superiority in image classification tasks advances the development of BCI [67]. In our case, we transformed the time-series EEG data into spectral features that demonstrate different mental states like drowsiness (theta wave, 4-8 Hz), awake (beta wave, 13-30 Hz), concentration (gamma wave, 40-45 Hz). These spectral features could reveal the mental state for each channel, which were plotted spatially in topographical plots to represent the mental state different brain regions. Therefore, CNN can further capture the spatial features across the spectral data from different brain regions for enhanced classification.

C. Long Short-Term Memory Layer

Other than the spatial features from the topographical plots, there may exists temporal dependencies between successive EEG topographical plots. These temporal dependencies may reveal whether there are precursors to certain mental states. Hence, to further capture the potential temporal effects, the LSTM layer is adopted. It is a type of recurrent neural network composed of a cell, an input gate, an output gate, and a forget gate. The memory cells enable the network to retain information across extended sequences, thus having the advantages of avoid gradient vanishing problems and capture long-term temporal dependencies [68]. The main application area of LSTM is time series recognition. In our case, the CNN already extracts the spatial features from the EEG topographical plots. Hence, the LSTM layer aims to further capture the temporal effects and dependencies based on the learned features of the convolutional modules.



(a) 2D CNN-LSTM with stacked image

Time distributed with 5 epochs

(b) 2D CNN with stacked image

(c) 2D CNN with separated convolution

Five repeated sub-models for separated convolution

(d) 3D CNN with stacked image

Fig. 4. Network architecture for proposed first-stage models

D. Data input and model configuration

In this study, we propose four different deep learning architectures for the adaptive automation model with two data stacking methods. These architectures include (1) a 2D CNN-LSTM that also captures the temporal information between epochs, (2) a 2D CNN with stacked image, (3) a 2D CNN that perform separated convolutions on EEG spectral features from different EEG band waves, and (4) a 3D CNN that captures spatial information from the stacked topographical plots. The following illustrates their data inputs and the respective model configuration.

1) Data stacking and splitting: To prepare data for inputting into the deep learning architecture, we adopted two different approaches, namely image stacking and separated inputs.

For image stacking, topographical plots of the five band waves collected from the same epoch are stacked together. As CNN can be implemented with 2D or 3D convolution operations. The previous takes the input (width, height, channels) while the latter takes (depth, width, height, channels), we considered two different stacking methods to suit 2D and 3D convolution operations. For the 2D case, images are stacked on the channel axis, resulting in an image of (128, 128, 15) for each epoch. For the 3D case, images are stacked on the depth axis, giving each epoch an image of (5, 128, 128, 3). Note that the output depth d_{out} of a convolution with kernel size of 3 and max-pooling operation is calculated by $(d_{in} - 2)/2$. It limits the number of convolutional modules so that complex patterns may hardly be learned. Therefore, we resize the depth of each 3D image into (16, 128, 128, 3) by dividing the factor of $d_{current}/d_{desired} = 5/16$. With depth of 16, the 3D CNN can accommodate two convolutional modules between input and output layers.

Other than stacking images with the same epoch, we also adopted separated inputs and convolutional modules for each EEG band waves. This approach feed images from different band waves to a sub-model with individual input layers and performed convolutions separately. Then, the sub-models are concatenated together for an output. The details of this architecture will be discussed in **Section IV(D)(2)**.

Furthermore, the spatiotemporal 2D CNN-LSTM model requires image sequences as the input. Therefore, we transformed the stacked images of every five successive epochs as a sequence for the LSTM layer to capture the temporal relationship. The input size of the model is thus (time step, width, height, channels), i.e., (5, 128, 128, 15).

To ensure a balanced dataset is used for training, the stratified sampling method splits the entire dataset into train, validation, and test sets. Firstly, we divide the dataset of EEG topographical plots into subsets for each participant and each LOA. Secondly, we shuffle each subset with a buffer size equal to the size of the subset to reduce variance and maintain the generalisability of the model. Thirdly, we extract 10% of data from each subset and concatenate them as the test set. Fourthly, a five-fold cross-validation is adopted. Each remaining data subset is separated into five sets. One of the five sets is used as the validation set of each cross-validation iteration. The remainder are used as the training set.

2) Model configuration: In neural networks, an adequately defined network structure best fits the data for the best prediction performance. Hence, we leverage a grid search to optimise model performance for first-stage models using KerasTuner 2.5.0. The grid search includes the learning rate, batch size, number of layers, and filters. TABLE III shows the grid search matrix.

Each search trial is terminated if there is no loss reduction for three consecutive epochs. The grid search was conducted in the Windows 10 Enterprise 64-bit operating environment with AMD Ryzen 9 7950X 4.50 GHz CPU, 128 GB RAM, and NVIDIA GeForce RTX 4090 24 GB. The computational results indicate that the optimal network structure is (32, 64, 128, 256) for 2D CNN and (64, 128) for 3D CNN, with a learning rate of 0.001 and batch size of 64. The optimal number of LSTM units is 128. The optimal configuration is then adopted for model training. Fig. 4. shows the network architecture of the proposed first-stage models. Each of the model structure are described in detail below.

Model (1) – 2D CNN-LSTM with stacked image: It starts with an input layer of 5 * 128 * 128 * 15, where the foremost ‘5’ represents the number of time step in each image sequence. The EEG topographical plots are rescaled to [0, 1] by a rescaling layer of 1/255. Four convolution modules of kernel (2, 2) and filters (32, 64, 128, 256) are adopted with a 2D convolution layer, a batch normalisation layer, and a 2D max-pooling layer each. Dropout layers are added to the end of the second (dropout rate: 0.3) and the fourth (dropout rate: 0.4) convolution modules to prevent overfitting. A TimeDistributed wrapper is added to each module and a flatten layer is applied. Then, a LSTM layer with 128 units is appended to capture the temporal relationship. Finally, two fully connected layers with 64 and 3 units are used for classification.

TABLE IV

PERFORMANCE COMPARISON BETWEEN THE PROPOSED MODELS AND THE BENCHMARKING MODELS

Algorithm	Test accuracy	
	First-stage	Second-stage
Proposed image-based/spatiotemporal models		
2D CNN-LSTM (Stacked image)	0.9918	0.9907
2D CNN (Stacked image)	0.9885	0.9825
2D CNN (Separated convolutional modules)	0.9875	0.9815
3D CNN (Stacked image)	0.9245	0.9435
Benchmarking models		
LightGBM	0.6839	0.8834
Random forest	0.6556	0.8825
XGBoost	0.6489	0.8711
SVM (RBF)	0.5803	0.8138
Decision tree	0.5752	0.8155
Logistic regression	0.4478	0.6849
SVM (Linear)	0.4254	0.6886
SVM (Polynomial)	0.4162	0.6933

Note – Ranked by descending order of test accuracy of the first-stage model. SVM: Support vector machine, RBF: Radial Basis Function. Channel F4 and CP6 represent the frontal and parietal lobe respectively, which show significant difference ($p < 0.05$) between three LOA.

Model (2) – 2D CNN with stacked image: It starts with an input layer of 128 * 128 * 15. Likewise, the EEG topographical plots are rescaled to [0, 1] with the same convolution module structure as Model (1) Then, a flattening layer and two fully connected layers with 64 and 3 units are appended to make predictions.

Model (3) – 2D CNN with separated convolution modules: It starts five separated sub-models. Each sub-model begins with an input layer of 128 * 128 * 3, which is the dimension of an EEG topographical plot image of a band wave. Then, the data are rescaled by 1/255 and performed convolutions in the same manner as the stacked image approach. After the flattening layer of each sub-model, the five sub-models are concatenated. Finally, two fully connected layers with 64 and 3 units are used for classification.

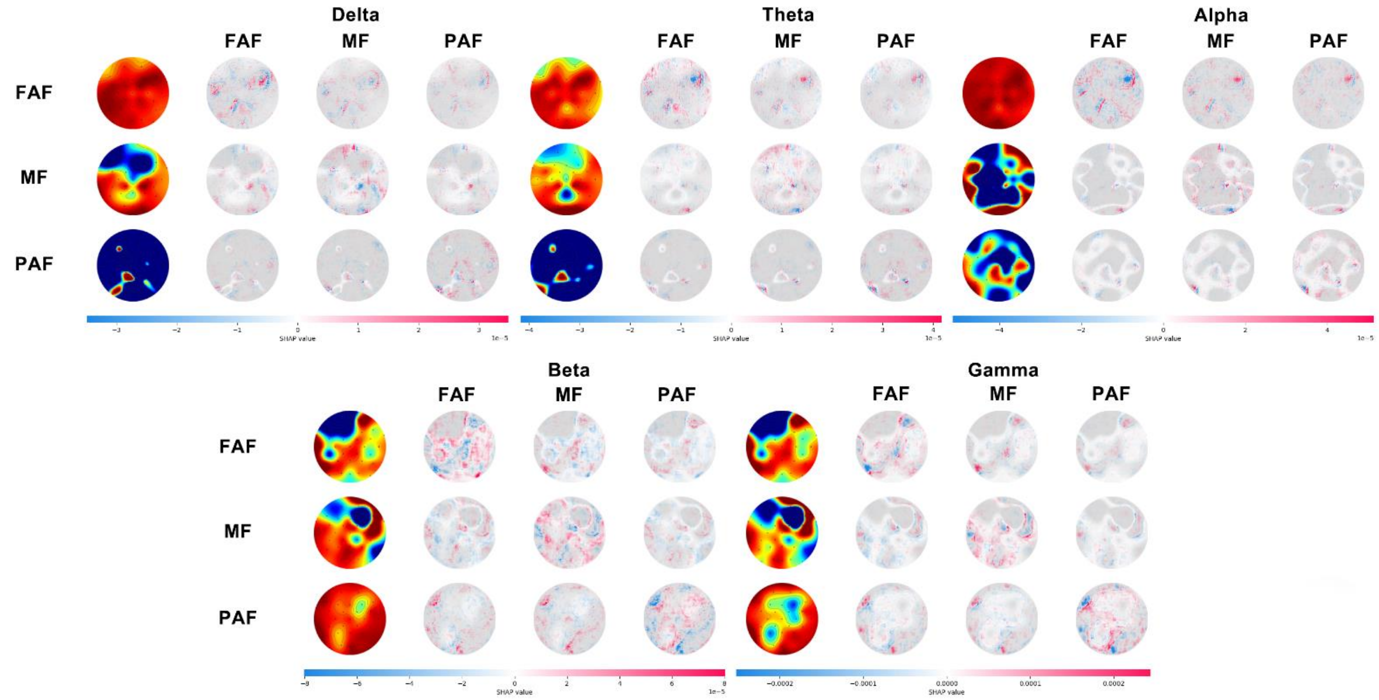


Fig. 5. Model explanations on a sample from each LOA in first-stage classification (Left: True class and actual image; Top: Label for each class).

Model (4) – 3D CNN with stacked image: It begins of an input layer of 16 * 128 * 128 * 3. Image inputs are rescaled to [0, 1] by 1/255. Two convolution modules of kernel (2, 2, 2) and filters (64, 128) are adopted with a 3D convolution layer, a batch normalisation layer, and a 3D max-pooling layer each. A 3D global average pooling layer is appended to the end of the second convolution module. Then, a dropout layer with a dropout rate of 0.3 follows. Finally, two fully connected layers with 64 and 3 units are used for classification.

As the dataset of the second stage is the subset of that of the first stage, the second-stage classification adopts the same set of network architectures in the first stage, except the last layer being changed to a single neuron for binary classification, i.e., OOTL versus non-OOTL.

All models are coded using Python 3.8.18 with TensorFlow Keras 2.5.0. The maximum epoch for training is set as 100. The training is terminated if the validation accuracy does not improve for ten epochs. Adam optimiser is used with a learning rate of 0.001 and a batch size of 64. It is executed using the same platform as the grid search.

3) Shapley additive explanations for model interpretation: As mentioned in **Section II(C)**, the uncertainty of ‘why’ and ‘how’ the ‘black-box’ deep learning models perform classification may hinder one’s confidence in applying them in safety-critical domains. In this regard, we adopt the DeepExplainer of SHAP to explain the deep 2D CNN image classifiers. SHAP is a game theoretical approach utilising Shapley values to evaluate the impacts of each feature on the model output to explain the prediction [*69*]. In image classification tasks, SHAP explains how different portions of an image affect the model output by visualising the positive and negative Shapley values for each classification class. In our case, with EEG topographical plot inputs, one can then realise which brain regions and frequency band waves help the model to classify the mental state as human OOTL. Apart from increasing trust in deep learning models, implementing SHAP also provide theoretical insights on how brain activity reflects human OOTL can be gained.

We randomly selected 100 sets of plots as background examples of the DeepExplainer. Then, a sample is chosen randomly from each LOA (in the first stage model) and from non-OOTL and OOTL classes (in the second stage model) to explain the prediction output. It is implemented in Python 3.8.18 using SHAP 0.41.0.

E. Baseline comparison with non-image classification models

To evaluate the performance of the proposed model and its configuration, we employ other machine learning algorithms using non-image-based inputs, including random forest, decision tree, support vector machine, logistic regression, XGBoost, and LightGBM as baseline models. For non-image inputs, we have tabular data in 3 (LOA) * 32 (Channels) * 5 (EEG frequency bands). However, it is known that a large number of features may adversely affect the model performance due to reduced feature impact. Literature suggests that the parietal lobe integrates information such as visual information that needs to be obtained outside the cockpit and internal auditory warnings [70]. Meanwhile, the frontal lobe reflects cognition, memory, decision-making, and problem-solving processes [71]. Therefore, channel CP6 from the parietal lobe and F4 from the frontal lobe are selected for analysis [72, 73]. These benchmarking algorithms are coded using Python 3.8.18 with SciKit-learn 1.0.2, XGBoost 1.6.2, and LightGBM 4.2.0.

V. Results and Discussions

In this section, we present the results of our proposed deep CNN models, other similar models for mind wandering detection, and the non-image-based benchmarking models and discuss their performance. In line with enabling trustworthy

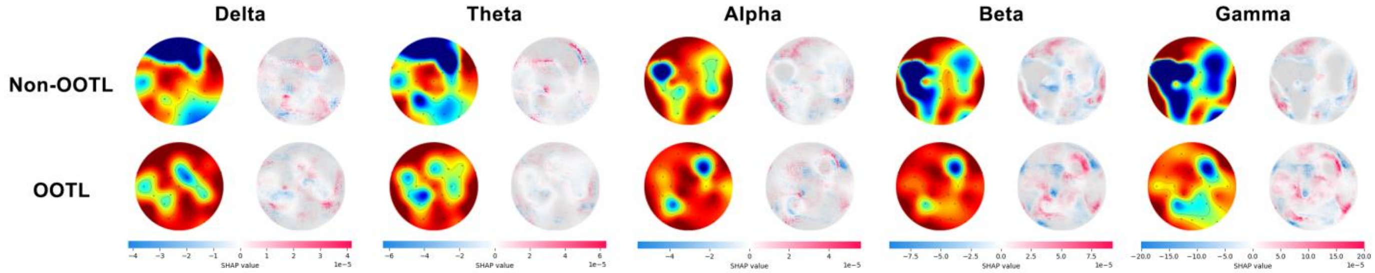


Fig. 6. Model explanations on a sample from OOTL and non-OOTL class in second-stage classification (Left: True class and actual image).

deep learning models for aviation applications, we also interpret the explanations done by SHAP on the proposed 2D CNN and discuss the insights drawn from the explanation.

A. Classification performance

TABLE IV shows the model performance of the proposed image-based deep learning models and non-image-based benchmarking models. The image-based models outperform all non-image-based benchmarking models in both the first and second stage classification. The 2D CNN-LSTM with stacked image input attains the highest test accuracy at 0.9918 and 0.9907 for both the first and second stage, respectively. It is followed by the models without the LSTM module, but the method of image stacking and convolution does not significantly influence the test accuracies. Conversely, the benchmarking models attain an average accuracy of 0.5542 ± 0.1096 and 0.7916 ± 0.0892 for the first and second stage, respectively. The LightGBM attains the highest accuracy among the benchmarking models at 0.6839 in the first stage and 0.8834 in the second stage. The SVM with polynomial function attains the lowest accuracy at 0.4162 in the first stage, while the logistic regression yields the lowest accuracy at 0.6849 in the second stage. Other than the benchmarking models, we also compare with several models that detects mind wandering. The SVM models of [74] concluded that the performance cannot achieve accuracy higher than chance levels, which aligns with our baseline SVM models, while other similar models also got only an average accuracy of 65-70% [75, 76]. Compared to the deep learning model for augmented EEG data in time-frequency domain of [77], our model outperformed their model with a slightly higher accuracy. To sum up, our proposed models have the highest accuracy among these studies.

The above results indicate that the image-based approach boosts the classification accuracy by at least 30% in the first stage and at least 10% in the second stage. It suggests that the CNN models, which capture the spatial information of each EEG channel using image-based PSD data, can significantly enhance the ability of the classifier to distinguish the mental state differences between FAF, PAF, and MF. The LSTM layer further learned the temporal patterns in the model to achieve better accuracy. In addition, using the absolute PSDs as model inputs may not be as effective as using relative values presented in images. This phenomenon is because PSDs cannot be directly interpreted for mental state assessment but require a baseline condition for comparison. Hence, as all data is scaled to RGB (0-255) in the image-based approach, this approach may thus attain an improved mental state classification performance.

B. Model interpretation with Shapley Additive Explanations

Fig. 5 and Fig. 6. shows the classification explanations on the randomly selected sample from each LOA and OOTL/non-OOTL class, respectively. It shows the actual PSD plots and the shaded plots. Red and blue shades indicate those parts contribute positively and negatively to the prediction of that classification class, respectively.

In Fig. 5, the delta and theta waves of FAF in the frontal and parietal lobes are higher than that of PAF and MF, indicating a state of sleep and drowsiness. The corresponding areas are shaded in red, suggesting that these potential sleep and drowsiness contribute positively to FAF predictions. It aligns with existing knowledge that high delta and theta waves are usually found in a drowsy state. Therefore, FAF can be a potential indicator of OOTL but not necessarily indicating OOTL. The state of OOTL can be further verified in the second stage model.

Conversely, the beta and gamma waves of MF and PAF in the frontal and parietal lobes are higher than that of FAF, which suggests pilots are more awake and concentrated in these two states. These areas are contributing positively to the predictions of MF and PAF. PAF shows a higher PSD in the parietal lobe, which suggests that pilots in PAF integrate various sensory information to maintain their SA. Meanwhile, MF shows a higher PSD in the right frontal lobe, which could be attributed to monitoring behaviours and sustained attention.

Furthermore, the alpha wave may facilitate the model to distinguish between MF and PAF. We realised that MF has a lower alpha power for the parietal lobe than PAF. It implies that pilots are less relaxed in MF than in PAF, which is reasonable as pilots in MF need to manoeuvre the aircraft using the sidestick as well. Note that both are shaded in red, which suggests that the parietal lobe contributes to the prediction of MF and PAF, respectively. It further signifies that PAF pilots are more relaxed than MF, which may improve their decision-making performance.

In Fig. 6., OOTL are characterised by heightened delta and theta waves in parietal lobe being red shaded by SHAP. Furthermore, the elevated beta and gamma waves of the non-OOTL sample in the frontal lobe is attributed to red shades, which aligns with the theoretical background that pilots who are not OOTL shall be more awake and concentrated. Conversely, the OOTL sample has blue shades in elevated beta and gamma regions in the frontal lobe, while having red shades in the regions with lower beta and gamma band powers. It also aligns with the fact that lower beta and gamma band powers shall favour the classification of OOTL, since pilots shall be less awake and concentrated when they are

OOTL. Even there might be abnormality of elevated beta and gamma regions in the OOTL sample, our results indicated that the classification model remains valid through shading those features with blue shade, thus classifying it as OOTL. These interpretations can enhance the trustworthiness for the model to be safely applied in aviation as the prediction output can be linked to the theoretical background for interpretation.

VI. Conclusion

This paper proposes an adaptive automation framework to keep the pilots in the loop and mitigate under/overload during cruising flights. An explainable spatiotemporal deep learning framework with 2D and 3D convolutional layers, as well as an LSTM layer are designed and trained using EEG PSD collected from a simulator-based experiment to identify human OOTL. Our computational results indicate that the proposed image-based approach attains a high accuracy at 0.9918 in the first stage and 0.9907 in the second stage, which outperforms the non-image-based benchmarking models by a 30.79% and 10.73% enhancement in accuracy in the first and second stages, respectively. Furthermore, we also employ SHAP to explain the classification outputs to strengthen the model's trustworthiness for safety-critical aviation applications. Cockpit designers can utilise the domain knowledge and managerial insights drawn from SHAP in cockpit automation design so that human operators can interact with automation more seamlessly.

This study is also subject to several limitations. Our EEG data is collected from a flight simulator with 24 cadet pilots who have experience with flight simulators. While flight simulators are widely used in flight training, the fixed-base flight simulators may only partially replicate the real-world flight experience without a motion platform. Cadet pilots also have fewer flight/simulator hours when compared to licensed pilots. In addition, certain assumptions, despite being verified empirically, are used in data labelling. Hence, future studies may replicate the study with increased sample size of licensed pilots preferably in naturalistic settings. The model can be further integrated with flight automation and used in conjunction with commercial grade dry EEGs in real-world application setting for model validation on a flight with sufficient cruising time, as the dry EEGs can be set up quickly for implementation despite a slight drop in data quality. These enhancements in data collection, labelling, and future applications shall improve the model and mitigate unwanted consequences of human OOTL during cruising flights.

Acknowledgment

Our gratitude is extended to all experiment participants for their contribution to data curation of this study.

References

[1] J. Gouraud, A. Delorme, and B. Berberian, "Autopilot, Mind Wandering, and the Out of the Loop Performance Problem," *Frontiers in Neuroscience,* vol. 11, 2017, doi: 10.3389/fnins.2017.00541.

[2] A. Sebok and C. D. Wickens, "Implementing Lumberjacks and Black Swans Into Model-Based Tools to Support Human–Automation Interaction," *Human Factors,* vol. 59, no. 2, pp. 189-203, 2017, doi: 10.1177/0018720816665201.

[3] C. Y. Yiu, K. K. H. Ng, S. C. M. Yu, and C. W. Yu, "Sustaining aviation workforce after the pandemic: Evidence from Hong Kong aviation students toward skills, specialised training, and career prospects through a mixed-method approach," *Transport Policy,* vol. 128, pp. 179-192, 2022, doi: 10.1016/j.tranpol.2022.09.020.

[4] M. Tatasciore and S. Loft, "Can increased automation transparency mitigate the effects of time pressure on automation use?," *Applied Ergonomics,* vol. 114, p. 104142, 2024, doi: 10.1016/j.apergo.2023.104142.

[5] I. Gegoff, M. Tatasciore, V. Bowden, J. McCarley, and S. Loft, "Transparent Automated Advice to Mitigate the Impact of Variation in Automation Reliability," *Human Factors,* 2023, doi: 10.1177/00187208231196738.

[6] B. Gebru *et al.*, "A Review on Human–Machine Trust Evaluation: Human-Centric and Machine-Centric Perspectives," *IEEE Transactions on Human-Machine Systems,* vol. 52, no. 5, pp. 952-962, 2022, doi: 10.1109/THMS.2022.3144956.

[7] C. Billings, J. Lauber, H. Funkhouser, E. Lyman, and E. Huff, "NASA aviation safety reporting system," 1976.

[8] T. B. Sheridan and R. Parasuraman, "Human-Automation Interaction," *Reviews of Human Factors and Ergonomics,* vol. 1, no. 1, pp. 89-129, 2005, doi: 10.1518/155723405783703082.

[9] T. O'Neill, N. McNeese, A. Barron, and B. Schelble, "Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature," *Human Factors,* vol. 64, no. 5, pp. 904-938, 2020, doi: 10.1177/0018720820960865.

[10] B. Berberian, B. Somon, A. Sahaï, and J. Gouraud, "The out-of-the-loop Brain: A neuroergonomic approach of the human automation interaction," *Annual Reviews in Control,* vol. 44, pp. 303-315, 2017, doi: 10.1016/j.arcontrol.2017.09.010.

[11] L. Onnasch, C. D. Wickens, H. Li, and D. Manzey, "Human Performance Consequences of Stages and Levels of Automation:An Integrated Meta-Analysis," *Human Factors,* vol. 56, no. 3, pp. 476-488, 2014, doi: 10.1177/0018720813501549.

[12] Q. Li, L. H. Chi, H. M. Him, L. K. Lok, N. K. K. H., and C. Y. and Yiu, "The Effects of Aeronautical Decision-Making Models on Student Pilots' Situational Awareness and Cognitive Workload in Simulated Non-Normal Flight Deck Environment," *The International Journal of Aerospace Psychology,* vol. 33, no. 3, pp. 197-213, 2023/07/03 2023, doi: 10.1080/24721840.2023.2231506.

[13] L. Bainbridge, "Ironies of automation," *Automatica,* vol. 19, no. 6, pp. 775-779, 1983, doi: 10.1016/0005-1098(83)90046-8.

[14] J. S. Warm, R. Parasuraman, and G. Matthews, "Vigilance Requires Hard Mental Work and Is Stressful," *Human Factors,* vol. 50, no. 3, pp. 433-441, 2008, doi: 10.1518/001872008X312152.

[15] D. R. Thomson, P. Seli, D. Besner, and D. Smilek, "On the link between mind wandering and task performance over time," *Consciousness and Cognition,* vol. 27, pp. 14-26, 2014, doi: 10.1016/j.concog.2014.04.001.

[16] Y. Q. Wingelaar-Jagt, T. T. Wingelaar, W. J. Riedel, and J. G. Ramaekers, "Fatigue in Aviation: Safety Risks, Preventive Strategies and Pharmacological Interventions," *Frontiers in Physiology,* Review vol. 12, 2021, doi: 10.3389/fphys.2021.712628.

[17] S. Ayas, B. Donmez, and X. Tang, "Drowsiness Mitigation Through Driver State Monitoring Systems: A Scoping Review," *Human Factors,* 2023, doi: 10.1177/00187208231208523.

[18] E. Balta, A. Psarrakis, and A. Vatakis, "The effects of increased mental workload of air traffic controllers on time perception: Behavioral and physiological evidence," *Applied Ergonomics,* vol. 115, p. 104162, 2024, doi: 10.1016/j.apergo.2023.104162.

[19] S. Rothfuß, M. Wörner, J. Inga, A. Kiesel, and S. Hohmann, "Human–Machine Cooperative Decision Making Outperforms Individualism and Autonomy," *IEEE Transactions on Human-Machine Systems,* vol. 53, no. 4, pp. 761-770, 2023, doi: 10.1109/THMS.2023.3274916.

[20] C. Y. Yiu, K. K. H. Ng, F. T. S. Chan, and Q. Li, "Vaccines, associated risk and air transport industry post-COVID-19: A structural equation modelling-based empirical study in Hong Kong," *Research in Transportation Business & Management,* vol. 50, p. 101038, 2023, doi: 10.1016/j.rtbm.2023.101038.

[21] Q. Li, K. K. H. Ng, C. Y. Yiu, X. Yuan, C. K. So, and C. C. Ho, "Securing air transportation safety through identifying pilot's risky VFR flying behaviours: An EEG-based neurophysiological modelling using machine learning algorithms," *Reliability Engineering & System Safety,* vol. 238, p. 109449, 2023, doi: 10.1016/j.ress.2023.109449.

[22] E. van Weelden, M. Alimardani, T. J. Wiltshire, and M. M. Louwerse, "Aviation and neurophysiology: A systematic review," *Applied Ergonomics,* vol. 105, p. 103838, 2022, doi: 10.1016/j.apergo.2022.103838.

[23] C. Diaz-Piedra, H. Rieiro, A. Cherino, L. J. Fuentes, A. Catena, and L. L. Di Stasi, "The effects of flight complexity on gaze entropy: An experimental study with fighter pilots," *Applied Ergonomics,* vol. 77, pp. 92-99, 2019, doi: 10.1016/j.apergo.2019.01.012.

[24] C. Y. Yiu, N. K. K. H., L. Qinbiao, and X. and Yuan, "Gaze behaviours, situation awareness and cognitive workload of air traffic controllers in radar screen monitoring tasks with varying task complexity," *International Journal of Occupational Safety and Ergonomics,* pp. 1-12, 2025, doi: 10.1080/10803548.2025.2453312.

[25] R. Parasuraman and C. D. Wickens, "Humans: Still Vital After All These Years of Automation," *Human Factors,* vol. 50, no. 3, pp. 511-520, 2008, doi: 10.1518/001872008x312198.

[26] P. Liu, "Reflections on Automation Complacency," *International Journal of Human–Computer Interaction,* pp. 1-17, 2023, doi: 10.1080/10447318.2023.2265240.

[27] S.-Y. Han, N.-S. Kwak, T. Oh, and S.-W. Lee, "Classification of pilots' mental states using a multimodal deep learning network," *Biocybernetics and Biomedical Engineering,* vol. 40, no. 1, pp. 324-336, 2020, doi: 10.1016/j.bbe.2019.12.002.

[28] M. R. Endsley, "Ironies of artificial intelligence," *Ergonomics,* pp. 1-13, 2023, doi: 10.1080/00140139.2023.2243404.

[29] S. Zhai, S. Gao, L. Wang, and P. Liu, "When both human and machine drivers make mistakes: Whom to blame?," *Transportation Research Part A: Policy and Practice,* vol. 170, p. 103637, 2023, doi: 10.1016/j.tra.2023.103637.

[30] M. R. Endsley, "Understanding Automation Failure," *Journal of Cognitive Engineering and Decision Making,* 2024, doi: 10.1177/15553434231222059.

[31] M. R. Endsley and E. O. Kiris, "The Out-of-the-Loop Performance Problem and Level of Control in Automation," *Human Factors,* vol. 37, no. 2, pp. 381-394, 1995, doi: 10.1518/001872095779064555.

[32] M. Jipp and P. L. Ackerman, "The Impact of Higher Levels of Automation on Performance and Situation Awareness: A Function of Information-Processing Ability and Working-Memory Capacity," *Journal of Cognitive Engineering and Decision Making,* vol. 10, no. 2, pp. 138-166, 2016, doi: 10.1177/1555343416637517.

[33] D. Manzey, J. Reichenbach, and L. Onnasch, "Human Performance Consequences of Automated Decision Aids: The Impact of Degree of Automation and System Experience," *Journal of Cognitive Engineering and Decision Making,* vol. 6, no. 1, pp. 57-87, 2012, doi: 10.1177/1555343411433844.

[34] N. Strand, J. Nilsson, I. C. M. Karlsson, and L. Nilsson, "Semi-automated versus highly automated driving in critical situations caused by automation failures," *Transportation Research Part F: Traffic Psychology and Behaviour,* vol. 27, pp. 218-228, 2014, doi: 10.1016/j.trf.2014.04.005.

[35] G. Calhoun, "Adaptable (Not Adaptive) Automation: Forefront of Human–Automation Teaming," *Human Factors,* vol. 64, no. 2, pp. 269-277, 2021, doi: 10.1177/00187208211037457.

[36] J. Sauer, C.-S. Kao, and D. Wastell, "A comparison of adaptive and adaptable automation under different levels of environmental stress," *Ergonomics,* vol. 55, no. 8, pp. 840-853, 2012, doi: 10.1080/00140139.2012.676673.

[37] L. J. Prinzel, F. G. Freeman, M. W. Scerbo, P. J. Mikulka, and A. T. Pope, "Effects of a Psychophysiological System for Adaptive Automation on Performance, Workload, and the Event-Related Potential P300 Component," *Human Factors,* vol. 45, no. 4, pp. 601-614, 2003, doi: 10.1518/hfes.45.4.601.27092.

[38] S. D. Reddy, S. Goyal, and T. K. Reddy, "Riemannian Approach Based Depression classification using Transfer Learning for MEG signals," in *2023 IEEE 4th Annual Flagship India Council International Subsections Conference (INDISCON),* 5-7 Aug. 2023 2023, pp. 1-4, doi: 10.1109/INDISCON58499.2023.10270192.

[39] M. Deng *et al.*, "An analysis of physiological responses as indicators of driver takeover readiness in conditionally automated driving," *Accident Analysis & Prevention,* vol. 195, p. 107372, 2024, doi: 10.1016/j.aap.2023.107372.

[40] F. Wu, W. Mai, Y. Tang, Q. Liu, J. Chen, and Z. Guo, "Learning Spatial-Spectral-Temporal EEG Representations with Deep Attentive-Recurrent-Convolutional Neural Networks for Pain Intensity Assessment," *Neuroscience,* vol. 481, pp. 144-155, 2022, doi: 10.1016/j.neuroscience.2021.11.034.

[41] D. Das Chakladar, S. Dey, P. P. Roy, and D. P. Dogra, "EEG-based mental workload estimation using deep BLSTM-LSTM network and evolutionary algorithm," *Biomedical Signal Processing and Control,* vol. 60, p. 101989, 2020, doi: 10.1016/j.bspc.2020.101989.

[42] K. Muhammad, A. Ullah, J. Lloret, J. D. Ser, and V. H. C. d. Albuquerque, "Deep Learning for Safe Autonomous Driving: Current Challenges and Future Directions," *IEEE Transactions on Intelligent Transportation Systems,* vol. 22, no. 7, pp. 4316-4336, 2021, doi: 10.1109/TITS.2020.3032227.

[43] Y. Gu, Y. Jiang, T. Wang, P. Qian, and X. Gu, "EEG-Based Driver Mental Fatigue Recognition in COVID-19 Scenario Using a Semi-Supervised Multi-View Embedding Learning Model," *IEEE Transactions on Intelligent Transportation Systems,* vol. 25, no. 1, pp. 859-868, 2024, doi: 10.1109/TITS.2022.3211536.

[44] V. Gupta, T. P. Kendre, T. K. Reddy, and V. Arora, "Comparative Performance Analysis of Scalp EEG and Ear EEG based P300 Ambulatory Brain-Computer Interfaces using Riemannian Geometry and Convolutional Neural Networks," in *2022 National Conference on Communications (NCC)*, 24-27 May 2022 2022, pp. 314-319, doi: 10.1109/NCC55593.2022.9806815.

[45] S. Singh, V. Gupta, T. K. Reddy, B. Bhushan, and L. Behera, "Meditation and Cognitive Enhancement: A Machine Learning Based Classification Using EEG," in *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 9-12 Oct. 2022 2022, pp. 1973-1978, doi: 10.1109/SMC53654.2022.9945131.

[46] E. Q. Wu *et al.*, "Detecting Fatigue Status of Pilots Based on Deep Learning Network Using EEG Signals," *IEEE Transactions on Cognitive and Developmental Systems,* vol. 13, no. 3, pp. 575-585, 2021, doi: 10.1109/TCDS.2019.2963476.

[47] E. Q. Wu *et al.*, "Inferring Cognitive State of Pilot's Brain Under Different Maneuvers During Flight," *IEEE Transactions on Intelligent Transportation Systems,* vol. 23, no. 11, pp. 21729-21739, 2022, doi: 10.1109/TITS.2022.3189981.

[48] C. Lv, J. Nian, Y. Xu, and B. Song, "Compact Vehicle Driver Fatigue Recognition Technology Based on EEG Signal," *IEEE Transactions on Intelligent Transportation Systems,* vol. 23, no. 10, pp. 19753-19759, 2022, doi: 10.1109/TITS.2021.3119354.

[49] A. Chougule, J. Shah, V. Chamola, and S. Kanhere, "Enabling Safe ITS: EEG-Based Microsleep Detection in VANETs," *IEEE Transactions on Intelligent Transportation Systems,* vol. 24, no. 12, pp. 15773-15783, 2023, doi: 10.1109/TITS.2022.3230259.

[50] A. Balaji, U. Tripathi, V. Chamola, A. Benslimane, and M. Guizani, "Toward Safer Vehicular Transit: Implementing Deep Learning on Single Channel EEG Systems for Microsleep Detection," *IEEE Transactions on Intelligent Transportation Systems,* vol. 24, no. 1, pp. 1052-1061, 2023, doi: 10.1109/TITS.2021.3125126.

[51] P. Angkan *et al.*, "Multimodal Brain–Computer Interface for In-Vehicle Driver Cognitive Load Measurement: Dataset and Baselines," *IEEE Transactions on Intelligent Transportation Systems,* pp. 1-16, 2024, doi: 10.1109/TITS.2023.3345846.

[52] P. Mi *et al.*, "Driver Cognitive Architecture Based on EEG Signals: A Review," *IEEE Sensors Journal*, pp. 1-1, 2024, doi: 10.1109/JSEN.2024.3471699.

[53] D. H. Lee, J. H. Jeong, B. W. Yu, T. E. Kam, and S. W. Lee, "Autonomous System for EEG-Based Multiple Abnormal Mental States Classification Using Hybrid Deep Neural Networks Under Flight Environment," *IEEE Transactions on Systems, Man, and Cybernetics: Systems,* vol. 53, no. 10, pp. 6426-6437, 2023, doi: 10.1109/TSMC.2023.3282635.

[54] X. Yu, C.-H. Chen, and H. Yang, "Air traffic controllers' mental fatigue recognition: A multi-sensor information fusion-based deep learning approach," *Advanced Engineering Informatics,* vol. 57, p. 102123, 2023, doi: 10.1016/j.aei.2023.102123.

[55] J. Dong, S. Chen, M. Miralinaghi, T. Chen, P. Li, and S. Labi, "Why did the AI make that decision? Towards an explainable artificial intelligence (XAI) for autonomous driving systems," *Transportation Research Part C: Emerging Technologies,* vol. 156, p. 104358, 2023, doi: 10.1016/j.trc.2023.104358.

[56] M. R. Endsley, "Supporting Human-AI Teams:Transparency, explainability, and situation awareness," *Computers in Human Behavior,* vol. 140, p. 107574, 2023, doi: 10.1016/j.chb.2022.107574.

[57] E. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 32, no. 11, pp. 4793-4813, 2021, doi: 10.1109/TNNLS.2020.3027314.

[58] R. Dwivedi *et al.*, "Explainable AI (XAI): Core Ideas, Techniques, and Solutions," *ACM Comput. Surv.,* vol. 55, no. 9, p. Article 194, 2023, doi: 10.1145/3561048.

[59] C. Y. Yiu *et al.*, "A Digital Twin-Based Platform towards Intelligent Automation with Virtual Counterparts of Flight and Air Traffic Control Operations," *Applied Sciences,* vol. 11, no. 22, p. 10923, 2021, doi: 10.3390/app112210923.

[60] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods,* vol. 134, pp. 9-21, 2004, doi: 10.1016/j.jneumeth.2003.10.009.

[61] A. Mognon, J. Jovicich, L. Bruzzone, and M. Buiatti, "ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features," *Psychophysiology,* vol. 48, no. 2, pp. 229-240, 2011, doi: 10.1111/j.1469-8986.2010.01061.x.

[62] S. Sanei and J. A. Chambers, *EEG signal processing*. John Wiley & Sons, 2013.

[63] H. Taheri Gorji, N. Wilson, J. VanBree, B. Hoffmann, T. Petros, and K. Tavakolian, "Using machine learning methods and EEG to discriminate aircraft pilot cognitive workload during flight," *Scientific Reports,* vol. 13, no. 1, p. 2507, 2023, doi: 10.1038/s41598-023-29647-0.

[64] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 33, no. 12, pp. 6999-7019, 2022, doi: 10.1109/TNNLS.2021.3084827.

[65] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics,* vol. 36, no. 4, pp. 193-202, 1980/04/01 1980, doi: 10.1007/BF00344251.

[66] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into Imaging,* vol. 9, no. 4, pp. 611-629, 2018/08/01 2018, doi: 10.1007/s13244-018-0639-9.

[67] M. T. Sadiq, M. Z. Aziz, A. Almogren, A. Yousaf, S. Siuly, and A. U. Rehman, "Exploiting pretrained CNN models for the development of an EEG-based robust BCI framework," *Computers in Biology and Medicine,* vol. 143, p. 105242, 2022, doi: 10.1016/j.compbiomed.2022.105242.

[68] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation,* vol. 9, no. 8, pp. 1735-1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[69] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," presented at the Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 2017.

[70] C. Y. Yiu *et al.*, "Towards safe and collaborative aerodrome operations: Assessing shared situational awareness for adverse weather detection with EEG-enabled Bayesian neural networks," *Advanced Engineering Informatics,* vol. 53, p. 101698, 2022, doi: 10.1016/j.aei.2022.101698.

[71] Q. Li, K. K. H. Ng, S. C. M. Yu, C. Y. Yiu, and M. Lyu, "Recognising situation awareness associated with different workloads using EEG and eye-tracking features in air traffic control tasks," *Knowledge-Based Systems,* vol. 260, p. 110179, 2023, doi: 10.1016/j.knosys.2022.110179.

[72] M. A. Schier, "Changes in EEG alpha power during simulated driving: a demonstration," *International Journal of Psychophysiology,* vol. 37, no. 2, pp. 155-162, 2000, doi: 10.1016/S0167-8760(00)00079-9.

[73] S. Barua, M. U. Ahmed, C. Ahlström, and S. Begum, "Automatic driver sleepiness detection using EEG, EOG and contextual information," *Expert Systems with Applications,* vol. 115, pp. 121-135, 2019, doi: 10.1016/j.eswa.2018.07.054.

[74] C. Y. Jin, J. P. Borst, and M. K. van Vugt, "Distinguishing vigilance decrement and low task demands from mind-wandering: A machine learning analysis of EEG," *European Journal of Neuroscience,* vol. 52, no. 9, pp. 4147-4164, 2020, doi: 10.1111/ejn.14863.

[75] J. M. Groot *et al.*, "Probing the neural signature of mind wandering with simultaneous fMRI-EEG and pupillometry," *NeuroImage,* vol. 224, p. 117412, 2021, doi: 10.1016/j.neuroimage.2020.117412.

[76] S. Chaudhary, P. Pandey, K. P. Miyapuram, and D. Lomas, "Classifying EEG Signals of Mind-Wandering Across Different Styles of Meditation," in *Brain Informatics*, Cham, M. Mahmud, J. He, S. Vassanelli, A. van Zundert, and N. Zhong, Eds., 2022: Springer International Publishing, pp. 152-163.

[77] S. Hosseini and X. Guo, "Deep Convolutional Neural Network for Automated Detection of Mind Wandering using EEG Signals," presented at the Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, Niagara Falls, NY, USA, 2019.

Cho Yin Yiu is currently a Ph.D. candidate in the Human Factors and Ergonomics Laboratory, Department of Aeronautical and Aviation Engineering (AAE), The Hong Kong Polytechnic University (PolyU). He received his B.Eng. (Hons) in Aviation Engineering and Minor in Applied Psychology from PolyU in 2022.

Kam K.H. Ng (Member, IEEE) is currently the Associate Head and an Associate Professor in the Department of Aeronautical and Aviation Engineering (AAE), The Hong Kong Polytechnic University (PolyU). He received his Ph.D. Degree from the Department of Industrial and Systems Engineering, PolyU in 2019.

Qinbiao Li is currently a Research Assistant Professor in the Human Factors and Ergonomics Laboratory, Department of Aeronautical and Aviation Engineering (AAE), The Hong Kong Polytechnic University (PolyU). He received his Ph.D. Degree from AAE, PolyU in 2024.

Xin Yuan is currently a Ph.D. candidate in the Human Factors and Ergonomics Laboratory, Department of Aeronautical and Aviation Engineering (AAE), The Hong Kong Polytechnic University. She received her M.Eng. degree from Shandong University, China in 2022.