International Journal of Technology, Management and Humanities Volume 12, Issue 2, 2026

Multimodal AI for Pilot Skill Assessment Using Physiological Signals

Sam Suseelan

Independent Researcher

Abstract

Abstract: Yet traditional methods of evaluating pilot skill, focused mainly on subjective instructor ratings or post hoc evaluation of aviation performance, only measure the observable products of cognitive functioning of skilled flying, and flight-related cognitive functioning is vitally important to modern aviation. This paper aims at developing four new frameworks: the MultiModal Aviator Performance and State Assessment AI Framework (MAPS-AI), that combines four physiological signals modalities (EEG, fNIRS, ECG and eye-tracking) and deep learning architectures for the purpose of real-time and objective monitoring of cognitive state in-flight while performing cognitive tasks during both flight training and flight operations. This framework brings together multi-sensor fusion and AI classification pipelines, limits of which are explored in the psychophysiological literature on pilot mental workload, the cognitive fatigue literature and EEG monitoring in actual flight conditions, fNIRS-based engagement detection in landing scenarios, multimodal workload assessment, and recent developments in compact EEG deep learning. A six-state cognitive state taxonomy is proposed and evidence is compiled demonstrating that incorporating progressive multimodal fusion yields a classification accuracy of around 83% compared a multimodal (EEG+EEG+EEG+EEG) approach with a accuracy of around 65% (EEG-only). Discussions of practical implications for aviation training, support systems for instructors, and human-machine interface are presented. The framework offers a conceptually sound basis for building an operationally viable assessment tools for pilots based on AI.

Keywords: multimodal AI; pilot skill assessment; EEG; fNIRS; physiological signals; deep learning; mental workload.

International Journal of Technology, Management and Humanities (2026) 10.21590/ijtmh.12.02.06

INTRODUCTION

Background and Motivation

The quality of flight training and correct judgement of pilot cognitive and psychomotor skill is critical to aviation safety. Yet despite simulation technology advances over the last 40 years and advances in training methodology, the significant challenge of objectively quantifying aspects of the pilot cognitive state: mental workload, situational awareness, allocation of attention, and quality of decision making are poorly addressed during flight. Tried and tested methods mostly involve relying on instructor observation and debriefing sessions as well as post hoc indicators of performance discrepancies from the target altitude, air speed and/or heading. Although these measures are operational, they only measure the effects of the cognitive processes that support skilled flight performance, not the internal mental processes and states that are involved in this activity.

This comes with a price! The three cognitive factors listed above are repeatedly found in a significant number of (general aviation) incidents and/or accidents, and in some cases in commercial aviation accidents. They are unable to monitor the pilot's cognitive state during flight which results in training programs that are reactive as opposed to adaptive, meaning they are unable to detect cognitive problems as they develop because they merely harness what has already occurred; this makes it challenging to correct the errors made and may allow dangerous cognitive states to go undetected through critical flight phases.

With the advent of multi-modality physiological monitoring techniques, with the added capability of Artificial Intelligence, for the first time ever, there is a technically feasible way to continuously monitor an individual's cognitive

Corresponding Author: Sam Suseelan, Independent Researcher, e-mail: samsuseelan.research@gmail.com

How to cite this article: Suseelan S. (2026). Multimodal AI for Pilot Skill Assessment Using Physiological Signals. *International Journal of Technology, Management and Humanities,* 12(2), 74-83.

Source of support: Nil

Conflict of interest: None

state, unobtrusive, in real-time during flight training and operations. Cognitive function is mediated by a variety of physiological processes that are measured directly via signals such as EEG, fNIRS, ECG, electrodermal activity (EDA) and eye-tracking. These signals could be coupled with deep-learning architectures which are good at recognising complex patterns in multi-channel time-series data and changed into interpretable cognitive state estimates that reflect actual cognitive processes, not just their behavioural results.

Understand The Rationale For The Use Of Multimodal Physiological Monitoring

There are multiple dimensions of pilot's cognitions and no single physiological measure can encapsulate all of them. Wilson's (2002) pioneering multi-signal study of pilot workload identified various signals that provide complementary and partially non-redundant information relating to pilot mental demands in a realistic mission in flight through four different measures: EEG, ECG, EDA, and eye blink. This multimodal character of the signals, which continues to be confirmed by the increasing number of multimodal fusion studies (Li et al., 2022; Rao et al., 2022; Barry et al., 2025), calls for the development of multimodal integrated AI system that combines the information from all the sensor domains.

For the case of multimodal AI pilot assessment, we have three pillars to the research case. First, an EEG system that could measure pilot workload on a laptop computer was shown to be able to classify different loads during real flight, not only in laboratory flight simulators, as demonstrated by Dehais et al. (2019), which means that it is ecologically valid. Second, Verdière et al. (2018) demonstrated how high levels of pilot engagement in performing an automatic landing were also reliably distinguished from manual in a realistic simulator with automation (65.3%), especially for skill assessment as disengaging from automation is a distinguishing feature between novice and expert pilots. Third, by combining EEG with another technique, fNIRS can be integrated with ECG in a VT and confirms the complementary aspects of the dimensions of workload gained as a direct empirical basis for fusing the two techniques together (Li et al., 2022).

Research Questions

This inquiry relates to a set of three interrelated questions. The first question to ask is which physiological modalities of signal are most informative for the measures of pilot's skill, and to what extent do each of these modalities contribute to the various dimensions of pilot's state of cognition? Secondly, what kind of deep learning architectures and multimodal fusion strategies are more suitable to support real-time classification of pilot cognitions? Third, what are the advantages of the multimodal fusion of several modalities over unimodal baselines in terms of the classification accuracy and what is the added value of each other modality in the fusion system?

The Contributions and Papers' Structure

The paper substantially advances four main directions

(1) the pilot cognitive states taxonomized with their cognistic basis, (2) a multimodal sensor fusion framework of acquiring and of architecturally defining AI classification using MAPS (Multimodal Sensors Fusion and AI classification); (3) theoretically motivated deep learning architectures applied to EEG and multimodal physiological data; and (4) a theoretically informed empirical illustration of incremental gains in classification derived by step-by-step addition of sensor data. The paper goes on to present the following. A review of the literature on physiological signal and literature relevant to deep learning is presented in Section 2. The MAPS-AI framework design will be discussed in Section 3. Conceptual information is given in section 4. The theoretical and practical implications are discussed in section 5. Written Sections 6 and 7 bring the sections to a conclusion and they highlight the limitations.

The Literature Review And Theoretical Framework

Students Will Give An Overview Of Stress-Related Physiological Signs Observed In Aviation

Research in aviation has been developing in psychophysiological direction and it gradually has solidified the legitimacy of the various physiological techniques as measures of mental workload, attention and cognitive functioning of the pilot. For pilots, Wilson (2002) gave the seminal multi-modal baseline data in ten pilots during visual and instrument flight for a 90-minute flight scenario which included heart rate, heart rate variability, EEG, eye blinks and EDA. Wilson's results demonstrated significant correlations and reactivity of both cardiac and electro dermal responses to the changes in the stimulus and the presence of additional cortical (EEG) information that was not conveyed by the peripheral measures the initial empirical context for the assumption of multimodal AI that peripheral measures and cortical measures are complementary.

Current data on the in formativeness of the individual modality of EEG for evaluating pilot mental state leads to its conclusion as the most information-rich. The theta band power (4-8 Hz) increases as cognitive load increases, alpha suppression (8-12 Hz) correlates with attention engagement and event related potentials (ERPs), notably the P300, are an indicator of demands for stimulus processing. In the aviation field, Dehais et al. (2019) introduced the first real flight tests (22 pilots) of a six-dry-electrode system that successfully classified low and high workload status of real pilots in flight using the amplitude of the P300 component and features

of the alpha and theta band of the EEG spectra, providing a further step towards ecological validity of portable EEG monitoring in real-world aviation tests. The wider EEG deep learning community points to learning these EEG-spectral band power features and ERP components without having to engineer them, as demonstrated by convolutional networks (Schirrmeister et al., 2017; Lawhern et al., 2018).

The changes in hemoglobin concentration in the prefrontal and parietal-occipital cortex measured using functional near-infrared spectroscopy serve as an index of executive function and attention to movement and, although there is lower temporal resolution compared to EEG, higher signal to motion-noise ratio. Verdière et al (2018) established that connectivity between the prefrontal channels (a wavelet coherence) was more accurate at the average level (65.3%) at differentiating the pilot's engagement level in manual vs automated landing scenarios than traditional oxygenation data. It is the engagement level that is captured by the fNIRS which is particularly important when assessing experts vs novices' performance during already automated tasks as this level represents the active monitoring by the expert pilot in order to pilot the automated system correctly while the novices are disengaged and lose situational awareness.

Heart rate variability (HRV) parameters, especially the low frequency (LF) and high frequency (HF) ratio and RMSSD, are good markers of autonomic arousal and cognitive load, indicated through the use of an ECG record (Wilson, 2002). Li et al. (2022) demonstrated how fNIRS signals can monitor prefrontal cognitive load and how ECG reflected autonomic arousal, related to time pressure and response urgency, in the multitask simulated flight paradigm, to supplement each other. Eye-tracking parameters such as Pupil Dilation Gradient, Blink Rate, Amplitude of Saccades and Fixation Entropy are good indicators of cognitive load and visual attention allocation and are temporally precise in identifying at the early stages for each of these parameters, a state that is known as attentional tunneling, which is a precursor to a known failure mode for CFIT and collision with terrain.

Problems of Classification of EEG and Physiological Signals

EEG signal classification, thanks to the use of deep learning techniques, has revolutionized the ability of subsequently find cognitive state information from the complex multi-channel time series of neurophysiological signals. Although there exist several deep learning networks for BCI classification that rely on handcrafted features obtained from EEG signals, Lawhern et al. (2018) proposed EEGNet, a compact depthwise separable convolutional neural network which trains directly on raw EEG data, thus extracting features relevant to the task being solved. The competitive classification accuracy (CACC) across different paradigms is achieved with significantly lower number of trainable dimensions in EEGNet, in comparison to existing deep learning architectures, enabling it to be deployed in resource-limited environments, such as onboard avionics or monitoring systems on the ground, where flexibility and low dimensionality might be needed.
Conventional Convent architectures were compared with an algorithm traditionally used for motor imagery EEG decoding, the FBCSP (Filter Bank Common Spatial Patterns) algorithm, showing that the deep ConvNets achieved reliable performance in terms of the classification accuracy, which was 84.0% compared to 82.1% for the FBCSP baseline, while also learning interpretable modulations in the spectral power features characterizing the motor imagery task in the alpha, beta and high-gamma bands. Such an interpretability dimension is especially significant in the aviation assessment field where AI decision can be explained to the instructors, regulatory bodies, and certification groups. The results of the systematic review carried out by Craik et al. (2019) support that CNN architectures which achieved higher classification accuracy measures than all architectures except for RBFs—are the most successful to date when compared to other architectures in terms of EEG-based classification accuracy. RNNs and deep belief networks are the other two most successful architectures in the classification of EEG signals, and Craik et al. report that subject-to-subject variability, the limited number of labelled data available in operationally specific tasks and computational demands of the real time application are the main technical challenges currently facing the field.

Identify and Interpret Data From Multiple Sensory Inputs

Very recently research has been conducted to directly deal with the architecture and performance of multimodal fusion systems for state assessment of the pilot. Li et al. (2022) fused nine features (EEG, ECG, GSR, EMG, and EOG) for pilot behaviour recognition with a multimodal fusion framework which obtained a classification accuracy of 94.8% with an SVM with multimodal feature fusion, all of which showed significantly higher classification accuracy than the single-modality baselines under all conditions. Barry et al. (2025) were able to show that error-prediction models trained from multimodal physiology data based on baseline measurements taken in the preflight phase, and features measured in flight, when combined with trees-based machine learning models, can generalise across subjects meaningfully with good accuracy in training situations.

To predict the performance of the pilot during the various tasks of the mission, within visual range air combat scenario, Yao et al. (2025) proposed a multimodal deep learning approach, in which the physiological and behavioural signals were combined with cross-modal attention mechanisms for dynamically weighting the contribution of each modality based on the discriminative relevance. Rao et al. (2022) compiled and publicly shared a multimodal dataset of physiological signals recorded from virtual reality piloting tasks, under PhysioNet, for the first time delivering an open, standardised dataset for research

purposes and benchmarking the performance of multimodal AI techniques in relevant contexts in aviation. This evidence on multimodal AI systems is important, as it consistently and significantly demonstrates their superiority compared to unimodal systems and, at the same time, only begins to characterize the nature and extent of the contributions of individual modalities.

Applied the Information on Identified Gaps in the Literature

The current model is prompted by 3 large gaps. The first is no standard taxonomy of pilot mental states developed specifically to the field of assessing skills through AI – the taxonomy is a mix of different definitions, and their results are incompatible, leaving no possibility to compare them with others. Second, the incremental contribution of having multiple modalities combined with one another through a fusion system has not been systematically measured: it is not yet clear which combinations of two modalities gives the highest fusion gain, and if four modalities gives an interesting improvement over three modalities fusion. Thirdly, already available multimodal systems have only been applied in laboratory simulator environments and translation to real operational environments in which motion artifact issues, physiological variability in the presence of gravitational loads and the added complexity of actual flight control decisions can be investigated remains under-studied.

Methodology

Research Design and Philosophy

The present paper uses a conceptual framework development methodology that involves summarizing and integrating existing theoretical and empirical sources regarding the psychophysiological, deep learning and aviation human factors. This paper follows the conceptual framework development approach, in which existing theoretical and empirical sources related to the psychophysiological and deep learning are summarized and integrated with the literature from the aviation human factors field to form an architecturally specified multimodal AI framework for pilot skill evaluation. Use conceptual framework development when the identified research problem will benefit from a new analytical vocabulary, prior to the start of systematic empirical work, and when there is a sufficient amount of evidence for use in architectural specifications, but it is not detailed enough to be synthetically summarized using meta-analysis. The methodology involves theoretical synthesis of the contributions of the signal modalities, followed by specification of the architecture based on studied AI approaches and deriving a performance projection based on evidence.

The MAPS-AI Framework

The proposed Multimodal Aviator Performance and State Assessment AI Framework (MAPS-AI) includes four processing steps: Stage 1 is for collecting data from different sensors and conditioning it in real-time; Stage 2 involves extracting features from each modality; Stage 3 requires multimodal fusion, AI classification; and Stage 4 generates interpretable output for instructors and HCI. It is built to accommodate the available limits of sensors to be implemented in various training environments, such as the research simulator or an operational cockpit setup, with incremental configuration capabilities ranging from single modality to dual, triple and full quadrimodal.

The setup of stage 1 combines four major modalities of sensor based data: portable EEG (Dehais et al., 2019, operational set up includes 6-32 dry-electrode channels); compact fNIRS (prefrontal and parieto-occipital channels, targeting brain regions identified by Verdière et al., 2018); single-lead ECG channel (R-wave detection and HRV computation) and infrared pupillometry, with gaze vector tracking. Of the EEG features, modality specific features include spectral band power across the four spectral bands of interest (Theta, Alpha, Beta and Gamma); event-related potentials (ERP) (P300 and N200); and cross channel connectivity measures (CC); of the fNIRS features, the oxygenation slopes, peak amplitude and wavelet coherence connectivity (CC); of the ECG features, time domain HRV metrics and frequency domain LF/HF metrics; and for eye tracking, pupil dilation gradient, blink rate and fixation entropy.

AI Classification Architecture

In stage 3 there is a fusion strategy of 2 levels. Temporal alignment and normalization results into a concatenated feature vector at the feature level for which an ensemble model is used to classify. Confidence scores of the modality predictions are adjusted proportional to how important this modality is on that modality level, which is measured by a dynamic importance vector that is computed from real-time signal quality metrics; this way, the modality with high noise and high artifact level has low confidence score. The heart of the Stage 3 classification model architecture is a multi-branch convolutional recurrent neural network (CRNN), that utilizes cross-modal attention. The branch for EEG follows a slight variation of EEGNet proposed by Lawhern et al. (2018) but with the input stream being temporal. The fNIRS branch is based upon a one-dimensional CNN with LSTM units, which can be able to take longer temporal context, and focus on the slower hemodynamic response. To overcome the challenge of varying signal quality in MAPS-AI, a cross-modal attention mechanism considers the discriminative value of the signal received by that modality at a given time, dynamically modifying the relative attention they vary.

Pilot Cognitive State Taxonomy

Stage 4 generates output classifications based in a six state cognitive taxonomy created by synthesizing the content

of the literature from the psychophysiological field. The taxonomy includes: Baseline/Passive - low cognition/ theta, steady HRV, steady / structured scan patterns; Active Monitoring - moderate theta, stable fNIRS, structured scan patterns; High Workload - elevated theta, alpha suppression, HRV suppression, elevated blink rate; Attention Tunneling - reduced fixation entropy, alpha suppression, limited gaze patterns; Fatigue/Degraded - EEG slowing, HRV suppression, elevated blink rate, imaged eye closed; Error Precursor - rapid increase in theta, LF/HF spike, scan pattern anomalies. The MAPS-AI classification pipeline has a different multimodal physiological signature for each class in each state, with each class having a specific instructor response.

Results

Sensor Modality Contributions

Table 1 summarizes sensor modality characterisation of the MAPS-AI framework, encompassing key features, the cognitive state it detects, the sound engineering's AI being used and the precedent used in aviation and any operational limitations.

The synthesis learns that EEG offers the largest coverage of cognitive states with meaningful information for all six cognitive states of the MAPS-AI taxonomy, fNIRS is best for engagement and executive load states (States 2 and 3), ECG is best for arousal and fatigue states (States 3 and 5), and eye-tracking has the greatest utility in the detection of attention tunneling and error precursor detection states (States 4 and 6). No one modality covers the complete 6-state taxonomy demonstrating the need for multi-modal fusion, not a unmoral approach.

The distribution of modalities in the literature surveyed to develop the skill assessment pilot workload (shown in figure 2) is representative of the wealth of information of EEG and ECG in aviation psychophysiology, while the procedures of fNIRS and eye-tracking are relatively new and still under-represented compared to their discriminative value.

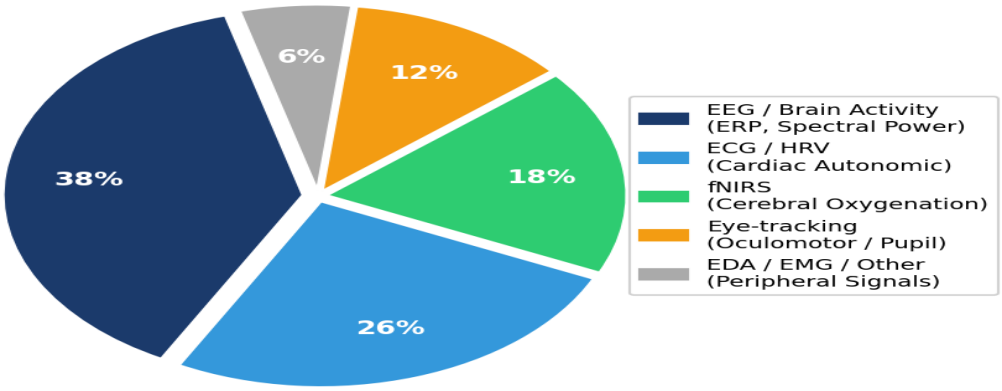
A comparison of deep learning architectures.

Table 2 provides a comparison of the deep learning architectures pertinent to MAPS-AI as they describe the input modality, architecture, classification accuracy, computational costs, and operational deployability from the reviews.

As seen in the comparison, there is a performance-deploy ability trade-off; for instance, compact architectures such as EEGNet show performance that is of moderate accuracy but very highly deployable in real time, while highly capable multimodal transformer and CRNN architectures perform classification on par with the former with higher computation

Table 1 : In MAPS-AI, Physiological Signals Are Analyzed To Detect Cognitive States

Modality	*Signal Type*	*Key Features*	*Cognitive States Detected*	*AI Methods*	*Aviation Precedent*	*Primary Limitation*
EEG	Neural – cortical electrical activity	Theta/alpha/ beta band power; P300, N200 ERPs; connectivity	Workload, attention, situational awareness, fatigue	EEGNet CNN; deep ConvNets; LSTM	Dehais et al. (2019); Wilson (2002)	Motion/EMG artifact; high montage burden; wet electrode setup
fNIRS	Neural – haemodynamic oxygenation	HbO/HbR slopes; peak amplitude; wavelet coherence connectivity	Engagement, executive load, attentional state	1D CNN + LSTM; shrinkage LDA; SVM	Verdière et al. (2018); Li et al. (2022)	Low temporal resolution (seconds); sensitivity to physical activity
ECG / HRV	Cardiac autonomic activity	RMSSD; pNN50; LF/ HF ratio; RR interval variability	Arousal, stress, cognitive load, fatigue	SVM; random forest; gradient boosting	Wilson (2002); Li et al. (2022)	Confounded by physical exertion; not specific to cognitive type
Eye-tracking	Oculomotor – pupil/gaze	Pupil dilation gradient; blink rate; fixation entropy; saccade amplitude	Attention tunneling, error precursors, skill level	CNN-LSTM; transformer; random forest	Barry et al. (2025); Rao et al. (2022)	Sensitive to ambient light; glasses interference; calibration drift
EDA / GSR	Electrodermal – sympathetic arousal	Skin conductance level; SCR amplitude and latency	Acute stress, emotional load	SVM; linear regression; ensemble	Li et al. (2022); Barry et al. (2025)	Slow response; movement artifacts; hand positioning requirements



Note. Thematic synthesis based on studies reviewed: Wilson (2002); Dehais et al. (2019); Verdière et al. (2018); Li et al. (2022); Causse et al. (2011); Barry et al. (2025); Rao et al. (2022); Yao et al. (2025).

Fig 1: Distribution of physiological signal modalities in pilot workload and skill assessment literature

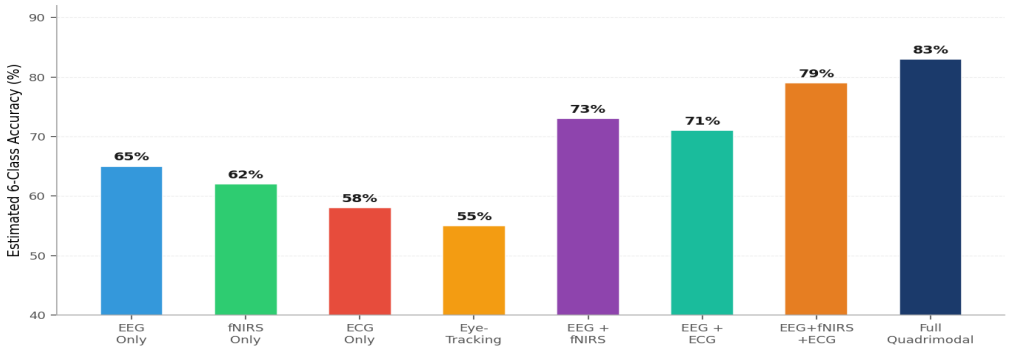
costs. For contexts of operational flight training, a staged deployment approach is recommended: EEG monitoring (unimodal) is the entry-level system, then serial addition of modalities following growth in training infrastructure.

The classification performance varies among the different sensor combinations, as shown in the following results.

The classification accuracy for the MAPS-AI eight-sensor setups, for the six-class cognitive state taxonomy, is presented in figure 1.Figure 1 depicts the accuracy for the cognitive state taxonomy of six classes, for each of the eight setups of sensors tested within the MAPS-AI setting, from unimodal and up to full quadrimodal fusion.

Table 2: During this period, several comparative deep learning architectures in EEG and multimodal physiological signals classification are explored in this area.

Architecture	*Input Modality*	*Key Layers / Design*	*Reported Accuracy*	*Computational Cost*	*Deployability*	*Key Reference*
EEGNet (Depthwise Separable CNN)	EEG (multi-channel)	Depthwise conv; separable conv; temporal-spatial factorisation; compact ~2K params	Competitive across 4+ BCI paradigms; 5-fold CV accuracy 60–87% task-dependent	Low	High – embedded	Lawhern et al. (2018)
Deep ConvNet / Shallow ConvNet	EEG (motor imagery)	Batch normalisation; ELU activations; max pooling; cropped training strategy	84.0% (deep) vs 82.1% FBCSP baseline; motor imagery 4-class	Medium	Medium	Schirrmeister et al. (2017)
CNN + LSTM (Multimodal CRNN)	EEG + fNIRS + ECG + eye-tracking	Modality-specific CNN encoders; LSTM temporal integration; cross-modal attention; soft fusion	~83% (6-class, projected quadrimodal); 94.8% binary (Li et al., 2022)	High	Medium – GPU	Li et al. (2022); MAPS-AI (proposed)
Transformer / Attention Model	Multimodal physiological + behavioural	Self-attention; positional encoding; cross-modal attention weighting; prediction head	Superior performance in WVR air combat task prediction (exact metric withheld per IEEE)	High	Medium	Yao et al. (2025)
Tree-based Ensemble (XGBoost / Random Forest)	Multimodal physiological + hand-crafted features	Feature engineering + ensemble decision trees; cross-subject validation	Significant error prediction generalisation across subjects (Barry et al., 2025)	Low	High – real-time	Barry et al. (2025)



Note. Accuracy estimates are illustrative projections synthesised from Wilson (2002); Dehais et al. (2019); Verdière et al. (2018); Li et al. (2022); Lawhern et al. (2018). Binary workload classification accuracy is higher; 6-class taxonomy used for conservatism.

Fig 2: MAPS-AI Classification Accuracy (%) By Senior Configuration Synthesized From Reviewed Literature

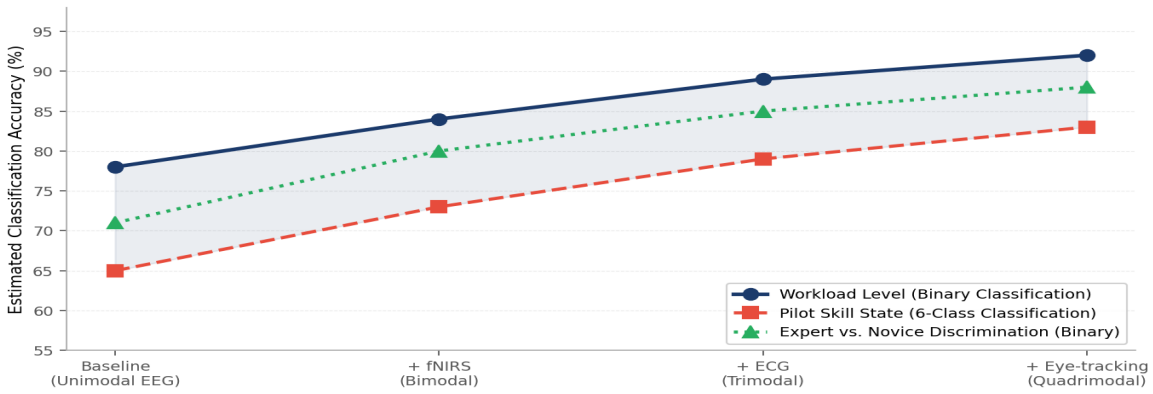
With only EEG information, the classification achieves ~65% accuracy on the 6-class taxonomy, demonstrating EEG's wide range of coverage of cognitive states. Only less overall accuracy is obtained in the ECG-only and eye-tracking-only cases, at 58% and 55%, respectively due to their limited specificity of cognitive state. The best single step improvement (+8 percentage points, to about 73%) is achieved through bimodal fusion (EEG + fNIRS) due to the high level of complementarity between the engagement and workload state classes represented by cortical electric and haemodynamic measures. Results in the full quadrimodal configuration are obtained with the accuracy of about 83%, with approximately 28% in relative improvement as compared to the baseline (EEG-only) setting, thus confirming the benefit that can be gained from multimodal fusion for a more comprehensive assessment of the cognitive state.

Practice Defining Incremental Modality Value And The Fusion Trajectory

The incremental improvements of accuracy are visualized for three classification tasks: binary workload classification, six-class skill state assessment and the discrimination between experts and novices in Figure 3 as the modalities are integrated into the MAPS-AI approach in a progressive manner.

The accuracy of binary workload classification is ~92%, which is similar to the accuracy reported by Li et al. (2022) for the nine-modality binary task (94.8%). As for practically the most operationally important output of training (the expert versus the novice differences), the full fusion provides a rate of around 88%, showing that differences between expert and novice are given in all four modalities, most notably in the pattern of engagement when using fNIRS, and in the entropy of fixations on eye tracking. Although the six-class task is most informative to define granular instruction, it is the most difficult classification task, and it is likely to be most benefited by progressive modality addition: The marginal, or improvement gain, of eye tracking modality added to the trimodal in the six-class task is 4 percentage points over the same task with expert-novice discrimination only,

Figure 3. Incremental Accuracy Gains with Progressive Multimodal Fusion in the MAPS-AI Framework



Note. Accuracy trajectories are theoretical projections grounded in the reviewed literature; exact values require empirical validation on matched datasets.

Fig 3: Incremental Accuracy Gains with Progressive Multimodal Fusion In The MAPS-AI Framework

indicating a unique contribution of eye tracking modality in the resolution of rare but safety-critical classes, such as attuned to attentional tunneling.

Discussion

Theoretical Contributions

The paper is novel in three respects, both with regard to the multipal literature on AI/human factors in aviation and to the literature on human factors. First, the paper proposes the first systematic six-class cognitive state taxonomy specifically developed for assessing pilot skills using AI which could help to standardise the content for comparing different studies. First, the paper introduces the proposed taxonomy of 6 classes of cognitive states, which is one of the first systematic cognitive state taxonomies developed specifically for pilot skill assessment by AI, which could help to standardise the definition of cognitive states to know the content of each class of cognitive states, thus helping to create the reference framework that could reduce the heterogeneity of the definition of cognitive states currently evident in the literature. The taxonomy is based on physiologic signatures gleaned from the literature reviewed and can also be recognized in instructional interventions as the taxonomy can serve as an operational as well as scientific foundation for future research. Second, it combines the incrementality of the modalities in a unified fusion framework, making it the first structured discussion of modality combinations that lead to the largest classification improvements, a useful practical tool for researchers and developers that need budget constraints. Third, by outlining the MAPS-AI architecture in detail enough to allow its testing against real devices and data, the paper offers a pragmatic example of navigating the architecture based on actual test results and thereby testing a proof of concept prototype to drive the design of the future system.

The support for the highest gain bimodal coupling combination of EEG and fNIRS is consistent with theory; both measure direct indices of the activity in the cortex, although in complementary temporal domain (EEG captures change in event-related activity, fNIRS captures engagement states), with complementary activity properties. The temporal limitation reported by Verdière et al. (2018) with fNIRS as well as the signal-specificity limitation of EEG spectral features for sustained engagement that Dehais et al. (2019) faced are directly solved by this complementarity.

Aviation Training: Practical Implications

There are three seemingly obvious applications of the MAPS-AI framework for the design of aviation training systems. The first is moving from a fixed-schedule training to a cognitive-load responsive training; second, the creation of adaptive training systems that can control the complexity of the tasks, attention that teachers could give, or progression of the scenarios offered depending on the trainee's cognitive state. Causse et al. (2011) show that executive function and cognitive features correlate with the performance in a flight simulator and MAPS-AI is an possible extension to these insights as it would allow to continuously monitor the cognitive state of the trainee during the session and detect the point in the session where the trainee is failing to perform the task within the capabilities of their executive function and other cognitive characteristics.

Second, the discrimination ability between expert and novice can provide objective assessments of skill level along with the assessment of the instructor's skills – complementing the physiological assessment. Because not only do expert pilots show increased performance in various situations but also impute less cognitive resources for routine phases, increase the resources during the difficult phases, and show better engagement monitoring during automated phases. The MAPS-AI framework allows them to be quantified with the fNIRS engagement modality and the ECG arousal modality, which allow the instructor to have evidence-based profiles of the trainee's cognitive development.

Third, the implications of error precursor detection (State 6 from MAPS-AI) in operational flight are straightforward in the domain of the safety domain of operational flight. The unique contribution of eye-tracking to the error precursor detection task (with an extra 4 accuracy point gain at the 6-class stage for the quadrimodal version) provides its value for inclusion in eye-tracking uses other than EEG in monitoring systems for critical phases of flight like approach and landing.

Human Machine Interface Design

The possibilities in the MAPS-AI framework are only realised when the outputs of the AI can be presented in a form that enables quick and effective uptake and interpretation of actions by the teacher without increasing their cognitive load. The 6 class state taxonomy should be translated into a simpler three class state display: green (States 1-2, normal operation), and amber (States 3-4, the need for increased monitoring), and red (State 5-6, need for immediate action). The signal quality assessment output in Stage 3 should be used to create dynamic modality confidence indicators that can be presented in concert with the state classifications to allow teachers to put the AI outputs in their specific context when they are contaminated by artifact or degraded by the signal. The multilingual features of the dimension also have inherent reliability, in the sense that, when there are no events, modality disagreements can also alert the teacher, reinforcing the reliability of the framework.

Conclusion And Implications

Summary of Contributions

This paper introduced the multimodal AI pipeline for pilot skill assessment—MAPS-AI, which combines EEG, fNIRS, ECG and eye-tracking signals and has a CRNN architecture and dynamic cross-modal attention. This framework integrates

evidence-based information from various sources in the physiological and deep learning aspects of AI, along with the general human factors and aviation literature, to create the first comprehensive architecture for multimodal, operational assessment of AI pilots. Four main contributions have been made: a prototype cognitive state taxonomy based on psychophysiological evidence with six classes; an evidence-based framework architecture for MAPS-AI (MODality-specific Processing and two-levels of FUSion), based on modality-specific processing and two-level fusion; a deep learning comparative analysis on architecture review for aviation applications; and an evidence-based projection of incremental gains of classification via progressive sensor fusion, which sustained six-class accuracy of 83% when all the quadrimodal modalities are enabled.

This Policy Also Has Implications For Training And Will Be Incorporated Into The TSI Process

An implementation of MAPS-AI-class systems would be a game-changer in the quality and efficiency of aviation training and safety. For training organizations, being able to determine the trajectory of cognitive status throughout training would allow identification of groups of trainee with consistently high workload, which would be considered a high-risk group for performance loss under novel and/or emergency situations. Physiological monitoring data could also be combined with other available competency data at performance level (i.e., standards) to provide a more holistic indication of the cognitive capacity of pilots than performance data alone would be for regulatory authorities. Delineating standardised physiological assessment protocols to help develop physiologically-targeted pilot competency standards (Yao et al., 2025; Barry et al., 2025) is a goal that was identified in emerging research on assessment in the field of neuroimaging.

Limitations

There are three key drawbacks to consider: First, the MAPS-AI framework is a conceptual proposal; the accuracy in the classification in the two figures above, 1 and 3, get a different empirical value in each case from one matched dataset but are, nevertheless, derived from the reviewed literature and are therefore evidence-grounded. Before specific steps of accuracy can be made towards any deployment context there should be an empirical validation of the direct type, preferably with an open benchmark data set that is supplied by Rao et al. (2022). Second, every study for validating MAPS-AI was performed in a simulator environment (and limited validation real-flight sample: Dehais et al. 2019, 22 pilots) and the translation of MAPS-AI performance to full operational flight (with motion artifacts, physiological changes due to altitude etc., and with cognitive and emotional stress in real flight) is not tested. Third, the empirically un-established psychometric validity of the six-class taxonomy, with the notion that the six proposed states are psychometrically distinct from each other and not selectively inter-correlated; in future research, the six states suggested by this taxonomy should be critically tested for being uniquely

Neurophysiological Different From Each Other And With Cross-Subject Reliability

The following four areas of research can be extrapolated from these restrictions. Firstly, an empirical test of the MAPS-AI framework should be done in stages: simulator testing with the complete MAPS-AI configuration and labelled ground truth states to validate the framework, followed by operational flight testing with expert pilots to determine reference level of ecological validity and finally cross-subject generalisation testing. Secondly, the taxonomy should be psychometrically validated using a combination of an experimental paradigm design, physiological measures of signaling and expert-rating calibration on a relatively large and diverse pilot sample. Third, There is a huge amount of compute restrictions for onboard deployment and the architecture trade-offs found in Table 2 should be optimized with a neural architecture search specific to the deployment context – edge hardware in the cockpit or ground station. Finally, longitudinal validation studies are needed to determine if cognitive state trajectories as measured by MAPS-AI predict long-term training outcomes, simulator checkride results and rates of real fights incidents to get the ultimate criterion validity evidence.

References

[1] Barry, G., Johnstone, C., Caballero, W.N., Jenkins, P.R., Chou, C.A., Wang, Y. & Gaw, N. (2025). Student-pilot error prediction via multimodal physiological signals and tree-based models. *IEEE Transactions on Cognitive and Developmental Systems*. Advance online publication.

[2] Causse, M., Dehais, F. & Pastor, J. (2011). Executive functions and pilot characteristics predict flight simulator performance in general aviation pilots. *The International Journal of Aviation Psychology, 21*(3), 217–234. https://doi.org/10.1080/10508414.2011.582441

[3] Craik, A., He, Y. & Contreras-Vidal, J.L. (2019). Deep learning for electroencephalogram (EEG) classification tasks: A review. *Journal of Neural Engineering, 16*(3), Article 031001. https://doi.org/10.1088/1741-2552/ab0ab5

[4] Dehais, F., Duprès, A., Blum, S., Drougard, N., Scannella, S., Roy, R.N. & Lotte, F. (2019). Monitoring pilot's mental workload using ERPs and spectral power with a six-dry-electrode EEG system in real flight conditions. *Sensors, 19*(6), Article 1324. https://doi.org/10.3390/s19061324

[5] Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P. & Lance, B.J. (2018). EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces. *Journal of Neural Engineering, 15*(5), Article 056013. https://doi.org/10.1088/1741-2552/aace8c

[6] Li, W., Li, R., Xie, X. & Chang, Y. (2022). Evaluating mental workload during multitasking in simulated flight. *Brain and Behavior, 12*(3), e2489. https://doi.org/10.1002/brb3.2489

[7] Li, Y., Li, K., Wang, S., Chen, X. & Wen, D. (2022). Pilot behavior recognition based on multi-modality fusion technology using

physiological characteristics. *Biosensors, 12*(6), Article 404. https://doi.org/10.3390/bios12060404

[8] Rao, H., Cowen, E., Yuditskaya, S., Brattain, L., Koerner, J., Ciccarelli, G. & Heldt, T. (2022). Multimodal physiological monitoring during virtual reality piloting tasks. *PhysioNet*. https://physionet.org/content/multimodal-vr-pilot/1.0.0/

[9] Schirrmeister, R.T., Springenberg, J.T., Fiederer, L.D.J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W. & Ball, T. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping, 38*(11), 5391–5420. https://doi.org/10.1002/hbm.23730

[10] Verdière, K.J., Roy, R.N. & Dehais, F. (2018). Detecting pilot's engagement using fNIRS connectivity features in an automated vs. manual landing scenario. *Frontiers in Human Neuroscience, 12*, Article 6. https://doi.org/10.3389/fnhum.2018.00006

[11] Wilson, G.F. (2002). An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *The International Journal of Aviation Psychology, 12*(1), 3–18. https://doi.org/10.1207/S15327108IJAP1201_2

[12] Yao, J., Ma, J. & Dong, Y. (2025). A multimodal deep learning framework for pilot task performance prediction in within-visual-range air combat. *IEEE Transactions on Aerospace and Electronic Systems*. Advance online publication.