IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 8 November 2023, accepted 27 December 2023, date of publication 3 January 2024,
date of current version 11 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3349495

TOPICAL REVIEW

Advancing Aviation Safety Through Machine Learning and Psychophysiological Data: A Systematic Review

IBRAHIM ALRESHIDI[1,2,3], IRENE MOULITSAS[1,2], AND KARL W. JENKINS[1]

[1]Centre for Computational Engineering Sciences, Cranfield University, MK43 0AL Cranfield, U.K.
[2]Machine Learning and Data Analytics Laboratory, Digital Aviation Research and Technology Centre (DARTeC), Cranfield University, MK43 0AL Cranfield, U.K.
[3]College of Computer Science and Engineering, University of Ha'il, Hail 81451, Saudi Arabia

Corresponding authors: Ibrahim Alreshidi (ibrahim.alreshidi@cranfield.ac.uk) and Irene Moulitsas (i.moulitsas@cranfield.ac.uk)

ABSTRACT In the aviation industry, safety remains vital, often compromised by pilot errors attributed to factors such as workload, fatigue, stress, and emotional disturbances. To address these challenges, recent research has increasingly leveraged psychophysiological data and machine learning techniques, offering the potential to enhance safety by understanding pilot behavior. This systematic literature review rigorously follows a widely accepted methodology, scrutinizing 80 peer-reviewed studies out of 3352 studies from five key electronic databases. The paper focuses on behavioral aspects, data types, preprocessing techniques, machine learning models, and performance metrics used in existing studies. It reveals that the majority of research disproportionately concentrates on workload and fatigue, leaving behavioral aspects like emotional responses and attention dynamics less explored. Machine learning models such as tree-based and support vector machines are most commonly employed, but the utilization of advanced techniques like deep learning remains limited. Traditional preprocessing techniques dominate the landscape, urging the need for advanced methods. Data imbalance and its impact on model performance is identified as a critical, under-researched area. The review uncovers significant methodological gaps, including the unexplored influence of preprocessing on model efficacy, lack of diversification in data collection environments, and limited focus on model explainability. The paper concludes by advocating for targeted future research to address these gaps, thereby promoting both methodological innovation and a more comprehensive understanding of pilot behavior.

INDEX TERMS Aviation safety, machine learning, deep learning, mental states classification, pilot behavior, systematic review, psychophysiological signals, EEG.

I. INTRODUCTION

As the global aviation industry undergoes transformative technological advancements, the role of pilots is concurrently evolving from simply operating machinery to making critical decisions in high-stakes, dynamic environments [1]. In light of the complex nature of contemporary aviation operations, a comprehensive understanding of pilot behavior becomes paramount for enhancing aviation safety. Machine Learning (ML) technologies, particularly when integrated with psychophysiological data such as electroencephalogram (EEG), present a promising route for in-depth investigation into this vital area. These cutting-edge methodologies enable researchers to acquire nuanced insights into various facets of pilot behavior, including cognitive states and emotional responses. This paper serves as a systematic literature review, conducted in accordance with the rigorous methodological guidelines [2], [3], [4]. It aims to offer an exhaustive synthesis of existing research on the application of ML techniques and psychophysiological data for understanding pilot behavior.

The associate editor coordinating the review of this manuscript and approving it for publication was Gang Wang.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

A. IMPOTANCE OF AVIATION SAFTEY

As a critical component of modern transportation infrastructure, the aviation industry plays an indispensable role in both global commerce and individual mobility. The industry facilitates the movement of millions of passengers and vast amounts of cargo annually, thereby serving as a linchpin in the global economy. Given this scale of operation, the imperative for ensuring aviation safety cannot be overstated; the consequences of failure are cataclysmic, both in terms of human life and economic impact [5].

However, the achievement of optimal safety levels is a complex endeavor, influenced by a myriad of factors ranging from technological innovation to regulatory oversight [6]. Advances in technology have undeniably contributed to enhanced safety mechanisms, from state-of-the-art air traffic control systems to predictive maintenance algorithms that preempt mechanical failures. Nonetheless, the industry is not immune to challenges [7], [8], [9], [10]. Factors such as increasing air traffic, geopolitical tensions, and even natural disasters pose new kinds of risks that require continuous scrutiny and innovation in safety protocols [11].

Moreover, the stakes are not merely quantitative but also qualitative. A single aviation accident can have a ripple effect, undermining public confidence in air travel and triggering economic repercussions that extend far beyond the aviation sector. Regulatory bodies, therefore, are in a perpetual state of vigilance, working in tandem with airlines, aircraft manufacturers, and other stakeholders to formulate and implement safety guidelines that are both rigorous and adaptive to changing circumstances [12].

In summary, aviation safety is a multifaceted and ever-evolving concern that requires a holistic approach, embracing technological, human, and systemic factors. The high stakes involved, both in terms of human lives and economic implications, make it a subject of paramount importance that warrants ongoing research and continual improvement.

B. ROLE OF PILOT BEHAVIOUR IN AVIATION

In the intricate system of aviation safety, the role of pilot behavior emerges as a focal point, governed by an intricate interplay of cognitive processes, emotional states, and physiological responses. Pilots, situated at the nexus of multifarious human-machine interactions, bear the colossal responsibility of safeguarding not just the aircraft and its passengers, but also the integrity of the entire aviation system. Their actions, or lack thereof, can have immediate and far-reaching consequences that extend from the cockpit to the broader aviation ecosystem [13].

With the advent of increasingly automated flight systems, the role of pilots has evolved significantly. While automation has undeniably enhanced safety and efficiency, it has also engendered new forms of cognitive workload and psychological stress. Pilots are no longer solely vehicle operators but have become complex decision-makers tasked with managing an array of automated systems. They must maintain situational awareness and be prepared to intervene effectively in unexpected circumstances [14]. This shift has introduced challenges related to attention allocation, decision-making under pressure, and even ethical considerations, such as how to respond in unavoidable emergency situations.

Psychophysiological markers, such as EEG data, have emerged as invaluable tools for gaining insights into pilots' internal states, particularly during high-stakes scenarios like take-offs, landings, and emergency situations. These data types allow researchers to delve into the nuances of cognitive load, attentional focus, and emotional regulation, which are crucial for understanding how pilots make decisions under stress [15], [16].

Moreover, the role of pilot behavior has systemic implications that ripple through the aviation safety ecosystem, influencing everything from regulatory frameworks to the design of new technologies [17], [18], [19]. For example, a nuanced understanding of how pilots handle attentional tunneling could inform the design of more intuitive cockpit interfaces. Similarly, insights into emotional and physiological responses to unexpected events could be invaluable for the development of realistic training simulations.

In summary, the multifaceted and systemic impact of pilot behavior necessitates its thorough investigation. Given its complexity and far-reaching implications, it warrants not just academic exploration, but also practical, real-world applications, ideally supported by advanced methodologies like ML and psychophysiological data analysis.

C. MACHINE LEARNING AND PSYCHOPHYSIOLOGICAL DATA IN AVIATION RESEARCH

The advent of ML technologies represents a pivotal milestone in aviation research, especially in the nuanced domain of pilot behavior. These advanced computational techniques offer a comprehensive framework for analyzing intricate, high-dimensional psychophysiological data sets like EEG, which are often beyond the scope of traditional statistical methods to interpret in a meaningful manner [20].

ML algorithms, encompassing a broad array of models such as tree-based, support vector machine (SVM), and various neural networks, have proven to be immensely effective in predicting and understanding multiple facets of pilot behavior. These include, but are not limited to, cognitive workload, emotional states, and even task engagement. The capacity to leverage the voluminous and complex variables available in psychophysiological data sets speaks volumes about the transformative potential of ML in this research domain [21]. The applications of these capabilities extend far beyond academic inquiry and are making inroads into real-world applications, including but not limited to, predictive monitoring, adaptive cockpit interfaces, and even real-time decision support systems.

Furthermore, the confluence of ML with psychophysiological data yields an interdisciplinary approach that capitalizes

on the strengths inherent in both domains. Psychophysiological data provides an unparalleled window into the complex internal states of pilots, including cognitive and emotional variables [22]. ML, on the other hand, serves as the analytical framework capable of extracting granular insights from this data. This synergistic relationship has given rise to groundbreaking studies that have significantly extended our understanding of human performance and decision-making within aviation contexts [23], [24], [25], [26], [27], [28].

The structure of this paper is meticulously designed to provide a holistic overview of the current state of research on the application of ML techniques to psychophysiological data for understanding pilot behavior. Following this introductory section, the paper delineates its systematic review methodology, presents a comprehensive synthesis of key findings, offers an extensive discussion contextualizing these results within the broader landscape of aviation safety and pilot behavior, and concludes by summarizing the salient insights while identifying research gaps that offer promising avenues for future inquiry.

II. METHODOLOGY

The methodology of this systematic review serves as the architectural framework, designed to furnish robust, transparent, and reproducible outcomes. Adhering scrupulously to the guidelines [2], [3], [4], this section delineates the meticulous steps taken to answer the posited research questions. It provides an exhaustive description of the protocols followed in the search, selection, and analysis of literature, in addition to quality assessment. Fig. 1 presents a graphical description of the procedure.

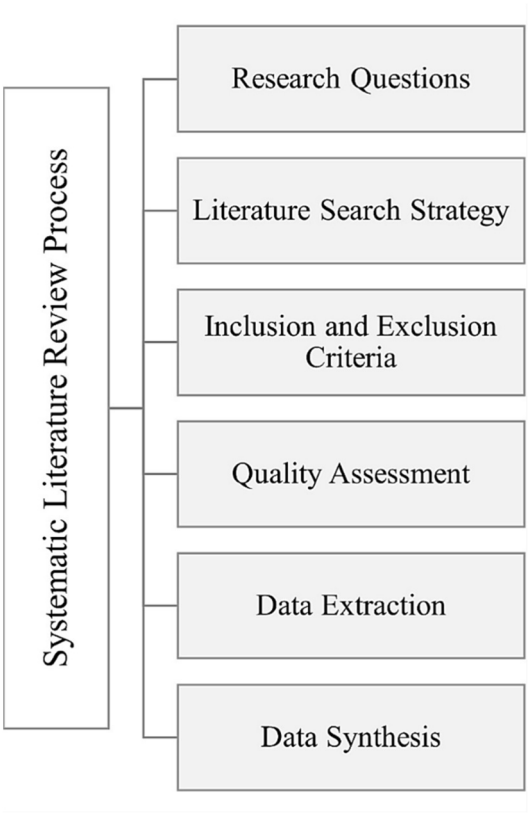

FIGURE 1. The adopted steps of the systematic review.

A. RESEARCH QUESTIONS

The present systematic review is directed by a set of carefully formulated research questions. These questions are designed not merely to clarify what is already known but to illuminate areas requiring further exploration. The principal research questions are:

- **RQ1: What are the primary focus areas in the application of ML to psychophysiological data for understanding pilots' behavior?**
  - What behavioral and cognitive states are most studied?
- **RQ2: How are preprocessing, data types, and feature extraction approached in existing studies on psychophysiological data for pilot behavior?**
  - Which psychophysiological data types are most used?
  - What artifacts are commonly found in the psychophysiological data?
  - What preprocessing techniques are prevalent?
  - What features are commonly extracted?
- **RQ 3 What are the types of models utilized to understand the pilot behavior?**
  - Which evaluation mechanism and metrics were utilized to assess the models?
- **RQ4: What is the comparative performance of various ML and DL models in predicting pilot behavior?**
  - What implications do these performance metrics hold?
- **RQ5: What are the methodological limitations in existing studies?**
  - What future research directions are suggested by the methodological limitations?

B. LITERATURE SEARCH STRATEGY

The integrity of a systematic review is profoundly dependent on the comprehensiveness and rigor of its literature search strategy. To ensure a robust selection of studies pertinent to the research questions, this review adopted a multi-faceted search strategy, encompassing several academic databases and employing a sophisticated set of search queries.

1) SEARCH QUERIES

Keywords and Boolean operators were strategically aligned to construct queries that are both expansive and incisive. Search terms were primarily derived from the research questions. Subsequently, terms related to ML were incorporated based on authoritative sources such as [29]. Phrases such as "machine learning," "psychophysiological data," "EEG," and "pilot behavior" were intricately woven together through

Boolean operators like "AND" and "OR," fashioning a search net designed for both breadth and precision.

2) ACADEMIC DATABASES

The review encompassed an exhaustive search across a selection of databases renowned for their scholarly contributions, namely IEEE Xplore, Scopus, PubMed, ScienceDirect, and Google Scholar. These databases were strategically chosen for their credibility and extensive coverage of academic articles in the fields of engineering, science, and technology. In Scopus and ScienceDirect, a comprehensive scan was conducted on titles, abstracts, and keywords for each retrieved study. For IEEE Xplore, the focus was primarily on metadata. It is worth noting that PubMed was queried by scanning both titles and abstracts, while in Google Scholar, only titles were examined. Such differentiation in search strategies was necessitated by the unique syntax and capabilities of each database. Accordingly, modifications were made to the initial search string to suit the particular idiosyncrasies of each database.

3) TIME FRAME

The time frame selected for the search reflects a balance between historical depth and contemporary relevance. A window of the last fifteen years was delineated, allowing for an appraisal of seminal works while also encompassing the most recent advancements. This temporal scope ensures that the review remains at the cusp of contemporary scientific thought.

C. INCLUSION AND EXCLUSION CRITERIA

The efficacy of a systematic review is substantially influenced by the criteria governing the inclusion and exclusion of studies. These criteria act as sieves that sift through the amassed literature, retaining articles of relevance and discarding those that do not align with the objectives of the review.

Inclusion Criteria:

1. **Peer-Reviewed Journals and Conferences:** Only articles published in peer-reviewed journals or conference proceedings were considered to ensure the research's quality and credibility.
2. **Pilot Behavior:** Research specifically targeting pilot behavior, either in real-world or simulated environments, was included.
3. **Machine Learning Models:** Studies employing ML or deep learning (DL) algorithms for data analysis were considered.
4. **Full-Text Availability:** Studies were required to be fully accessible, either through open access or institutional subscriptions, for comprehensive analysis.

Exclusion Criteria:

1. **Non-Peer-Reviewed Sources:** Articles from non-peer-reviewed sources, such as blogs, opinion pieces, or commercial publications, were excluded.
2. **Non-Aviation Contexts:** Research targeting sectors other than aviation, or general human behavior, was excluded.
3. **Non-English Publications:** Research published in languages other than English was not considered.
4. **Unspecified or Ambiguous Methods:** Studies lacking transparent methodology were excluded to ensure the integrity and reproducibility of the review.

D. QUALITY ASSESSMENT

Quality assessment is pivotal in the context of systematic reviews for ensuring that the conclusions drawn are based on rigorous and reliable studies. Each included study was thoroughly evaluated using a predetermined set of criteria:

1. **Relevance to Research Questions:** Studies were assessed based on the extent to which their objectives and outcomes align with the questions posed by this review. Those highly relevant to the review's research questions are considered to offer more meaningful contributions to the aggregated findings.
2. **Quality of Data:** The robustness of psychophysiological measures and the ML techniques used were scrutinized.
3. **Clarity and Completeness:** The level of detail and clarity with which the study's methodology and findings are presented were also considered. Well-documented studies contribute to the review's overall credibility and facilitate future replication efforts.

E. DATA EXTRACTION

The data extraction phase constitutes a critical juncture in the systematic review pipeline, serving as the foundational bedrock for ensuing rigorous analytical undertakings. This section meticulously outlines the orchestrated methodology and structured approach employed for gleaning pertinent data from the studies that met the previously established inclusion and exclusion criteria.

1) SEARCH PROCESS

To synthesize a collection of studies pertinent to the research aims, a rigorously formulated search query was executed across selected academic databases. This initial search yielded a total of 3352 potential studies for inclusion. Following this, a dedicated de-duplication process was undertaken, resulting in the removal of 2107 duplicate entries. This left 1245 studies for further examination.

Subsequently, a comprehensive screening process was carried out, wherein titles, abstracts, and keywords of these 1245 studies were meticulously evaluated against the inclusion and exclusion criteria. This narrowed down the list to 104 studies deemed potentially relevant. A subsequent full-text screening was conducted, further subjected to quality assessment protocols, leading to the exclusion of an additional 37 studies. At this juncture, the compilation stood at 67 studies.

Furthermore, to ensure a thorough and exhaustive review, the references cited in these 67 studies were also examined. This supplemental search led to the inclusion of an additional 13 studies that met the review's criteria. Thus, the final pool of studies included in this systematic review totals 80. A visual representation of this sequential selection process is illustrated in Fig. 2.

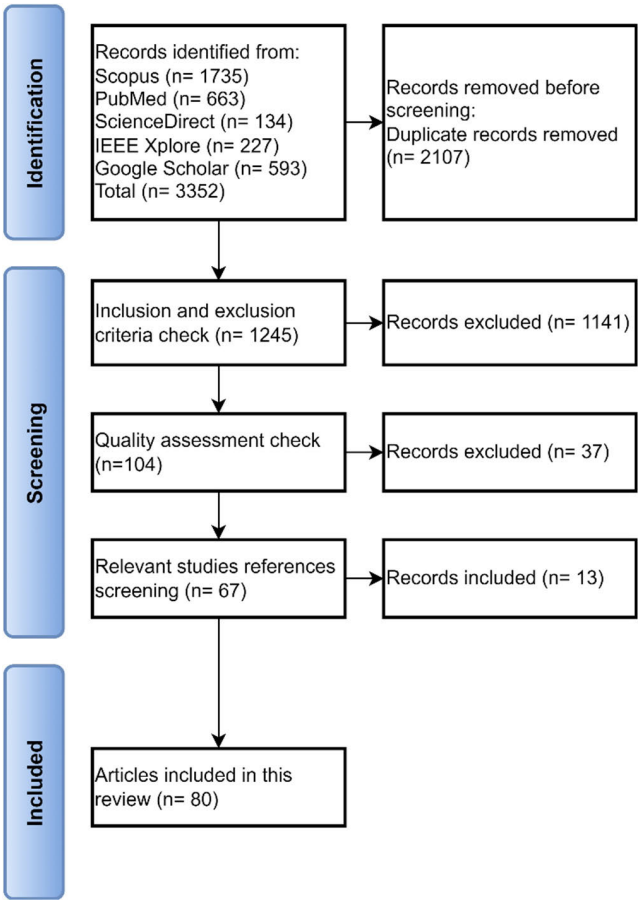

FIGURE 2. PRISMA flow diagram.

2) DATA EXTRACTION PROTOCOL

The data extraction process was designed to capture a rich set of information from each study, thereby enabling a nuanced analysis aligned with the research questions. For each study included in this systematic review, the following data were extracted:

1. **Article Title:** The title of the article was noted to provide a preliminary understanding of the study's focus and scope.
2. **Year of Publication:** The publication year was recorded to assess the temporal distribution of research efforts and to identify trends or shifts in research focus over time.
3. **Publication Venue:** The venue where the article was published.
4. **Behavioral Aspects:** Specific behavioral states or traits such as workload, fatigue, attention, and emotional states like stress or anxiety were identified and recorded.
5. **Model Type:** Information regarding the types of models employed, such as ML, DL, or Statistical Models, was extracted. This facilitated a comparative analysis of the methodologies adopted in the existing literature.
6. **Model Categories:** Within the ML models, specific categories such as tree-based models, SVM, and probabilistic models were noted to enrich the discussion on methodological diversity.
7. **Performance Metrics:** Metrics such as accuracy, recall, precision, and F1-score were extracted where available. This data aimed to provide a detailed account of the performance evaluations conducted in each study.
8. **Psychophysiological Data Types:** Types of psychophysiological data such as EEG, electrocardiogram (ECG), and galvanic skin response (GSR) were recorded to understand the range of data employed in assessing pilot behavior.
9. **Preprocessing Techniques:** Methods used for preprocessing, such as independent component analysis (ICA) or bandpass filtering, were also captured. This allowed for a comprehensive review of the techniques used to refine psychophysiological data before model training.
10. **Features Extracted:** The types of features extracted from the psychophysiological data, like power spectral density (PSD), wavelet coefficients (WC), or statistical measures, were noted. This contributed to the discussion on feature engineering practices in the existing literature.
11. **Limitations and Future Work:** An assessment of each study's limitations and suggestions for future research contribute to an understanding of gaps in the current body of literature. This information is crucial for setting the stage for future explorations.

F. DATA SYNTHESIS

The extracted data were subjected to a multi-layered synthesis process aimed at offering a nuanced understanding of the literature. The first layer involved a descriptive statistical analysis of basic metrics such as year of publication and publication types of studies. The second layer honed in on the behavioral aspects, where specific behavioral states like workload, fatigue, and attention, as well as emotional states, were analyzed. The aim was to ascertain the breadth of human performance-limiting states explored in existing literature and identify under-researched areas. The final layer of synthesis focused on the methodological paradigms employed across the studies. Models used, types of psychophysiological data, preprocessing techniques, and performance metrics were categorized and analyzed to discern prevailing trends and potential gaps.

The synthesized data were visually represented through charts and tables, facilitating a clearer interpretation and comparison of findings. Moreover, the synthesis incorporated a narrative approach, integrating the quantitative and qualitative findings to offer a cohesive and comprehensive view of the research landscape on the application of ML and psychophysiological data in understanding pilot behavior.

III. RESULTS

This section serves as the empirical focal point of this systematic review, presenting a rigorous analysis of the data extracted from the 80 included studies. Adhering to the data extraction protocol delineated in the methodology section, this segment synthesizes the findings across multiple dimensions, including the types of ML models employed, their performance metrics, and the psychophysiological data types used for predicting pilot behavior. Furthermore, this section provides a granular breakdown of methodological choices in existing literature, including data preprocessing techniques, artifacts identified, and features extracted. The results presented herein aim to offer a comprehensive understanding of the current state of the art, serving as a foundational base for the subsequent discussion section where these findings will be interpreted, contextualized, and evaluated.

A. QUALIFIED STUDIES OVERVIEW: A SYSTEMATIC ENUMERATION OF EMPIRICAL INVESTIGATIONS

In order to provide a comprehensive overview of the empirical investigations qualified for inclusion in this review, multiple criteria have been considered for categorizing the studies. An initial enumeration of the studies is presented in Table 1, which lists each study by a unique Study ID, along with its citation and title. This table serves as a systematic reference, facilitating cross-referencing throughout this review.

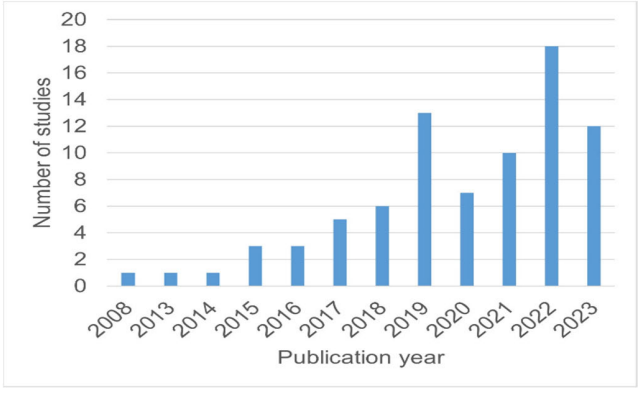

FIGURE 3. Study publication distribution using a yearly calendar.

In addition to tabulated data, Fig. 3 offers a temporal mapping of the studies, illustrating the number of publications per year. Upon examination of Fig. 3, it is evident that there has been a notable surge in the number of studies published from 2015 onwards, signaling an increased research focus on the subject matter. This could be attributed to various factors such as technological advancements, policy changes, or shifts in research priorities.

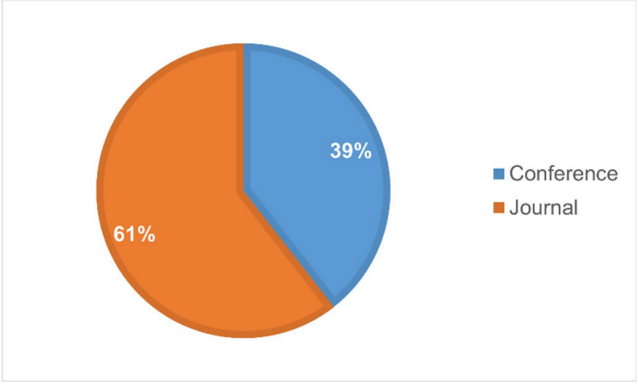

FIGURE 4. Publication type.

Fig. 4 supplements this by delineating the division between journal articles and conference papers among the selected studies. According to Fig. 4, a majority of the research is published in journal articles, which often undergo rigorous peer-review processes. The prevalence of journal articles could be indicative of the maturity and established nature of this research area.

B. TAXONOMY OF PILOT'S BEHAVIORAL AND COGNITIVE STATES

The taxonomy of behavioral and cognitive states in aviation-based empirical studies is visualized in Fig. 5, serving as a cornerstone for this analysis. It segments the research focus into five overarching categories: 'Cognitive Load Indicators,' 'Performance Metrics,' 'Attention Dynamics,' 'Emotional Responses,' and 'Miscellaneous.' Among these, 'Cognitive Load Indicators' are markedly dominant, comprising a substantial 75% of the selected studies. This predominance creates a striking contrast with the other categories, each of which constitutes a fraction of the total research corpus. Such an imbalance underscores a significant skew in existing research, leaning heavily towards quantifiable cognitive metrics.

A more granular examination reveals that within 'Cognitive Load Indicators,' 'Workload' accounts for 65% of the studies, followed by 'Fatigue' at 25%. Less represented sub-categories like 'Stress,' 'Skill Level,' 'Drowsiness,' and 'Attention Reserve' warrant attention for their minimal inclusion. In the 'Emotional Responses' domain, 'Emotion' captures 60% of the focus, with 'Reaction' and 'Situational Awareness' evenly sharing the remaining 40%. 'Attention Dynamics' is chiefly concerned with 'Distraction' at 34% and 'Attention' at 22%, but critically underrepresents performance-limiting states such as 'Diverted Attention'

TABLE 1. List of the qualified studies.

Study ID	Paper	Title
S1	[30]	Pilot Skill Level and Workload Prediction for Sliding-Scale Autonomy
S2	[31]	Classification of pilots' mental states using a multimodal deep learning network
S3	[32]	Sensing and Assessing Cognitive Workload Across Multiple Tasks
S4	[33]	Monitoring Pilot's Mental Workload Using ERPs and Spectral Power with a Six-Dry-Electrode EEG System in Real Flight Conditions
S5	[34]	A Machine Learning Approach to the Detection of Pilot's Reaction to Unexpected Events Based on EEG Signals
S6	[35]	Multimodal Approach for Emotion Recognition Based on Simulated Flight Experiments
S7	[36]	Fatigue detection in commercial flight operations: Results using physiological measures
S8	[37]	Prediction of Cognitive States during Flight Simulation using Multimodal Psychophysiological Sensing
S9	[38]	Probe-Independent EEG Assessment of Mental Workload in Pilots
S10	[39]	A Composite Cognitive Workload Assessment System in Pilots Under Various Task Demands Using Ensemble Learning
S11	[40]	Comparison between artificial neural network and multilinear regression models in an evaluation of cognitive workload in a flight simulator
S12	[41]	Psychophysiological Sensing and State Classification for Attention Management in Commercial Aviation
S13	[42]	Emotions Assessment on Simulated Flights
S14	[43]	Pilots' Fatigue Status Recognition Using Deep Contractive Autoencoder Network
S15	[44]	Empirical evidence for the relationship between cognitive workload and attentional reserve
S16	[45]	Effectiveness of Physiological and Psychological Features to Estimate Helicopter Pilots' Workload: A Bayesian Network Approach
S17	[46]	Quantifying cognitive workload in simulated flight using passive, dry EEG measurements
S18	[47]	Pilots' Brain Cognitive State Inference Based on Remaining Life HSMM
S19	[48]	The Complementary Role of Activity Context in the Mental Workload Evaluation of Helicopter Pilots: A Multi-tasking Learning Approach
S20	[49]	Measurement and modelling of the behavior of military pilots
S21	[50]	Cognitive state recognition using wavelet singular entropy and ARMA entropy with AFPA optimized GP classification
S22	[51]	Classification of Electrophysiological Signatures With Explainable Artificial Intelligence: The Case of Alarm Detection in Flight Simulator
S23	[25]	Pilot Behavior Recognition Based on Multi-Modality Fusion Technology Using Physiological Characteristics
S24	[52]	Measurement and identification of mental workload during simulated computer tasks with multimodal methods and machine learning
S25	[23]	EEG-based analysis for pilots' at-risk cognitive competency identification using RF-CNN algorithm
S26	[53]	Identification of Pilots' Fatigue Status Based on Electrocardiogram Signals
S27	[54]	Assessment of Ocular and Physiological Metrics to Discriminate Flight Phases in Real Light Aircraft
S28	[55]	Using machine learning methods and EEG to discriminate aircraft pilot cognitive workload during flight
S29	[56]	EEG-Based Prediction of Cognitive Load in Intelligence Tests
S30	[57]	Detecting Fatigue Status of Pilots Based on Deep Learning Network Using EEG Signals
S31	[28]	Recognition of Pilot Mental workload in the Simulation Operation of Carrier-based Aircraft Using the Portable EEG
S32	[58]	Towards safe and collaborative aerodrome operations: Assessing shared situational awareness for adverse weather detection with EEG-enabled Bayesian neural networks
S33	[59]	Assessing cognitive mental workload via EEG signals and an ensemble deep learning classifier based on denoising autoencoders
S34	[60]	Cognitive Load Identification of Pilots Based on Physiological-Psychological Characteristics in Complex Environments
S35	[61]	Predicting hypoxic hypoxia using machine learning and wearable sensors
S36	[62]	Classification of flight phases based on pilots' visual scanning strategies
S37	[63]	Subject-specific mental workload classification using EEG and stochastic configuration network
S38	[26]	Machine learning-based approach for identifying mental workload of pilots
S39	[64]	Securing air transportation safety through identifying pilot's risky VFR flying behaviors: An EEG-based neurophysiological modelling using machine learning algorithms
S40	[65]	Pilot Selection in the Era of Virtual Reality: Algorithms for Accurate and Interpretable Machine Learning Models
S41	[66]	Recognition of the Mental Workloads of Pilots in the Cockpit Using EEG Signals
S42	[67]	Use of a DNN in Recording and Analysis of Operator Attention in Advanced HMI Systems
S43	[68]	Assessing Control Devices for the Supervisory Control of Autonomous Wingmen
S44	[69]	Real-time evaluation method of flight mission load based on sensitivity analysis of physiological factors
S45	[70]	Self-Affinity of an Aircraft Pilot's Gaze Direction as a Marker of Visual Tunneling
S46	[71]	Predicting Student Flight Performance with Multimodal Features
S47	[72]	Fatigue Detection and Analysis Based on Multi-channel Data Fusion
S48	[73]	A comparison of ECG and EEG metrics for in-flight monitoring of helicopter pilot workload
S49	[74]	A machine learning approach to reduce mental fatigue risk of pilots based on HRV data
S50	[75]	Autonomous System for EEG-Based Multiple Abnormal Mental States Classification Using Hybrid Deep Neural Networks Under Flight Environment
S51	[76]	Classification of Drowsiness Levels Based on a Deep Spatio-Temporal Convolutional Bidirectional LSTM Network Using Electroencephalography Signals
S52	[24]	Classification of Distraction Levels Using Hybrid Deep Neural Networks From EEG Signals
S53	[77]	Collaborative Communications Between A Human And A Resilient Safety Support System
S54	[78]	Continuous EEG Decoding of Pilots' Mental States Using Multiple Feature Block-Based Convolutional Neural Network
S55	[79]	Deep Learning-based Eye Gaze Estimation for Military Aviation
S56	[80]	Design of Wearable Eye Tracker with Automatic Cockpit Areas of Interest Recognition
S57	[81]	Detection and classification of pilots cognitive state using EEG
S58	[82]	EEG Based Eye Movements Multi-Classification Using Convolutional Neural Network
S59	[83]	EEG-Based Mental Workload Estimation
S60	[84]	In silico vs. Over the Clouds: On-the-Fly Mental State Estimation of Aircraft Pilots, Using a Functional Near Infrared Spectroscopy Based Passive-BCI

TABLE 1. *(Continued.)* **List of the qualified studies.**

S61	[85]	Long-Term BCI Training of a Tetraplegic User: Adaptive Riemannian Classifiers and User Training
S62	[86]	A New Perspective for the Training Assessment: Machine Learning-Based Neurometric for Augmented User's Evaluation
S63	[20]	Detection of mental fatigue state using heart rate variability and eye metrics during simulated flight
S64	[87]	In-Flight Automatic Detection of Vigilance States Using a Single EEG Channel
S65	[88]	Mental Workload Artificial Intelligence Assessment of Pilots' EEG Based on Multi-Dimensional Data Fusion and LSTM with Attention Mechanism Model
S66	[89]	A New Approach to Aircraft Handling Qualities Prediction
S67	[90]	Visual scanning strategies in the cockpit are modulated by pilots' expertise: A flight simulator study
S68	[91]	Recognition of fatigue status of pilots based on deep sparse auto-encoding network
S69	[92]	Reducing Commercial Aviation Fatalities Using Support Vector Machines
S70	[93]	Modeling and analysis of fatigue detection with multi-channel data fusion
S71	[94]	Neural correlates of user learning during long-term BCI training for the Cybathlon competition
S72	[27]	Self-Paced Dynamic Infinite Mixture Model for Fatigue Evaluation of Pilots' Brains
S73	[95]	Generalized Deep Learning EEG Models for Cross-Participant and Cross-Task Detection of the Vigilance Decrement in Sustained Attention Tasks
S74	[96]	Towards Safer Flights: A Multi-modality Fusion Technology-based Cognitive Load Recognition Framework
S75	[97]	A Method for Classification and Evaluation of Pilot's Mental States Based on CNN
S76	[98]	Prediction of Pilot's Reaction Time Based on EEG Signals
S77	[99]	Illuminating the Neural Landscape of Pilot Mental States: A Convolutional Neural Network Approach with Shapley Additive Explanations Interpretability
S78	[100]	Multimodal Approach for Pilot Mental State Detection Based on EEG
S79	[101]	A Comprehensive Analysis of Machine Learning and Deep Learning Models for Identifying Pilots' Mental States from Imbalanced Physiological Data
S80	[102]	Miscellaneous EEG Preprocessing and Machine Learning for Pilots' Mental States Classification: Implications

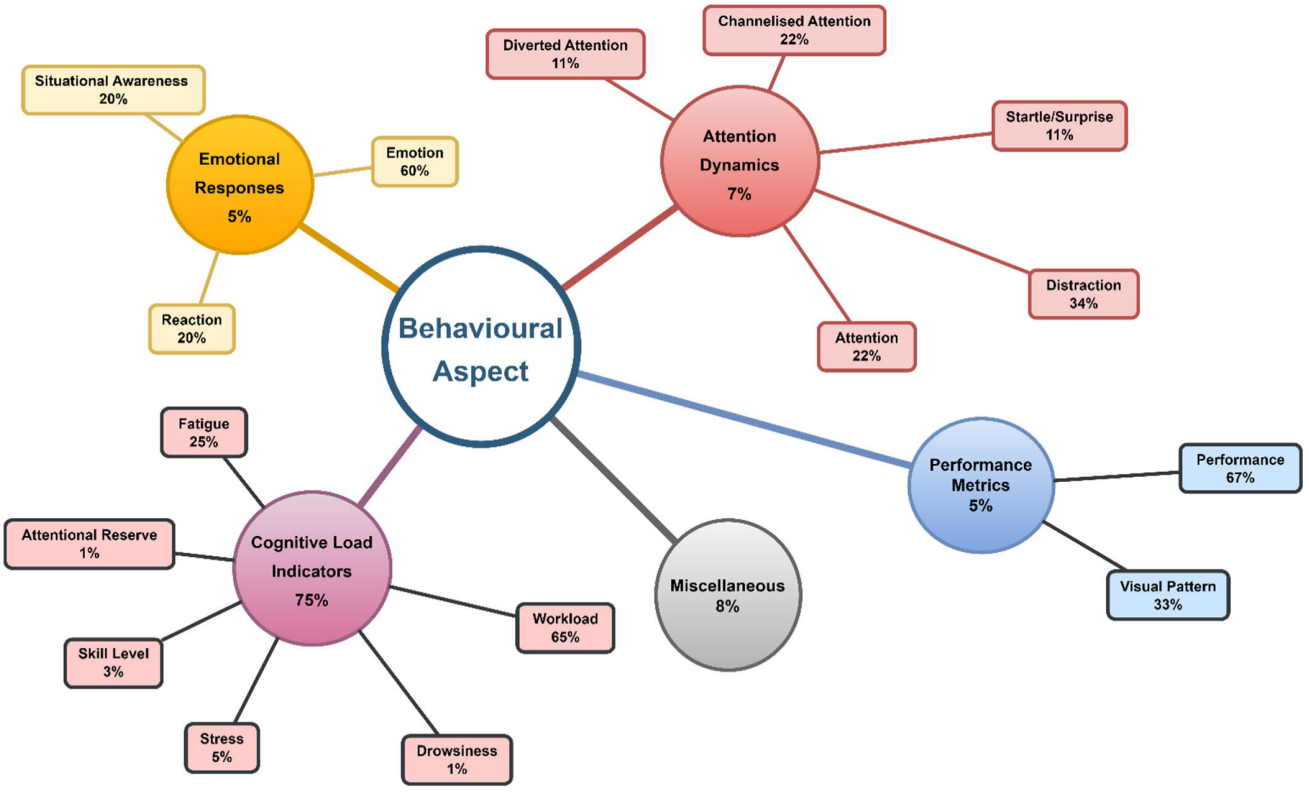

FIGURE 5. Behavior aspects categories.

and 'Startle/Surprise,' each barely surpassing a 10% share. The 'Miscellaneous' category, which accounts for 8% of the studies, is primarily composed of works where the behavioral aspect was neither the central focus nor explicitly articulated.

C. METHODOLOGICAL DESIGN: PSYCHOPHYSIOLOGICAL MEASURES, DATA PREPROCESSING, AND FEATURE EXTRACTION

The following subsection focuses on delineating the methodologies adopted in existing studies, with particular attention

to psychophysiological data employed, the artifacts identified, the methods used for data preprocessing, as well as the features extracted and their corresponding extraction techniques.

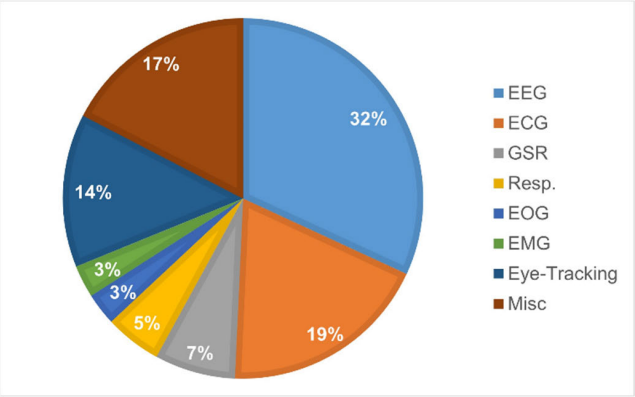

FIGURE 6. Comprehensive distribution of psychophysiological and other data types in existing literature on pilot behavior.

The distribution of psychophysiological data types used in research studies exhibits a notable range of diversity. As delineated in Fig. 6, EEG data are most commonly employed, accounting for 32% of the studies. This is followed by ECG data, which make up 19% of the studies. Interestingly, a 'Miscellaneous' category, comprising flight data and subjective measures such as NASA Task Load Index (NASA TLX), holds a non-trivial portion of 17%. Eye-Tracking and GSR follow suit, constituting 14% and 7%, respectively. On the lower end, Respiration (Resp.), Electrooculogram (EOG), and Electromyogram (EMG) data appear less frequently, each making up less than 5% of the studies.

Turning to Table 2, a detailed inspection reveals a rich array of artifacts and their corresponding preprocessing methods, sorted by psychophysiological data type. EEG data, for instance, are predominantly subjected to preprocessing methods such as ICA and bandpass filtering. Notably, some studies collected EOG, ECG, and EMG data simultaneously with EEG data and used them to identify heartbeats, muscle, and eye-related artifacts in the EEG data using ICA. For users of MATLAB, Artifact Subspace Reconstruction (ASR) is frequently employed. These techniques mitigate challenges posed by ocular and muscular artifacts common to EEG data collection. Interestingly, some studies did not employ any preprocessing techniques and proceeded directly to feature extraction. A range of other preprocessing techniques, including normalization, standardization, resampling, and detrending, were also employed. Some studies opted for manual inspection of the data to remove corrupted segments. ECG and GSR data, although less varied in preprocessing methods, also have unique sets of challenges and corresponding techniques. ECG data commonly undergo QRS detection to accurately identify heartbeats, while GSR data frequently are subjected to low-pass filtering to remove high-frequency noise.

TABLE 2. Summary of artifacts and corresponding preprocessing methods.

Data type	Artifact type	Preprocessing methods	Papers
EEG	• Eye movements • Muscle artifacts • Cardiac artifacts • Powerline interference	• Filtering • ICA • PCA • WT • ASR • Visual inspection and rejection	S2, S3, S4, S6, S8, S14, S25, S28, S29, S31, S32, S33, S39, S50, S51, S52, S54, S62
ECG	• Electrode motion • Powerline interference	• Filtering • WT	S1, S2, S6, S8, S13, S16, S19, S23, S24, S26
GSR	• Motion artifacts • Electrode artifacts	• Filtering	S2, S6, S8, S13, S16, S23, S24
Resp.	• Baseline drift • Motion artifacts	• Filtering	S2, S8, S16, S19, S23, S24
Eye-Tracking	• Blink artifacts • Saccadic artifacts	• Filtering • Interpolation	S3, S19, S27, S36, S40, S42, S45, S46, S47
Misc.	• Missing data • Outliers	• Data Smoothing • Interpolation	S1, S8, S13, S15, S23, S24

Wavelet Transform (WT), Principal Component Analysis (PCA), Miscellaneous (Misc.)

Complementing this, Table 3 offers a more nuanced examination of the features extracted from these psychophysiological data types. Within the domain of EEG, features such as PSD and WC are frequently extracted, often employing Fourier and WT. Some studies also extracted statistical features like mean, median, skewness, and kurtosis, often using time-domain methods. In addition, a cohort of studies explored the extraction of non-linear, spatial, and higher-level features like entropy, coherence, and phase-locking value. Several methodologies for feature extraction were noted, including Welch's method, Morlet wavelet, and Common Spatial Patterns. Furthermore, statistical tests

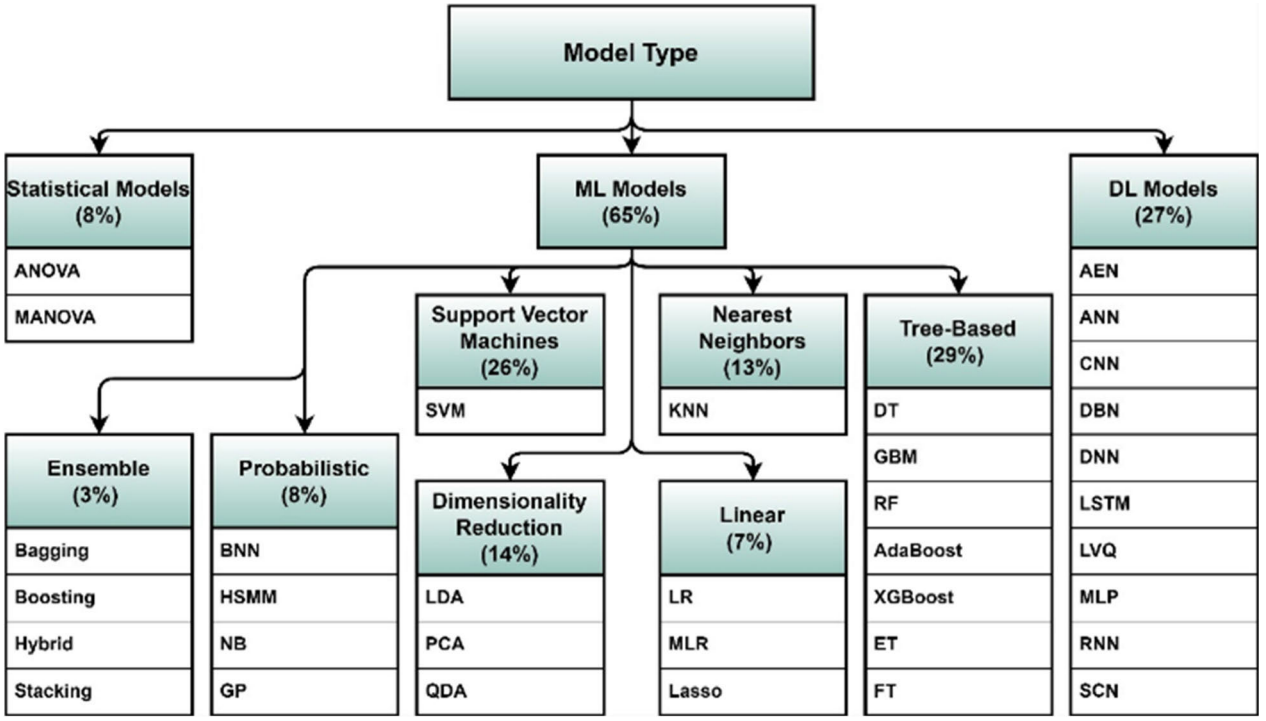

FIGURE 7. The model's types employed for identifying pilot's behavior. Abbreviation: Autoencoders (AEN), Artificial Neural networks (ANN), Deep Neural networks (DNN), Deep Belief Network (DBN), Multi-Linear Regression (MLR), Multilayer Perceptron (MLP), Naïve Bayes (NB), Random Forest (RF), Gradient Boosting Machines (GBM), Decision Trees (DT), Adaptive Boosting (AdaBoost), Support Vector Machines (SVM), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Least Absolute Shrinkage and Selection Operator (Lasso), Logistic Regression (LR), Long Short-Term Memory (LSTM), K-Nearest Neighbors (KNN), Hidden Semi Markov Models (HSMM), Quadratic Discriminant Analysis (QDA), Linear Discriminant Analysis (LDA), Naïve Bayes (NB), Gaussian Process (GP), Extra Tree (ET), Fine Tree (FT), Learning Vector Quantization (LVQ), Extreme Gradient Boosting (XGBoost), Bayesian neural networks (BNN), Stochastic Configuration Network (SCN).

and information-theoretic measures such as PCA, Analysis of Variance (ANOVA), Multivariate Analysis of Variance (MANOVA), Friedman tests, and Mutual Information Coefficient (MIC) were not uncommon. ML and DL methods also appeared as tools not only for classification but for feature extraction and selection as well.

In sum, the methodological paradigms underpinning the existing literature are diverse, intricate, and tailored to the unique challenges and opportunities presented by each type of psychophysiological data. These empirically-grounded observations provide a foundational base for subsequent interpretive and evaluative discussions.

D. TAXONOMY OF MODELS TYPES AND PERFORMANCE METRICS

The ensuing analysis is dedicated to providing a comprehensive breakdown of the types of predictive models currently deployed in the literature for the nuanced understanding of pilot behavior. Fig. 7 shows a more nuanced analysis of the model's types employed to identify the pilot's behavior. A compelling trend that demands attention is the preeminent use of ML models, which constitute a significant 65% of the total models utilized. This prevalence likely reflects the ML models' capability for handling complex, high-dimensional data. DL models are also noteworthy, albeit to a lesser extent, representing 27% of the models used. Statistical models account for the remaining 8%, indicating a less frequent but nonetheless important role in the research landscape.

Delving into the category of ML models, the analysis reveals a rich and varied methodological landscape. Leading this category are tree-based models, which account for 29% of ML models. Such models are frequently favored for their interpretability and robustness to noisy data. Following closely is SVM, which make up 26% of ML models, often chosen for their ability to handle high-dimensional spaces effectively. Dimensionality reduction models, which are crucial for simplifying complex datasets, comprise 14%. KNN algorithm is also significant, accounting for 13%, and are often employed for their simplicity and effectiveness in classification tasks. Probabilistic models, which offer nuanced probabilistic interpretations, account for 8%, while linear models, known for their ease of interpretation, make up 7%. Ensemble methods, which combine predictions from multiple models to improve performance, hold a smaller share of 3%. For further granularity, the tree-based models include a variety of algorithms like DT, XGBoost, and RF among others. Linear models predominantly feature LR and Lasso Regression, while Dimensionality Reduction models include techniques like LDA and PCA. Probabilistic models encompass BNN and GP, and ensemble methods feature techniques like Bagging and Boosting.

TABLE 3. Summary of features extracted and extraction methods.

Data type	Extracted features	Extraction and selection methods	Papers
EEG	• Frequency bands • Statistical features • ERPs • PSD	• FFT • WT • PCA • ICA • HHT • LSFT • WPD	S2, S3, S4, S5, S6, S8, S14, S25, S28, S29, S31, S32, S33, S39, S50, S51, S52, S54, S57, S58, S59, S62
ECG	• HR • HRV • LF • HF • RR intervals	• FFT • ANOVA • MANOVA • WT • Friedman test • Pearson correlation • Wilcoxon tests	S1, S2, S6, S8, S13, S16, S19, S23, S24, S26
GSR	• SCL • SCR	• ANOVA • MANOVA	S2, S6, S8, S13, S16, S23, S24
Resp.	• SDAbd • SDThor • DRFAbd • DRFThor	• ANOVA • MANOVA	S2, S8, S16, S19, S23, S24
Eye-Tracking	• Fixation duration • Saccade length • Pupil Dilation • AOIs	• MIC	S3, S19, S27, S36, S40, S42, S45, S46, S47
Misc.	• Altitude, Speed, Heading • Roll, Pitch, Yaw angles • Control inputs	• ANOVA • MANOVA	S1, S8, S13, S15, S23, S24

Area Of Interests (AOIs), Delta Ribcage-to-Abdomen (DRFAbd), Delta Ribcage-to-Thoracic (DRFThor), Event-related potential (ERP), Frequency Bands (Delta, Theta, Alpha, Beta, Gamma), Heart Rate (HR), Heart Rate Variability (HRV), Fast Fourier Transform (FFT), Lomb-Scargle Frequency Transform (LSFT), Hilbert-Huang Transform (HHT), High Frequency (HF), Low Frequency (LF), Standard Deviation Abdominal (SDAbd), Skin Conductance Level (SCL), Skin Conductance Response (SCR), Standard Deviation Thoracic (SDThor), Wavelet Packet Decomposition (WPD)

In the sphere of DL models, the existing studies unveil an extensive array of architectures, reflecting the burgeoning interest in leveraging complex neural network structures for identifying pilot behavior. Contributing to 27% of the total models deployed, DL techniques demonstrate their growing influence in this domain. The architectures span a variety of models, from traditional ANN to more complex and specialized types like CNN and LSTM. Each of these architectures brings unique advantages to the analysis of psychophysiological data. For instance, CNNs are often employed for their ability to automatically and adaptively learn spatial hierarchies of features, making them ideal for image and sequence data. LSTMs, on the other hand, excel in handling time-series data, capturing long-term dependencies which are often crucial in understanding behavioral states. The presence of DBN and RNN further enriches the landscape, signifying ongoing experimentation in the field to identify the most effective DL approaches for specific tasks. Even more specialized architectures like SCN and LVQ find mentions, indicating that the field is continually evolving with an expanding repertoire of advanced techniques. Statistical models, while less frequently employed, consist primarily of traditional techniques like ANOVA and MANOVA. These models are often used for hypothesis testing and the exploration of relationships between variables, providing a contrast to the predictive focus of ML and DL models.

TABLE 4. The metrics used to evaluate the models' performance.

Performance Metric	Paper
Accuracy	S1, S2, S4, S5, S6, S8, S9, S10, S12, S14, S16, S17, S21, S22, S23, S24, S25, S26, S27, S28, S30, S31, S32, S33, S34, S36, S37, S38, S39, S40, S42, S44, S45, S46, S47, S49, S50, S51, S52, S54, S56, S57, S58, S59, S60, S63, S65, S68, S75
Recall	S4, S10, S23, S24, S25, S26, S28, S31, S32, S33, S34, S35, S39, S40, S41, S44, S45, S50, S51, S59, S60, S63
Specificity	S4, S25, S33, S34, S35, S41, S51, S59
Precision	S10, S23, S24, S26, S28, S31, S32, S33, S39, S40, S44, S45, S50, S59, S60, S63
F1-score	S23, S26, S28, S31, S32, S39, S40, S44, S50, S63
AUC	S3, S10, S25, S28, S32, S35, S40, S44
MAE	S6, S11
MSE	S13, S23, S48
RMSE	S1, S6, S19, S42
Pearson Correlation	S29

Area under the Curve (AUC), Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE)

On the metric front, as shown in Table 4, accuracy is the most reported performance metric, featured in 65% of the studies, likely due to its simplicity and straightforward interpretation. Recall, which focuses on the model's ability

to identify all relevant instances, is reported in 29% of the studies, indicating its importance in applications where missing a positive instance is particularly costly. Precision appears in 21% of the studies, often employed alongside recall to provide a more complete picture of model performance. Specificity and F1-score, metrics that consider both false positives and negatives, are reported in 11% and 13% of the studies respectively. The AUC, RMSE, MSE, MAE, and Pearson's correlation metrics appear less frequently, suggesting their application in more specific or specialized contexts. Notably, some studies adopt a multi-metric approach, indicating a comprehensive methodology for performance evaluation.

In summary, the existing literature exhibits a varied and intricate array of predictive models and performance metrics, reflecting the methodological diversity inherent in the field. These findings serve as a robust foundation for subsequent interpretative discussions and scholarly evaluations, offering a comprehensive view of the methodological paradigms shaping current research.

E. COMPARATIVE PERFORMANCE OF MACHINE LEARNING AND DEEP LEARNING MODELS IN PREDICTING PILOT BEHAVIOR

The current subsection seeks to offer an exhaustive analytical examination of the average performance of diverse categories of ML and DL models. As a robust methodological approach, the performance accuracy for each category were extracted from the selected studies, meticulously averaged, and subsequently visualized in a bar chart, denoted as Fig. 8.

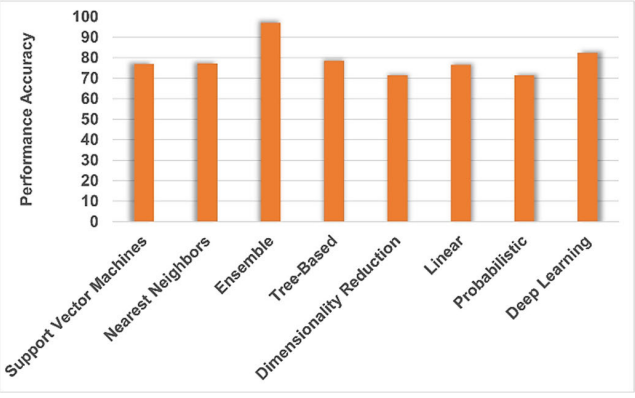

FIGURE 8. The performance accuracy of the models utilized in the literature.

The SVM and KNN models both share an identical average performance accuracy of 77%. While these numbers are certainly respectable, they do not represent the pinnacle of performance among the categories. Remarkably, Ensemble models eclipse other methodologies with an exceptional average performance rate of 97%. This exceptional performance could be attributed to the inherent capability of Ensemble models to combine multiple weak learners, thereby enhancing generalizability and robustness against overfitting.

Closely following Ensemble models, tree-based models exhibit an average performance rate of 78%. As illustrated in Fig. 9, XGBoost and GBM show a higher lower quartile at 64% and 86%, respectively, as well as a tighter interquartile range within this category, suggesting greater robustness in performance. Notably, ET appear to be exceptionally consistent, with all quintiles at 97%.

DL models also command attention with their average performance accuracy of 82%. For DL models, ANN and CNN display robust performances with medians at 80% and 83%, respectively. LSTM models show a lower quartile at 62% but reach as high as 87%, indicating potential for high performance but also room for improvement.

In contrast, Dimensionality Reduction and Probabilistic models both manifest a relatively lower average performance rate of 71%. Within Dimensionality Reduction models, LDA and QDA show a broad range in their performance. LDA has a lower quartile at 65% and an upper quartile at 81%, while QDA exhibits a wider distribution with a lower quartile at 48% and an upper quartile at 89%. Similarly, Linear models register an average performance rate of 77%, which is in line with SVM and KNN. Within Probabilistic models, BNN show remarkably consistent performance, with all quintiles at 67%. In contrast, GP and NB manifest wide performance ranges, from 43% to 96% and 37% to 91%, respectively.

IV. DISCUSSION

The Discussion section serves as a critical forum for interpreting the empirical findings presented in the results section. In line with the research questions posited, this section aims to offer an in-depth analysis of the current state of research on the application of ML models and psychophysiological data in understanding pilot behavior. It further contextualizes these findings within the broader academic discourse and identifies both methodological limitations and avenues for future research.

A. EVALUATION OF RESEARCH FOCUS ON PILOT'S BEHAVIOURAL AND COGNITIVE STATES (RQ1)

The analysis encapsulated in the results' subsection B offers a nuanced perspective on the existing body of research surrounding pilot behavior. While 'Cognitive Load Indicators' occupy a dominant position in the academic discourse, it is essential to interrogate the reasons behind such focused attention. One could speculate that the quantifiable nature of indicators like 'Workload' and 'Fatigue' makes them attractive candidates for empirical studies, possibly offering more straightforward avenues for data collection and analysis. However, this concentration exposes a conspicuous void in other pivotal areas. The paucity of research on performance-limiting states such as 'Channelized Attention,' 'Diverted Attention,' and 'Startle/Surprise' is particularly concerning. Given the critical nature of aviation operations and the potential ramifications of performance-limiting states on both safety and efficiency, this research gap represents a glaring omission.

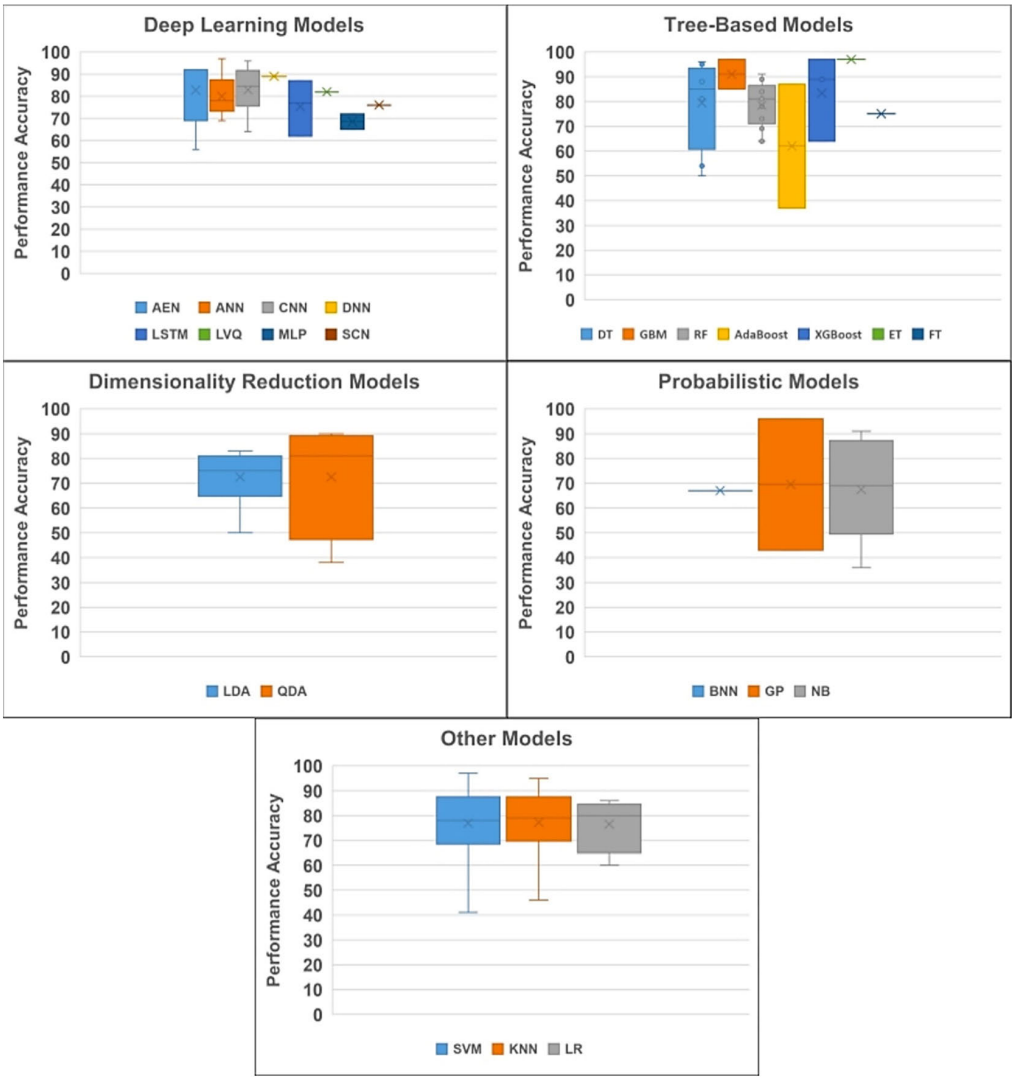

FIGURE 9. A box plot for each model type category.

In considering the methodological underpinnings of the existing literature, we encounter two predominant approaches: multi-level and binary classifications. Multi-level classifications, which are often employed to dissect complex behavioral aspects like 'Workload,' offer a more textured understanding but suffer from challenges related to comparability and standardization across studies. The absence of a universally accepted metric for defining 'low,' 'medium,' or 'high' levels of a behavioral aspect could potentially muddle the collective insights drawn from various studies. In contrast, binary classifications, commonly used for attributes like 'Fatigue,' offer clarity and are more easily interpretable. However, this reductionist approach might not capture the continuum of behavioral states pilots may experience, potentially leading to an incomplete or skewed understanding.

These methodological choices have far-reaching implications. For instance, the prevalent use of binary classifications might be well-suited for real-time monitoring systems in cockpits, where quick decisions are paramount. However, such systems, if based solely on existing binary-classification research, might lack the sensitivity to detect nuanced changes in a pilot's behavioral state, thereby reducing their overall efficacy. Thus, a balanced methodological approach seems warranted for future research. Adopting a hybrid model that incorporates both multi-level and binary classifications could offer a more holistic view, capturing both the nuanced complexities and the actionable insights needed in practical applications.

B. INTERPRETING METHODOLOGICAL PARADIGMS IN PILOT BEHAVIOR RESEARCH (RQ2)

The present analysis of existing studies offers a comprehensive perspective on the intricate methodologies adopted in the domain of pilot behavior research, revealing both the depth and the complexity of the current landscape. This diversity not only reflects the multidisciplinary nature of the field but also raises questions about methodological coherence

and standardization, providing fertile ground for academic scrutiny.

At the forefront of psychophysiological measures is EEG data, which constitutes 32% of the studies reviewed. This prevalence attests to EEG's high temporal resolution and its capability to capture complex neural activities, factors that have rendered it a popular choice among researchers. However, the data landscape is far from being monolithic. ECG data, which accounts for 19% of the studies, is also pivotal, often serving as an indicator of physiological stress and cognitive workload. The role of Eye-Tracking data is similarly significant, often employed to assess attentional states and situational awareness. These alternative data types underscore the multi-faceted nature of pilot behavior, which cannot be comprehensively understood through neural activities alone. The 'Misc.' category, comprising 17% of the studies and including flight data and subjective measures like the NASA TLX and Karolinska Sleepiness Scale, adds another layer of complexity. This category suggests an emerging trend towards the incorporation of multi-modal and subjective data, potentially offering a more rounded understanding of pilot behavior, a point that merits further investigation in future studies.

Diving into data preprocessing, the study identifies a wide array of techniques, each with its unique strengths and limitations. For EEG data, the prevalent use of ICA and bandpass filtering signifies a focus on mitigating ocular and muscular artifacts. However, the rise of ASR technique among MATLAB users signals the adoption of specialized methods that are tailored to specific research needs. Interestingly, some studies bypass preprocessing altogether, a choice that may have implications for data quality and interpretability. This diversity in preprocessing methods raises critical questions about the standardization and comparability of research outcomes. In the feature extraction stage, the landscape is equally diverse. While Fourier and WT are commonly employed for frequency-domain feature extraction from EEG data, the study also identifies a growing interest in statistical features and higher-level non-linear and spatial features. This methodological diversity is further enriched by the use of ML and DL techniques, not just for classification but also for feature extraction and selection.

The observed methodological paradigms thus present both opportunities and challenges. On the one hand, the diversity of methods enriches our understanding of pilot behavior from multiple psychophysiological perspectives. On the other hand, the lack of methodological standardization hampers cross-study comparisons and meta-analyses, an issue that warrants attention in future research. The existing literature on pilot behavior showcases a complex tapestry of methodological approaches, each designed to tackle the unique challenges posed by different types of psychophysiological data. This diversity offers a rich yet complex view of current research practices, providing a foundational base for subsequent academic discussions and critical evaluations.

C. INTERPRETATIVE DISCUSSION FOR MODEL TYPES AND EVALUATION METRICS (RQ3)

This discussion aims to delve into the use of detection models and performance metrics observed in existing literature, particularly in the context of utilizing psychophysiological data to identify pilot behavior. One of the most salient aspects is the predominant deployment of ML models, which constitute 65% of the total models utilized. This considerable emphasis on ML models raises pertinent questions about their comparative efficacy, especially in contexts where complex, high-dimensional data are involved.

In the domain of DL models, the diversity of architectures is particularly noteworthy. Contributing to 27% of the total models used, DL models signify their burgeoning influence in this area. Researchers have proposed various architectures, some combining CNN with Long LSTM networks for layered complexity. Other innovative proposals include deep contractive autoencoder networks with softmax classifiers, deep sparse autoencoder networks, and feature mapping layers in stacked denoising autoencoders. This suggests that the field is in a state of methodological flux, continuously exploring and adapting to find the most effective DL models for specific tasks. Traditional statistical models, although foundational, appear less frequently, making up 8% of the total models. Their limited use possibly suggests a methodological shift towards more data-driven models.

The metrics employed for performance evaluation also deserve critical examination. The prominence of accuracy, reported in 66% of the studies, could indicate a focus on overall classification effectiveness. However, the metric may not suffice in cases where the dataset is heavily imbalanced, underlining the need for more nuanced evaluation metrics like recall or precision. The adoption of multiple metrics in some studies indicates a multi-faceted approach to performance assessment but also points to a lack of standardization that could impede cross-study comparisons.

In conclusion, the existing literature exhibits a rich array of methodologies, from traditional statistical and ML models to advanced DL models, each with their unique merits and limitations. The variety of performance metrics used, while indicative of methodological diversity, also suggests the need for further standardization and comparative evaluation.

D. INTERPRETATIVE ANALYSIS BASED ON MODEL PERFORMANCE (RQ4)

The comprehensive results presented in the results' subsection E offer rich insights into the relative performance of various ML and DL models in the domain of pilot behavior prediction. The standout performance of Ensemble models, averaging at an exceptional rate of 97%, is particularly noteworthy. This could be attributed to the capacity of Ensemble models to synthesize insights from multiple weak learners, thereby enhancing their generalizability and robustness against overfitting. However, this high performance also raises questions about the diversity of base

learners employed in these ensemble models and how that contributes to their effectiveness. Tree-based models, with an average performance rate of 78%, offer another interesting point for discussion. While they perform well on average, the variance in performance across different types of tree-based models, such as RF and GBM, suggests that the choice of specific tree algorithms and their hyperparameters could be a crucial factor in achieving optimal performance.

The performance of DL models, averaging at 82%, is notable for its potential to capture intricate patterns in high-dimensional data. Yet, the distribution of performance across various DL architectures such as CNNs, LSTMs, and ANNs indicates that no single architecture dominates in terms of efficacy. This divergence could be indicative of the specialized nature of these architectures, optimized for specific kinds of data or tasks within the broader realm of pilot behavior prediction.

Dimensionality Reduction and Probabilistic models, with their lower average performance rates of 71%, warrant a discussion on their applicability and limitations. Given the complex, high-dimensional nature of psychophysiological data, these models may not capture the full scope of relevant features or patterns, thus limiting their performance. Future work might explore hybrid models that combine these methods with higher-performing models to improve accuracy. Moreover, the fact that some models show a wide distribution in their performance statistics, such as GP and NB, suggests a sensitivity to the specific conditions or configurations under which they are employed. This could be an important area for future investigation, particularly in identifying what those conditions or configurations are.

In summary, the detailed results on model performance and their distribution provide a multifaceted view of the current methodological landscape in predicting pilot behavior. They elucidate not just the strengths and weaknesses of various model categories but also point to numerous questions and directions for future research. This could include the exploration of hybrid models, methodological innovations to improve the performance of underperforming categories, and more nuanced applications tailored to the specific needs and challenges of psychophysiological data in pilot behavior analysis.

E. METHODOLOGICAL LIMITATIONS AND FUTURE RESEARCH DIRECTIONS (RQ5)

The assessment of the current literature reveals significant gaps and areas for improvement, necessitating a focused discussion on methodological limitations and future research directions. One pressing concern is the largely unexamined impact of preprocessing techniques on ML models. Although numerous preprocessing methods are employed across studies, the extent to which these choices influence ML models outcomes remains largely unexplored. This represents a critical avenue for future research, as a better understanding of this interplay could lead to more robust and generalizable models.

Another limitation is the reliance on traditional preprocessing techniques. The complexity of psychophysiological data, fraught with various artifacts, calls for the exploration of advanced preprocessing methods. Incorporating such methods could potentially lead to more accurate and reliable models for understanding pilot behavior, and thus should be a focus of future research efforts.

Additionally, the impact of employing data imbalance techniques on the performance of ML models has not been fully explored and evaluated. Given the frequent occurrence of imbalanced datasets in this domain, this lack of focus raises concerns about the generalizability and reliability of the reported results. Further, the disproportionate emphasis on accuracy as the principal metric for evaluating model performance becomes problematic, especially in cases involving imbalanced datasets. A focus on accuracy alone may not accurately reflect the model's ability to predict minority classes. Therefore, a multi-metric evaluation framework, incorporating additional metrics like recall, precision, and the F1-score, is crucial for a more balanced and comprehensive assessment of model performance.

In the domain of DL, the utilization of 1D-CNN appears to be underexplored in the context of psychophysiological data analysis for pilot behavior. The architecture of 1D-CNNs is well-suited for handling time-series data, offering the potential for enhanced feature extraction and, ultimately, more accurate predictions. Few studies in the selected corpus have addressed the critical issue of model interpretability or explainability, a paramount concern for real-world applications where understanding model decisions can have significant implications. This glaring omission in the current literature underscores the need for greater methodological rigor in future studies.

The literature's focus on data from specific environmental settings constrains the generalizability of the findings. Future studies could benefit from collecting and analyzing data from different environmental contexts, thereby enhancing the ecological validity of the research and providing a more comprehensive understanding of pilot behavior under varying conditions. Furthermore, the feature extraction methods employed in existing studies demonstrate a limited focus on traditional statistical and frequency-domain features. The exploration of spatial features, such as tangent space, remains largely untapped. Given the promise of such features in providing nuanced insights into cognitive states, their exploration could be a significant contribution to the field.

Lastly, the multidisciplinary nature of the field suggests an overarching need for cross-disciplinary collaboration to develop a more unified methodological framework. Such a framework could accommodate the complexities inherent in each type of psychophysiological data and the unique challenges they present. This standardization could, in turn, facilitate meta-analyses and cross-study comparisons, enriching our understanding of pilot behavior from multiple vantage points.

V. CONCLUSION

This systematic literature review endeavors to offer a nuanced and comprehensive understanding of the current state of research that applies ML models for the interpretation of psychophysiological data, specifically focusing on the behavior of pilots. A multifaceted array of findings have emerged from this review, which span the gamut from the types of psychophysiological data employed to the specific ML methodologies and their corresponding performance metrics. Firstly, this review uncovers a pronounced heterogeneity in the types of psychophysiological data employed across studies, with EEG data standing out as the most commonly used. This prominence of EEG data could be indicative of the broader acceptance of its reliability and efficacy in capturing cognitive states, yet it also raises questions about the underutilization of other types of data like ECG, GSR, and eye-tracking metrics.

Significantly, the review has identified a substantial gap in the behavioral aspects studied, most notably the underrepresentation of emotional responses and attention dynamics in the existing literature. These areas, although critical to understanding human performance-limiting states, have been less explored compared to workload and fatigue. Emotional states and attention levels are not only crucial for aviation safety but also enrich the understanding of pilot behavior in a more holistic manner. The current methodological approaches often categorize these aspects into broader categories, thereby potentially missing nuanced interrelations between different behavioral and cognitive states. Therefore, a more balanced academic inquiry into these areas is warranted for a more comprehensive understanding of pilot behavior.

When it comes to preprocessing methodologies, a diverse range exists; however, a notable gap lies in the absence of rigorous empirical evaluation exploring how these preprocessing choices could impact the outcomes of ML models. Given the intricacy of psychophysiological data, which often contains various types of noise and artifacts, understanding this relationship is not just academically interesting but also practically vital. Additionally, a remarkable methodological limitation is the scant attention given to the critical issue of model interpretability and explainability. Given that ML models are increasingly being considered for real-world applications in aviation, the lack of focus on this aspect is a significant shortcoming that future research must address.

The review also spotlights several key avenues for future investigation. It suggests that examining the impact of advanced preprocessing techniques, and how they interact with different model types, could offer new pathways to enhance model performance. The exploration of data imbalance correction methods, the use of spatial features like tangent space, and the incorporation of innovative model architectures such as 1D-CNNs represent other promising directions.

In sum, while the existing literature provides an invaluable starting point for the scientific understanding of pilot behavior through the lens of ML and psychophysiological data, there is ample room for methodological refinement and exploration. Addressing the identified gaps and underresearched areas will not only elevate the scientific rigor but also contribute to more nuanced, comprehensive, and practically applicable insights into pilot behavior. By focusing on these aspects, future research can aim to substantially advance the field, enriching both its academic depth and its practical applicability in the broader context of aviation safety and efficiency.

REFERENCES

[1] D. A. Wiegmann, J. Goh, and D. O'Hare, "The role of situation assessment and flight experience in pilots' decisions to continue visual flight rules flight into adverse weather," *Human Factors, J. Human Factors Ergonom. Soc.*, vol. 44, no. 2, pp. 189–197, Jun. 2002, doi: 10.1518/0018720024497871.

[2] B. Kitchenham, "Guidelines for performing systematic literature reviews in software engineering," Dept. Comput. Sci., Keele Univ., Keele, U.K., Tech. Rep. EBSE-2007-001, Jan. 2007.

[3] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering—A systematic literature review," *Inf. Softw. Technol.*, vol. 51, no. 1, pp. 7–15, Jan. 2009, doi: 10.1016/j.infsof.2008.09.009.

[4] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *BMJ*, vol. 339, no. 21 1, p. B2535, Jul. 2009, doi: 10.1136/bmj.b2535.

[5] C. V. Oster, J. S. Strong, and C. K. Zorn, "Analyzing aviation safety: Problems, challenges, opportunities," *Res. Transp. Econ.*, vol. 43, no. 1, pp. 148–164, Jul. 2013, doi: 10.1016/j.retrec.2012.12.001.

[6] I. L. Henderson, "Aviation safety regulations for unmanned aircraft operations: Perspectives from users," *Transp. Policy*, vol. 125, pp. 192–206, Sep. 2022, doi: 10.1016/j.tranpol.2022.06.006.

[7] D. D. Boyd, "A review of general aviation safety (1984–2017)," *Aerosp. Med. Human Perform.*, vol. 88, no. 7, pp. 657–664, Jul. 2017, doi: 10.3357/AMHP.4862.2017.

[8] G. Li and S. P. Baker, "Crash risk in general aviation," *J. Amer. Med. Assoc.*, vol. 297, no. 14, p. 1596, Apr. 2007, doi: 10.1001/jama.297.14.1596.

[9] G. Li, S. P. Baker, J. G. Grabowski, and G. W. Rebok, "Factors associated with pilot error in aviation crashes," *Aviat. Space Environ. Med.*, vol. 72, no. 1, pp. 52–58, 2001. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0035157741&partnerID=40&md5=8e0aed563fd772fa492f126484ad5e1e

[10] Commercial Aviation Safety Team. *SE211: Airplane State Awareness—Training for Attention Management*. Accessed: Dec. 25, 2022. [Online]. Available: http://www.skybrary.aero/index.php/SE211: _Airplane_State_Awareness_-_Training_for_Attention_Management_(R-D)

[11] D. E. M. Aurino, "Human factors and aviation safety: What the industry has, what the industry needs," *Ergonomics*, vol. 43, no. 7, pp. 952–959, Jul. 2000, doi: 10.1080/001401300409134.

[12] R. Bor, "The psychological impact of aircraft disasters," in *Passenger Behaviour*, 1st ed. New York, NY, USA: Taylor & Francis, 2003, pp. 1–18. [Online]. Available: https://www.taylorfrancis.com/chapters/mono/10.4324/9781315247328-20/psychological-impact-aircraft-disasters-robert-bor

[13] J. Behrend and F. Dehais, "How role assignment impacts decision-making in high-risk environments: Evidence from eye-tracking in aviation," *Saf. Sci.*, vol. 127, Jul. 2020, Art. no. 104738, doi: 10.1016/j.ssci.2020.104738.

[14] N. A. Stanton, P. R. G. Chambers, and J. Piggott, "Situational awareness and safety," *Saf. Sci.*, vol. 39, no. 3, pp. 189–204, 2001, doi: 10.1016/S0925-7535(01)00010-8.

[15] A. C. Marinescu, S. Sharples, A. C. Ritchie, T. S. López, M. McDowell, and H. P. Morvan, "Physiological parameter response to variation of mental workload," *Human Factors, J. Human Factors Ergonom. Soc.*, vol. 60, no. 1, pp. 31–56, Feb. 2018, doi: 10.1177/0018720817733101.

[16] K. Mohanavelu, S. Poonguzhali, K. Adalarasu, D. Ravi, V. Chinnadurai, S. Vinutha, K. Ramachandran, and S. Jayaraman, "Dynamic cognitive workload assessment for fighter pilots in simulated fighter aircraft environment using EEG," *Biomed. Signal Process. Control*, vol. 61, Aug. 2020, Art. no. 102018, doi: 10.1016/j.bspc.2020.102018.

[17] D. A. L. M. D. Sant'Anna and A. V. G. D. Hilal, "The impact of human factors on pilots' safety behavior in offshore aviation companies: A Brazilian case," *Saf. Sci.*, vol. 140, Aug. 2021, Art. no. 105272, doi: 10.1016/j.ssci.2021.105272.

[18] N. B. Sarter, R. J. Mumaw, and C. D. Wickens, "Pilots' monitoring strategies and performance on automated flight decks: An empirical study combining behavioral and eye-tracking data," *Human Factors, J. Human Factors Ergonom. Soc.*, vol. 49, no. 3, pp. 347–357, Jun. 2007, doi: 10.1518/001872007X196685.

[19] N. A. Stanton, K. L. Plant, K. M. A. Revell, T. G. C. Griffin, S. Moffat, and M. Stanton, "Distributed cognition in aviation operations: A gate-to-gate study with implications for distributed crewing," *Ergonomics*, vol. 62, no. 2, pp. 138–155, Feb. 2019, doi: 10.1080/00140139.2018.1520917.

[20] H. Qin, X. Zhou, X. Ou, Y. Liu, and C. Xue, "Detection of mental fatigue state using heart rate variability and eye metrics during simulated flight," *Human Factors Ergonom. Manuf. Service Industries*, vol. 31, no. 6, pp. 637–651, Nov. 2021, doi: 10.1002/hfm.20927.

[21] C. A. Morgan, D. E. Aikins, G. Steffian, V. Coric, and S. Southwick, "Relation between cardiac vagal tone and performance in male military personnel exposed to high stress: Three prospective studies," *Psychophysiology*, vol. 44, no. 1, pp. 120–127, Jan. 2007, doi: 10.1111/j.1469-8986.2006.00475.x.

[22] F. Chen, J. Zhou, Y. Wang, K. Yu, S. Z. Arshad, A. Khawaji, and D. Conway, *Robust Multimodal Cognitive Load Measurement* (Human–Computer Interaction Series). Cham, Switzerland: Springer, 2016, doi: 10.1007/978-3-319-31700-7.

[23] S. Jiang, W. Chen, Z. Ren, and H. Zhu, "EEG-based analysis for pilots' at-risk cognitive competency identification using RF-CNN algorithm," *Frontiers Neurosci.*, vol. 17, Apr. 2023, Art. no. 1172103, doi: 10.3389/fnins.2023.1172103.

[24] D.-H. Lee, S.-J. Kim, and Y.-W. Choi, "Classification of distraction levels using hybrid deep neural networks from EEG signals," in *Proc. 11th Int. Winter Conf. Brain-Comput. Interface (BCI)*, Feb. 2023, pp. 1–4, doi: 10.1109/BCI57258.2023.10078681.

[25] Y. Li, K. Li, S. Wang, X. Chen, and D. Wen, "Pilot behavior recognition based on multi-modality fusion technology using physiological characteristics," *Biosensors*, vol. 12, no. 6, p. 404, Jun. 2022, doi: 10.3390/bios12060404.

[26] K. Mohanavelu, S. Poonguzhali, A. Janani, and S. Vinutha, "Machine learning-based approach for identifying mental workload of pilots," *Biomed. Signal Process. Control*, vol. 75, May 2022, Art. no. 103623, doi: 10.1016/j.bspc.2022.103623.

[27] E. Q. Wu, M. Zhou, D. Hu, L. Zhu, Z. Tang, X.-Y. Qiu, P.-Y. Deng, L.-M. Zhu, and H. Ren, "Self-paced dynamic infinite mixture model for fatigue evaluation of pilots' brains," *IEEE Trans. Cybern.*, vol. 52, no. 7, pp. 5623–5638, Jul. 2022, doi: 10.1109/TCYB.2020.3033005.

[28] W. Zhu, C. Zhang, C. Liu, J. Yuan, X. Li, Y. Wang, and C. Jiang, "Recognition of pilot mental workload in the simulation operation of carrier-based aircraft using the portable EEG," in *Proc. 3rd Int. Conf. Human Mach. Interact.*, May 2023, pp. 43–49, doi: 10.1145/3604383.3604391.

[29] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[30] S. K. R. Nittala, C. P. Elkin, J. M. Kiker, R. Meyer, J. Curro, A. K. Reiter, K. S. Xu, and V. K. Devabhaktuni, "Pilot skill level and workload prediction for sliding-scale autonomy," in *Proc. 17th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2018, pp. 1166–1173, doi: 10.1109/ICMLA.2018.00188.

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

[32] M. D. Ziegler, A. Kraft, M. Krein, L. C. Lo, B. Hatfield, W. Casebeer, and B. Russell, "Sensing and assessing cognitive workload across multiple tasks," in *Proc. 10th Int. Conf. Found. Augmented Cognition, Neuroergonomics Oper. Neurosci., 18th Int. Conf. Human-Computer Interact. (HCI)*, vol. 9743. Cham, Switzerland: Springer Verlag, 2016, pp. 440–450, doi: 10.1007/978-3-319-39955-3_41.

[33] F. Dehais, A. Duprès, S. Blum, N. Drougard, S. Scannella, R. Roy, and F. Lotte, "Monitoring pilot's mental workload using ERPs and spectral power with a six-dry-electrode EEG system in real flight conditions," *Sensors*, vol. 19, no. 6, p. 1324, Mar. 2019, doi: 10.3390/s19061324.

[34] B. Binias, D. Myszor, and K. A. Cyran, "A machine learning approach to the detection of pilot's reaction to unexpected events based on EEG signals," *Comput. Intell. Neurosci.*, vol. 2018, pp. 1–9, 2018, doi: 10.1155/2018/2703513.

[35] V. C. C. Roza and O. A. Postolache, "Multimodal approach for emotion recognition based on simulated flight experiments," *Sensors*, vol. 19, no. 24, p. 5516, Dec. 2019, doi: 10.3390/s19245516.

[36] L. C. Thomas, C. Gast, R. Grube, and K. Craig, "Fatigue detection in commercial flight operations: Results using physiological measures," *Procedia Manuf.*, vol. 3, pp. 2357–2364, Jan. 2015, doi: 10.1016/j.promfg.2015.07.383.

[37] A. R. Harrivel, C. L. Stephens, R. J. Milletich, C. M. Heinich, M. C. Last, N. J. Napoli, N. Abraham, L. J. Prinzel, M. A. Motter, and A. T. Pope, "Prediction of cognitive states during flight simulation using multimodal psychophysiological sensing," in *Proc. AIAA Inf. Systems-AIAA Infotech @ Aerosp.*, Jan. 2017, p. 1135, doi: 10.2514/6.2017-1135.

[38] M. K. Johnson, J. A. Blanco, R. J. Gentili, K. J. Jaquess, H. Oh, and B. D. Hatfield, "Probe-independent EEG assessment of mental workload in pilots," in *Proc. 7th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, Apr. 2015, pp. 581–584, doi: 10.1109/NER.2015.7146689.

[39] H. Oh, B. D. Hatfield, K. J. Jaquess, L.-C. Lo, Y. Y. Tan, M. C. Prevost, J. M. Mohler, H. Postlethwaite, J. C. Rietschel, M. W. Miller, J. A. Blanco, S. Chen, and R. J. Gentili, "A composite cognitive workload assessment system in pilots under various task demands using ensemble learning," in *Foundations of Augmented Cognition* (Lecture Notes in Computer Science, Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Cham, Switzerland: Springer Verlag, 2015, pp. 91–100, doi: 10.1007/978-3-319-20816-9_10.

[40] M. Hannula, K. Huttunen, J. Koskelo, T. Laitinen, and T. Leino, "Comparison between artificial neural network and multilinear regression models in an evaluation of cognitive workload in a flight simulator," *Comput. Biol. Med.*, vol. 38, nos. 11–12, pp. 1163–1170, Nov. 2008, doi: 10.1016/j.compbiomed.2008.09.007.

[41] A. R. Harrivel, C. Liles, C. L. Stephens, K. K. Ellis, L. J. Prinzel, and A. T. Pope, "Psychophysiological sensing and state classification for attention management in commercial aviation," in *Proc. AIAA Infotech @ Aerosp.*, Jan. 2016, p. 1490, doi: 10.2514/6.2016-1490.

[42] V. C. Roza, O. Postolache, V. Groza, and J. M. D. Pereira, "Emotions assessment on simulated flights," in *Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA)*, Jun. 2019, pp. 1–6, doi: 10.1109/MeMeA.2019.8802194.

[43] E. Q. Wu, X. Y. Peng, C. Z. Zhang, J. X. Lin, and R. S. F. Sheng, "Pilots' fatigue status recognition using deep contractive autoencoder network," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 10, pp. 3907–3919, Oct. 2019, doi: 10.1109/TIM.2018.2885608.

[44] K. J. Jaquess, R. J. Gentili, L.-C. Lo, H. Oh, J. Zhang, J. C. Rietschel, M. W. Miller, Y. Y. Tan, and B. D. Hatfield, "Empirical evidence for the relationship between cognitive workload and attentional reserve," *Int. J. Psychophysiol.*, vol. 121, pp. 46–55, Nov. 2017, doi: 10.1016/j.ijpsycho.2017.09.007.

[45] P. Besson, C. Bourdin, L. Bringoux, E. Dousset, C. Maïano, T. Marqueste, D. R. Mestre, S. Gaetan, J.-P. Baudry, and J.-L. Vercher, "Effectiveness of physiological and psychological features to estimate helicopter pilots' workload: A Bayesian network approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 4, pp. 1872–1881, Dec. 2013, doi: 10.1109/TITS.2013.2269679.

[46] J. A. Blanco, M. K. Johnson, K. J. Jaquess, H. Oh, L.-C. Lo, R. J. Gentili, and B. D. Hatfield, "Quantifying cognitive workload in simulated flight using passive, dry EEG measurements," *IEEE Trans. Cogn. Develop. Syst.*, vol. 10, no. 2, pp. 373–383, Jun. 2018, doi: 10.1109/TCDS.2016.2628702.

[47] L. Yingxue and W. Qi, "Pilots' brain cognitive state inference based on remaining life HSMM," in *Proc. Chin. Control Decis. Conf. (CCDC)*, Jun. 2019, pp. 5289–5294, doi: 10.1109/CCDC.2019.8833272.

[48] I. Bargiotas, A. Nicolaï, P.-P. Vidal, C. Labourdette, N. Vayatis, and S. Buffat, "The complementary role of activity context in the mental workload evaluation of helicopter pilots: A multi-tasking learning approach," in *Human Mental Workload: Models and Applications*, vol. 1012. Cham, Switzerland: Springer, 2019, doi: 10.1007/978-3-030-14273-5_13.

[49] J. Kacer, P. Kutilek, V. Krivanek, R. Doskocil, P. Smrcka, and Z. Krupka, "Measurement and modelling of the behavior of military pilots," in *Modelling and Simulation for Autonomous Systems* (Lecture Notes in Computer Science, Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10756. Cham, Switzerland: Springer Verlag, 2018, pp. 434–449, doi: 10.1007/978-3-319-76072-8_32.

[50] Z. Cai, Q. Wu, D. Huang, L. Ding, B. Yu, R. Law, J. Huang, and S. Fu, "Cognitive state recognition using wavelet singular entropy and ARMA entropy with AFPA optimized GP classification," *Neurocomputing*, vol. 197, pp. 29–44, Jul. 2016, doi: 10.1016/j.neucom.2016.01.054.

[51] E. Massé, O. Bartheye, and L. Fabre, "Classification of electrophysiological signatures with explainable artificial intelligence: The case of alarm detection in flight simulator," *Frontiers Neuroinform.*, vol. 16, Jun. 2022, Art. no. 904301, doi: 10.3389/fninf.2022.904301.

[52] Y. Ding, Y. Cao, V. G. Duffy, Y. Wang, and X. Zhang, "Measurement and identification of mental workload during simulated computer tasks with multimodal methods and machine learning," *Ergonomics*, vol. 63, no. 7, pp. 896–908, Jul. 2020, doi: 10.1080/00140139.2020.1759699.

[53] T. Pan, H. Wang, H. Si, Y. Li, and L. Shang, "Identification of pilots' fatigue status based on electrocardiogram signals," *Sensors*, vol. 21, no. 9, p. 3003, Apr. 2021, doi: 10.3390/s21093003.

[54] S. Scannella, V. Peysakhovich, F. Ehrig, E. Lepron, and F. Dehais, "Assessment of ocular and physiological metrics to discriminate flight phases in real light aircraft," *Human Factors, J. Human Factors Ergonom. Soc.*, vol. 60, no. 7, pp. 922–935, Nov. 2018, doi: 10.1177/0018720818787135.

[55] H. T. 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," *Sci. Rep.*, vol. 13, no. 1, p. 2507, Feb. 2023, doi: 10.1038/s41598-023-29647-0.

[56] N. Friedman, T. Fekete, K. Gal, and O. Shriki, "EEG-based prediction of cognitive load in intelligence tests," *Frontiers Human Neurosci.*, vol. 13, p. 191, Jun. 2019, doi: 10.3389/fnhum.2019.00191.

[57] E. Q. Wu, P.-Y. Deng, X.-Y. Qiu, Z. Tang, W.-M. Zhang, L.-M. Zhu, H. Ren, G.-R. Zhou, and R. S. F. Sheng, "Detecting fatigue status of pilots based on deep learning network using EEG signals," *IEEE Trans. Cogn. Develop. Syst.*, vol. 13, no. 3, pp. 575–585, Sep. 2021, doi: 10.1109/TCDS.2019.2963476.

[58] C. Y. Yiu, K. K. H. Ng, X. Li, X. Zhang, Q. Li, H. S. Lam, and M. H. Chong, "Towards safe and collaborative aerodrome operations: Assessing shared situational awareness for adverse weather detection with EEG-enabled Bayesian neural networks," *Adv. Eng. Informat.*, vol. 53, Aug. 2022, Art. no. 101698, doi: 10.1016/j.aei.2022.101698.

[59] S. Yang, Z. Yin, Y. Wang, W. Zhang, Y. Wang, and J. Zhang, "Assessing cognitive mental workload via EEG signals and an ensemble deep learning classifier based on denoising autoencoders," *Comput. Biol. Med.*, vol. 109, pp. 159–170, Jun. 2019, doi: 10.1016/j.compbiomed.2019.04.034.

[60] H. Wang, N. Jiang, T. Pan, H. Si, Y. Li, and W. Zou, "Cognitive load identification of pilots based on physiological-psychological characteristics in complex environments," *J. Adv. Transp.*, vol. 2020, pp. 1–16, Nov. 2020, doi: 10.1155/2020/5640784.

[61] D. H. Snider, S. E. Linnville, J. B. Phillips, and G. M. Rice, "Predicting hypoxic hypoxia using machine learning and wearable sensors," *Biomed. Signal Process. Control*, vol. 71, Jan. 2022, Art. no. 103110, doi: 10.1016/j.bspc.2021.103110.

[62] V. Peysakhovich, W. Ledegang, M. Houben, and E. Groen, "Classification of flight phases based on pilots' visual scanning strategies," in *Proc. Symp. Eye Tracking Res. Appl.*, Jun. 2022, pp. 1–7, doi: 10.1145/3517031.3529641.

[63] L. Pang, L. Guo, J. Zhang, X. Wanyan, H. Qu, and X. Wang, "Subject-specific mental workload classification using EEG and stochastic configuration network (SCN)," *Biomed. Signal Process. Control*, vol. 68, Jul. 2021, Art. no. 102711, doi: 10.1016/j.bspc.2021.102711.

[64] 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," *Rel. Eng. Syst. Saf.*, vol. 238, Oct. 2023, Art. no. 109449, doi: 10.1016/j.ress.2023.109449.

[65] L. Ke, G. Zhang, J. He, Y. Li, Y. Li, X. Liu, and P. Fang, "Pilot selection in the era of virtual reality: Algorithms for accurate and interpretable machine learning models," *Aerospace*, vol. 10, no. 5, p. 394, Apr. 2023, doi: 10.3390/aerospace10050394.

[66] A. Hernández-Sabaté, J. Yauri, P. Folch, M. À. Piera, and D. Gil, "Recognition of the mental workloads of pilots in the cockpit using EEG signals," *Appl. Sci.*, vol. 12, no. 5, p. 2298, Feb. 2022, doi: 10.3390/app12052298.

[67] Z. Gomolka, E. Zeslawska, B. Twarog, D. Kordos, and P. Rzucidlo, "Use of a DNN in recording and analysis of operator attention in advanced HMI systems," *Appl. Sci.*, vol. 12, no. 22, p. 11431, Nov. 2022, doi: 10.3390/app122211431.

[68] G. T. Lorenz, J. S. Ehrenstrom, T. B. Ullmann, R. C. Palmer, N. L. Tenhundfeld, E. J. de Visser, B. T. Donadio, and C. C. Tossell, "Assessing control devices for the supervisory control of autonomous wingmen," in *Proc. Syst. Inf. Eng. Design Symp. (SIEDS)*, Apr. 2019, pp. 1–6, doi: 10.1109/SIEDS.2019.8735606.

[69] J. Chen, L. Xue, J. Rong, and X. Gao, "Real-time evaluation method of flight mission load based on sensitivity analysis of physiological factors," *Chin. J. Aeronaut.*, vol. 35, no. 3, pp. 450–463, Mar. 2022, doi: 10.1016/j.cja.2021.11.010.

[70] B. Berthelot, P. Mazoyer, S. Egea, J. M. André, È. Grivel, and P. Legrand, "Self-affinity of an aircraft pilot's gaze direction as a marker of visual tunneling," SAE Tech. Papers 2019-01-1852, Sep. 2019, pp. 1–11, doi: 10.4271/2019-01-1852.

[71] Z. Xi, O. Newton, G. McGowin, G. Sukthankar, S. Fiore, and K. Oden, "Predicting student flight performance with multimodal features," in *Proc. 13th Int. Conf. Social Comput., Behav.-Cultural Model. Predict. Behav. Represent. Modelling Simulation (SBP-BRiMS)*, in Lecture Notes in Computer Science, vol. 12268. Orlando, FL, USA: Univ. Central Florida, 2020, pp. 277–287, doi: 10.1007/978-3-030-61255-9_27.

[72] W. Huang and C. Wang, "Fatigue detection and analysis based on multi-channel data fusion," in *Cyber Security Intelligence and Analytics*. Cham, Switzerland: Springer, 2022, pp. 650–656, doi: 10.1007/978-3-030-97874-7_85.

[73] S. G. Hajra, P. Xi, and A. Law, "A comparison of ECG and EEG metrics for in-flight monitoring of helicopter pilot workload," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2020, pp. 4012–4019, doi: 10.1109/SMC42975.2020.9283499.

[74] S. Ding, X. Pan, R. Han, X. Zeng, Y. Li, and X. Zheng, "A machine learning approach to reduce mental fatigue risk of pilots based on HRV data," in *Proc. 12th Int. Conf. Qual., Rel., Risk, Maintenance, Saf. Eng. (QR2MSE)*, Jul. 2022, pp. 546–551, doi: 10.1049/icp.2022.2923.

[75] 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 Trans. Syst., Man, Cybern. Syst.*, vol. 53, no. 2, pp. 6426–6437, Oct. 2023, doi: 10.1109/TSMC.2023.3282635.

[76] J.-H. Jeong, B.-W. Yu, D.-H. Lee, and S.-W. Lee, "Classification of drowsiness levels based on a deep spatio-temporal convolutional bidirectional LSTM network using electroencephalography signals," *Brain Sci.*, vol. 9, no. 12, p. 348, Nov. 2019, doi: 10.3390/brainsci9120348.

[77] S. Samani, R. Jessop, and A. Harrivel, "Collaborative communications between a human and a resilient safety support system," in *Proc. IEEE Int. Conf. Auto. Syst. (ICAS)*, Hampton, VA, USA, Aug. 2021, pp. 1–5, doi: 10.1109/ICAS49788.2021.9551108.

[78] D.-H. Lee, J.-H. Jeong, K. Kim, B.-W. Yu, and S.-W. Lee, "Continuous EEG decoding of pilots' mental states using multiple feature block-based convolutional neural network," *IEEE Access*, vol. 8, pp. 121929–121941, 2020, doi: 10.1109/ACCESS.2020.3006907.

[79] L. R. D. Murthy and P. Biswas, "Deep learning-based eye gaze estimation for military aviation," in *Proc. IEEE Aerosp. Conf. (AERO)*, Mar. 2022, pp. 1–8, doi: 10.1109/AERO53065.2022.9843506.

[80] V. Socha, J. Videnský, S. Kusmirek, L. Hanáková, and V. Valenta, "Design of wearable eye tracker with automatic cockpit areas of interest recognition," in *Proc. New Trends Civil Aviation*, Oct. 2022, pp. 109–114, doi: 10.23919/NTCA55899.2022.9934332.

[81] Q. A. Khan, A. Hassan, S. Rehman, and F. Riaz, "Detection and classification of pilots cognitive state using EEG," in *Proc. 2nd IEEE Int. Conf. Comput. Intell. Appl. (ICCIA)*, Sep. 2017, pp. 407–410, doi: 10.1109/CIAPP.2017.8167249.

[82] H. Zhuang, B. Yang, B. Li, P. Zan, B. Ma, and X. Meng, "EEG based eye movements multi-classification using convolutional neural network," in *Proc. 40th Chin. Control Conf. (CCC)*, Jul. 2021, pp. 7191–7195, doi: 10.23919/CCC52363.2021.9550039.

[83] S. Samima and M. Sarma, "EEG-based mental workload estimation," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 5605–5608, doi: 10.1109/EMBC.2019.8857164.

[84] T. Gateau, H. Ayaz, and F. Dehais, "In silico vs. over the clouds: On-the-fly mental state estimation of aircraft pilots, using a functional near infrared spectroscopy based passive-BCI," *Frontiers Human Neurosci.*, vol. 12, p. 187, May 2018, doi: 10.3389/fnhum.2018.00187.

[85] C. Benaroch, K. Sadatnejad, A. Roc, A. Appriou, T. Monseigne, S. Pramij, J. Mladenovic, L. Pillette, C. Jeunet, and F. Lotte, "Long-term BCI training of a tetraplegic user: Adaptive Riemannian classifiers and user training," *Frontiers Human Neurosci.*, vol. 15, pp. 1–22, Mar. 2021, doi: 10.3389/fnhum.2021.635653.

[86] G. Borghini, P. Aricò, G. Di Flumeri, N. Sciaraffa, A. Colosimo, M.-T. Herrero, A. Bezerianos, N. V. Thakor, and F. Babiloni, "A new perspective for the training assessment: Machine learning-based neurometric for augmented user's evaluation," *Frontiers Neurosci.*, vol. 11, pp. 1–15, Jun. 2017, doi: 10.3389/fnins.2017.00325.

[87] F. Sauvet, C. Bougard, M. Coroenne, L. Lely, P. Van Beers, M. Elbaz, M. Guillard, D. Léger, and M. Chennaoui, "In-flight automatic detection of vigilance states using a single EEG channel," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 12, pp. 2840–2847, Dec. 2014, doi: 10.1109/TBME.2014.2331189.

[88] G. Jiang, H. Chen, C. Wang, and P. Xue, "Mental workload artificial intelligence assessment of pilots' EEG based on multi-dimensional data fusion and LSTM with attention mechanism model," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 36, no. 11, pp. 1–19, Sep. 2022, doi: 10.1142/S0218001422590352.

[89] D. H. Klyde, A. Lampton, D. G. Mitchell, C. Berka, and M. Rhinehart, "A new approach to aircraft handling qualities prediction," in *Proc. AIAA Sci. Technol. Forum Expo., AIAA SciTech Forum*, Hawthorne, CA, USA: American Institute of Aeronautics and Astronautics, Jan. 2021, pp. 1–15. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85100398142&partnerID=40&md5=2560a450c304cdd634ed564fc49baaa9

[90] C. Lounis, V. Peysakhovich, and M. Causse, "Visual scanning strategies in the cockpit are modulated by pilots' expertise: A flight simulator study," *PLoS ONE*, vol. 16, no. 2, Feb. 2021, Art. no. e0247061, doi: 10.1371/journal.pone.0247061.

[91] H. Shuang, W. Chuanfeng, and W. Qi, "Recognition of fatigue status of pilots based on deep sparse auto-encoding network," in *Proc. 36th Chin. Control Conf. (CCC)*, Jul. 2017, pp. 10945–10950, doi: 10.23919/ChiCC.2017.8029103.

[92] A. Mishra, K. K. Shrivastava, A. B. Anto, and N. A. Quadir, "Reducing commercial aviation fatalities using support vector machines," in *Proc. Int. Conf. Smart Syst. Inventive Technol. (ICSSIT)*, Nov. 2019, pp. 360–364, doi: 10.1109/ICSSIT46314.2019.8987752.

[93] W. Huang, C. Wang, H.-B. Jia, P. Xue, and L. Wang, "Modeling and analysis of fatigue detection with multi-channel data fusion," *Int. J. Adv. Manuf. Technol.*, vol. 122, no. 1, pp. 291–301, Sep. 2022, doi: 10.1007/s00170-022-09364-0.

[94] S. Tortora, G. Beraldo, F. Bettella, E. Formaggio, M. Rubega, A. Del Felice, S. Masiero, R. Carli, N. Petrone, E. Menegatti, and L. Tonin, "Neural correlates of user learning during long-term BCI training for the cybathlon competition," *J. Neuroeng. Rehabil.*, vol. 19, no. 1, pp. 1–19, Jul. 2022, doi: 10.1186/s12984-022-01047-x.

[95] A. Kamrud, B. Borghetti, C. S. Kabban, and M. Miller, "Generalized deep learning EEG models for cross-participant and cross-task detection of the vigilance decrement in sustained attention tasks," *Sensors*, vol. 21, no. 16, p. 5617, Aug. 2021, doi: 10.3390/s21165617.

[96] Y. Li, K. Li, S. Wang, Y. Li, J. Chen, and D. Wen, "Towards safer flights: A multi-modality fusion technology-based cognitive load recognition framework," in *Proc. IEEE 4th Int. Conf. Civil Aviation Saf. Inf. Technol. (ICCASIT)*, Oct. 2022, pp. 525–530, doi: 10.1109/ICCASIT55263.2022.9986937.

[97] Q. Wang, Z. Wang, R. Xiong, X. Liao, and X. Tan, "A method for classification and evaluation of pilot's mental states based on CNN," *Comput. Syst. Sci. Eng.*, vol. 46, no. 2, pp. 1999–2020, 2023, doi: 10.32604/csse.2023.034183.

[98] B. Binias, D. Myszor, H. Palus, and K. A. Cyran, "Prediction of pilot's reaction time based on EEG signals," *Frontiers Neuroinform.*, vol. 14, p. 6, Feb. 2020. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fninf.2020.00006

[99] I. Alreshidi, D. Bisandu, and I. Moulitsas, "Illuminating the neural landscape of pilot mental states: A convolutional neural network approach with Shapley additive explanations interpretability," *Sensors*, vol. 23, no. 22, pp. 1–21, Nov. 2023, doi: 10.3390/s23229052.

[100] I. Alreshidi, I. Moulitsas, and K. W. Jenkins, "Multimodal approach for pilot mental state detection based on EEG," *Sensors*, vol. 23, no. 17, p. 7350, Aug. 2023, doi: 10.3390/s23177350.

[101] I. Alreshidi, S. Yadav, I. Moulitsas, and K. Jenkins, "A comprehensive analysis of machine learning and deep learning models for identifying pilots' mental states from imbalanced physiological data," in *Proc. AIAA AVIATION Forum*, Jun. 2023, p. 4529, doi: 10.2514/6.2023-4529.

[102] I. M. Alreshidi, I. Moulitsas, and K. W. Jenkins, "Miscellaneous EEG preprocessing and machine learning for pilots' mental states classification: Implications," in *Proc. 6th Int. Conf. Adv. Artif. Intell.*, Oct. 2022, pp. 29–39.

IBRAHIM ALRESHIDI received the dual M.Sc. degree in software engineering and cybersecurity from George Washington University. He received the Graduate Certificate in computer security and information assurance from George Washington University. He has been a Faculty Member with the University of Ha'il, since 2012, where he taught computer science classes. Before that, he gained professional experience as a Developer with Arab National Bank. He is currently a Researcher with Cranfield University and a member of the Digital Aviation Research and Technology Centre (DARTeC). His research focuses on the adoption of machine learning techniques and psychophysiological data in the field of aviation.

IRENE MOULITSAS is a Senior Lecturer with Cranfield University and a member of the Digital Aviation Research and Technology Centre (DARTeC), with substantial experience in large scale computing, algorithms, simulation, and modeling. She is the Course Director of the M.Sc. computational software techniques in engineering. Her research has been funded through national and regional bodies in USA and EU and has focused on developing novel algorithms for enabling the efficient execution of large scientific computations on serial and parallel processing platforms. She has developed highly efficient algorithms and software that are publicly available for use by universities, research laboratories, and companies.

KARL W. JENKINS is a Professor of computational engineering and the Head of the Centre for Computational Engineering Sciences. He was with industry for Allot and Lomax Consulting Engineers and Davy Distington Ltd, working on various commercial CFD codes and training engineers in their use. He has worked on adaptive parallel grid techniques, and has developed parallel codes for academic use and for blue chip companies, such as Rolls Royce plc and was actively involved initiating one of the Cambridge regional e-science center projects entitled "Grid Technology for Distance CFD." He has published over 100 papers and has won the Gaydon prize for the most significant paper contribution at a leading symposium on combustion held in Chicago. He is a member of the United Kingdom Consortium on Turbulent Reacting Flows (UKCTRF). He has invited to give numerous international and domestic seminars and to participate as a discussion panel member at international HPC DNS/LES conferences in the U.S.

https://dspace.org/

Advancing aviation safety through machine learning and psychophysiological data: a systematic review

Alreshidi, Ibrahim; Moulitsas, Irene; Jenkins, Karl W.

2024-01-03
Institute of Electrical and Electronics Engineers (IEEE)
https://doi.org/10.1109/ACCESS.2024.3349495
Alreshidi I, Moulitsas I, Jenkins KW. (2024) Advancing aviation safety through machine learning and psychophysiological data: a systematic review. IEEE Access, Volume 12, January 2024, pp. 5132-5150

Downloaded from a DSpace repository. https://dspace.org/