IBRAHIM M. ALRESHIDI

Flight Crew's Cognitive States Detection Using Psychophysiological Measurements and Machine Learning Techniques

SCHOOL OF AEROSPACE, TRANSPORT AND MANUFACTURING
Computational Engineering Sciences

DOCTOR OF PHILOSOPHY (PhD)
Academic Year: 2020 - 2023

Supervisor: Dr Irene Moulitsas
Associate Supervisor: Karl W. Jenkins
October 2023

SCHOOL OF AEROSPACE, TRANSPORT AND MANUFACTURING
Computational Engineering Sciences

DOCTOR OF PHILOSOPHY (PhD)
Academic Year 2020 - 2023

IBRAHIM M. ALRESHIDI

Flight Crew's Cognitive States Detection Using Psychophysiological Measurements and Machine Learning Techniques

Supervisor: Dr Irene Moulitsas
Associate Supervisor: Karl W. Jenkins
October 2023

This thesis is submitted in partial fulfilment of the requirements for the degree of PhD.
(NB. Remove if the degree award is based solely on examination of the thesis)

© Cranfield University 2023. All rights reserved. No part of this publication may be reproduced without the written permission of the copyright owner.

Academic integrity declaration

I declare that:

- the thesis submitted has been written by me alone.
- the thesis submitted has not been previously submitted to this university or any other.
- that all content, including primary and/or secondary data, is true to the best of my knowledge.
- that all quotations and references have been duly acknowledged according to the requirements of academic research.

I understand that to knowingly submit work in violation of the above statement will be considered by examiners as academic misconduct.

ABSTRACT

In the ever-evolving landscape of aviation safety, the accurate assessment of pilots' mental states is of paramount significance. This thesis elucidates the critical role of Electroencephalogram (EEG) data in comprehending pilots' cognitive conditions. The dataset, sourced from attention-related human performance limiting states, was publicly available on the NASA open portal website and encompasses EEG, electrocardiogram, galvanic skin response, and respiration data.

The initial analyses delved into the challenges posed by noise within EEG recordings. After rigorous testing, it was observed that prevalent preprocessing techniques, specifically band-pass filtering coupled with Independent Component Analysis, were not always effective. This inefficiency underscored the need for more advanced methodologies to optimize machine learning outcomes. In response, subsequent research stages proposed a hybrid ensemble learning approach. This innovative approach integrated advanced automated EEG preprocessing with Riemannian geometry. Through rigorous experimentation and validation, it was determined that this methodology accentuated the profound advantages of refined preprocessing, significantly enhancing the accuracy and reliability of EEG data interpretation.

As the inquiry advanced, a more integrative approach was adopted, amalgamating EEG with other physiological data. A novel methodology, synergizing one-dimensional Convolutional Neural Networks with Long Short-Term Memory architectures, was unveiled. Additionally, the impact of employing methods to handle data imbalance on machine learning performance was thoroughly examined. In the concluding phases, the research placed a heightened emphasis on model interpretability. Through the integration of SHapley Additive exPlanations values, a bridge was constructed between intricate model predictions and nuanced human comprehension, delineating paramount features for distinct cognitive states.

To encapsulate, this thesis offers a meticulous dissection of EEG data manipulation, machine learning, and deep learning constructs, positing a blueprint for the augmentation of aviation safety through in-depth cognitive state evaluations.

Keywords:

Electroencephalography; EEG; Machine Learning; Deep Learning; Mental State Classification; Resampling Techniques; Aviation Safety; Pilot Behaviour; Ensemble Learning; Pilot Deficiencies; Artefact Detection; Tangent Space; EEG Preprocessing; Heterogeneous Data

ACKNOWLEDGEMENTS

Embarking on this PhD journey has been an enlightening experience, and along this path, I have been fortunate to receive the unwavering support of many.

First and foremost, I would like to express my deepest gratitude to my family. Their unfaltering belief in me, their sacrifices, and their relentless encouragement have been the bedrock upon which I've built my academic pursuits. To my friends, who have been there to share in both the highs and lows, your camaraderie, understanding, and encouragement have been invaluable. My colleagues too, have enriched this journey with their insights, shared moments of intellectual challenge, and camaraderie.

A significant portion of my growth can be attributed to my primary supervisor, Dr Irene Moulitsas. Her dedication, expert guidance, and patience have not just shaped this research, but have also moulded me as an academic and a thinker. I am equally grateful to Prof Karl W. Jenkins, my associated supervisor. His wealth of knowledge, critical feedback, and unwavering support have been vital in the progression and quality of my work.

Special mention must be made to Dr David Barry and Dr Glenn Leighton. Their discerning eyes and constructive feedback during my PhD annual progress reviews have been pivotal. Their invaluable insights and encouragement have significantly enhanced the depth and direction of my research.

I wish to extend my warm appreciation to Cranfield University. This esteemed institution has not only been the setting of my PhD journey but has also equipped me with invaluable resources, academic networks, and a nurturing environment that fostered growth and innovation.

Finally, my heartfelt thanks go to the University of Ha'il. Their trust in my potential, coupled with their generous sponsorship, has not only made this academic endeavour possible but has also been a testament to their commitment to fostering research and academic excellence.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AOIs	Area Of Interests
AEN	Autoencoders
AdaBoost	Adaptive Boosting
AI	Artificial Intelligence
ANN	Artificial Neural Network
AHPLS	Attention-related Human Performance-Limiting States
ASR	Artefact Subspace Reconstruction
BNN	Bayesian neural networks
CA	Channelized Attention
CS	Cosine Similarity
CPB	Comprehending Pilot’s Behaviour
CNN	Convolutional Neural Network
DA	Diverted Attention
DNN	Deep Neural networks
DBN	Deep Belief Network
DCAEN	Deep-stacked Contractive Autoencoder Network
DL	Deep Learning
DRFAbd	Delta Ribcage-to-Abdomen
DRFThor	Delta Ribcage-to-Thoracic
DT	Decision Tree
DR	Dimensionality Reduction
ECG	Electrocardiogram
EDA	Electrodermal Activity

EEG	Electroencephalogram
EMG	Electromyography
EOG	Electrooculogram
ERP	Event-related potential
ET	Extra Tree
ERT	Extremely Randomised Trees
FT	Fine Tree
FFNN	Feed-Forward Neural Network
FFT	Fast Fourier Transform
GBM	Gradient Boosting Machines
GSR	Galvanic Skin Response
GP	Gaussian Process
HHT	Hilbert-Huang Transform
HMM	Hidden Markov Model
HSMM	Hidden Semi Markov Models
HF	High Frequency
HR	Heart Rate
HRV	Heart Rate Variability
ICA	Independent Component Analysis
KNN	K-Nearest Neighbours
MLP	Multi-Layer Perceptron
MIC	Mutual Information Coefficient
MLR	Multi-Linear Regression
Misc.	Miscellaneous
NB	Naïve Bayes

NE	No Event
PCA	Principal Component Analysis
PSD	Power Spectral Density
QDA	Quadratic Discriminant Analysis
Resp.	Respiration
RNN	Recurrent Neural Networks
REA	Regularization Algorithms
RF	Random Forest
SCN	Stochastic Configuration Network
SDThor	Standard Deviation Thoracic
SCR	Skin Conductance Response
SCL	Skin Conductance Level
SHAP	Shapley Values and SHapley Additive exPlanations
SDAbd	Standard Deviation Abdominal
SMOTE	Synthetic Minority Over-sampling Technique
SS	Startle/Surprise
SVM	Support Victor Machine
LDA	Linear Discriminant Analysis
LSFT	Lomb-Scargle Frequency Transform
Lasso	Least Absolute Shrinkage and Selection Operator
LSTM	Long Short-Term Memory
LR	Logistic Regression
LIME	Local Interpretable Model-agnostic Explanations
LVQ	Learning Vector Quantization
LF	Low Frequency

WT	Wavelet Transform
WPD	Wavelet Packet Decomposition
XGBoost	eXtreme Gradient Boosting

1 Introduction

The story of aviation unfolds like an enthralling tapestry of human aspiration, innovation, and exploration. Starting with the Wright brothers' inaugural powered flight in 1903, the aviation landscape has been transformed beyond recognition (Anderson, 1997; Bilstein, 2001). Each epoch brought forth faster, more efficient, and technologically sophisticated aircraft, revolutionising global travel and communication (Heppenheimer, 1995; Salas et al., 2010).

Yet, beneath every technological marvel, the pilot remains the cornerstone. From the early aviators, who were pioneers in the truest sense, to modern-day pilots navigating state-of-the-art jetliners, their roles have seen significant evolution (Wiegmann et al., 2002). They transitioned from mere aircraft operators to crucial decision-makers, interpreting a myriad of onboard systems, and data streams, and ensuring the unerring safety of every soul onboard (Rosekind et al., 1999).

Today's pilots endure immense mental workloads due to the complexity of current commercial cockpits. The influx of sophisticated avionics, navigation systems, and communication protocols requires pilots to juggle an array of responsibilities beyond just manoeuvring the aircraft (Holford, 2022). Compounding factors like irregular schedules, rapid time zone changes, and gruelling physical demands of flying further strain pilots' cognitive capacities. The technologically advanced cockpits of modern aeroplanes have dramatically increased the cognitive demands placed on pilots. Pilots must continuously process immense amounts of data, maintain situational awareness, and make critical decisions while always prioritising the safety of all passengers and crew (Fitts & Jones, 1947; Wang et al., 2022).

Any lapse in a pilot's mental state, whether from fatigue, stress or cognitive overload, could lead to catastrophic consequences (Zaslona et al., 2018). Therefore, the aviation research community has highlighted an urgent need to predict and monitor pilot mental states, ensuring that every flight is not just efficient, but unequivocally safe (Kelly & Efthymiou, 2019). New tools and

technologies are required to assess cognitive workloads and detect any hazardous mental lapses among pilots during flight operations. Such capabilities will be critical to prevent accidents and further enhance aviation safety amidst the ever-growing complexities of modern commercial airliners.

Electroencephalography (EEG) and other physiological sensors have been valuable research tools in neuroscience since their development (Müller-Putz et al., 2015). EEG in particular captures electrical brain activity, enabling insights into cognition, emotions, and neurological function. Over time, EEG and related technologies have been used to better understand and diagnose conditions, ranging from epilepsy to complex sleep anomalies (Ahmad et al., 2023; Dissanayake et al., 2021).

In aviation, there is growing interest in applying tools like EEG and physiological sensors to assess pilot mental states during flight (Borghini et al., 2014; Charles & Nixon, 2019; Van Weelden et al., 2022). By measuring brain activity and other biometric data, researchers aim to detect cognitive workload, stress levels, fatigue, and attention lapses in pilots. As air traffic increases steadily, such neural and physiological monitoring capabilities become critical for aviation safety research (Marinescu et al., 2018; Peißl et al., 2018). Integrating EEG with other sensors provides a more comprehensive view of the pilot's cognitive status than any single tool (Han et al., 2020). Understanding the brain patterns and physiology of pilots in real-world conditions could lead to enhanced training, cockpit designs, and flight procedures that reduce risk. While still an emerging application area, EEG and related technologies have considerable promise to strengthen aviation safety amidst the complexities of modern commercial flying.

The digital era has brought about a new age of data analysis capabilities. At the forefront are machine learning (ML) and deep learning (DL) techniques, which possess unmatched abilities to discern patterns and relationships within massive datasets. These powerful computational methods are able to continuously learn and improve by training on huge stores of data. ML has become revolutionary across diverse sectors like healthcare, finance, transportation and more, enabling transformative applications from precision

medicine to stock market forecasting (Loh & Nguyen, 2022). Where once human analysis of large data was constrained, ML provides the analytical horsepower to reveal novel insights and move industries forward based on previously unattainable knowledge from big data (Fogel & Kvedar, 2017).

The integration of ML and DL techniques has enabled remarkable new possibilities for extracting insights from EEG data that were previously considered unattainable (Buckova et al., 2020; Corsi et al., 2023). These computational methods now allow researchers to decode complex neural activity patterns, predict cognitive states, and even detect early neurological changes associated with cognitive decline. However, fully realising the potential of applying ML to neuroscience and EEG analysis involves substantial ongoing research and debate. Key challenges include developing models that are interpretable and meaningful, ensuring data quality and integrity, and addressing the intricacies of training robust deep neural networks on brain activity data (Lan et al., 2018). While ML and DL have opened new vistas for EEG analysis, translating these tools into a practical understanding of the brain requires overcoming current limitations related to model transparency, data handling, and network optimization. Advancing the convergence of neuroscience and cutting-edge Artificial intelligence (AI) will rely on continuous innovation to tackle these open challenges through multidisciplinary collaboration.

1.1 Study Motivation

In the evolving world of technology, aviation has emerged as an industry where innovation is both a constant and a necessity. Despite significant technological progress, it remains a realm in which a substantial portion of operations within the cockpit are carried out by the flight crew, underscoring the critical role of human involvement in aviation (Kelly & Efthymiou, 2019). Concerningly, the Commercial Aviation Safety Team (CAST) disclosed that out of 18 international aviation accidents associated with a loss of aircraft control, 16 incidents were ascribed to flight crew attention-related human performance-limiting states (AHPLS) (NTSB, 2019 ; SKY_Brary, 2019). Moreover, an array of flight

accidents in the past has been attributed to pilot failure resulting from factors such as fatigue, stress, and emotional turmoil (McKay & Groff, 2016; NTSB, 2019 ; Wiegmann & Shappell, 2017; Yen et al., 2009). In this context, understanding AHPLS among flight crew members – Channelized Attention (CA), Diverted Attention (DA), and Startle/Surprise (SS) – becomes paramount.

The literature documents several attempts to predict these cognitive states using various techniques. Notably, it has been established that psychophysiological measures such as EEG can provide an in-depth understanding of the mental demands on flight crew (Hankins & Wilson, 1998). Furthermore, ML methods offer promising avenues for generating safety-critical knowledge, which could prevent future accidents (Oehling & Barry, 2019). However, limited research has been carried out in this direction. Although attempts have been made to classify cognitive states using EEG signals coupled with ML methods (Chaudhuri & Routray, 2020; Dehais et al., 2019; Gao et al., 2019; Jiao et al., 2018; Sonnleitner et al., 2014; Wu et al., 2019; Zhang et al., 2019), the efficacy of such systems remains under studied. Similarly, a plethora of research efforts have focused on combining EEG signals with other peripheral physiological measures such as Electrocardiogram (ECG), Galvanic Skin Response (GSR), and Respiration (Resp.) (Ahn et al., 2016; Han et al., 2020; Hogervorst et al., 2014; Liu et al., 2017; Zhang et al., 2017). Despite their supposed benefits, they have not been adequately substantiated with empirical evidence. Furthermore, the application and effect of different traditional EEG preprocessing techniques on the ML model's performance have not been extensively explored. This lack of exploration extends to the impact of these techniques on predicting mental states using combined EEG data in varied environments.

Artefact handling in EEG data, particularly in the context of AHPLS detection, remains an unresolved challenge, compounded by the issue of data imbalance which is notably under-studied in this area. The dataset for AHPLS detection is heavily class imbalanced, a factor that prior studies have marginally addressed alongside efficient artefact handling or meaningful feature engineering. While

there are some attempts to investigate a robust detection system covering all AHPLS cognitive states, these often aggregate the states rather than considering them individually. This aggregation approach overlooks the inherent complexities and imbalances present in each state's data, underscoring the need for further study that specifically addresses these disparities.

Lastly, despite the availability of the robust AHPLS dataset, which includes data from pilots in diverse mental states, very few studies have aimed to detect AHPLS using ML techniques (Harrivel et al., 2016; Harrivel et al., 2017; Terwilliger, 2020). Moreover, those that did, encountered substantial limitations and failed to deliver convincing results. This scenario, therefore, calls for the development of innovative, more effective methodologies to fill the gaps in the literature.

1.2 Aim and Objectives

The overarching aim of this research is to develop and validate ML models for the robust prediction of pilots' cognitive states, specifically attention-related human performance-limiting states, utilising multimodal sensor data and advanced EEG preprocessing techniques.

The aim of this research is achieved by fulfilling the following stated objectives:

Objective 1: Investigate the influence of preprocessing techniques on ML models in EEG-based cognitive state prediction. This objective aims to enhance understanding of how different preprocessing approaches affect model performance using diverse environments data.

Objective 2: Develop advanced methods for EEG data preprocessing and mental state detection in pilots. The focus here is on enhancing artefact handling and overall performance of predictive models in this specific application.

Objective 3: Investigate the application of diverse ML models for mental state prediction using integrated physiological data. This includes a focus on addressing data imbalances to enhance model accuracy and reliability.

Objective 4: Contribute to the development of interpretable ML models for EEG-based mental state prediction. The goal is to not only achieve high accuracy but also to provide meaningful insights into the decision-making processes of these models.

1.3 List of Publications

This thesis adheres to the paper format structure specified in Section 7.1 of the Cranfield University student handbook. It consists of six technical chapters, specifically Chapters 2 through 6. Each of these chapters is structured as independent, succinct reports presented in the style of publications.

Segments of this thesis have been shared with the broader scholarly community via publications in peer-reviewed journals and presentations at both national and international conferences. An overview of all the published results from this undertaking can be found in Table 1-1:

Table 1-1 List of publications

S/No	Manuscript Title	Journal/Conference	Status
1	Advancing Aviation Safety Through Machine Learning and Psychophysiological Data: A Systematic Review	IEEE Access Reference: Alreshidi I., Moulitsas I. & Jenkins KW (2024) Advancing Aviation Safety Through Machine Learning and Psychophysiological Data: A Systematic Review, IEEE Access, vol. 12, pp. 5132-5150, 2024, doi: 10.1109/ACCESS.2024.3349495.	**Published**

2	Miscellaneous EEG Preprocessing and Machine Learning for Pilots' Mental States Classification: Implications	6th International Conference on Advances in Artificial Intelligence (ICAAI) Reference: Alreshidi IM, Moulitsas I & Jenkins KW (2023) Miscellaneous EEG preprocessing and machine learning for pilots' mental states classification: implications. In: 6th International Conference on Advances in Artificial Intelligence (ICAAI 2022), Birmingham, 21-23 October 2022. https://doi.org/10.1145/3571560.3571565	**Published**
3	Multimodal Approach for Pilot Mental State Detection Based on EEG	Sensors MDPI Reference: Alreshidi I, Moulitsas I & Jenkins KW (2023) Multimodal approach for pilot mental state detection based on EEG, Sensors, 23 (17) Article No. 7350. https://doi.org/10.3390/s23177350	**Published**
4	A Comprehensive Analysis of Machine Learning and Deep Learning Models for Identifying Pilots' Mental States from	2023 AIAA AVIATION Forum to Focus on Revolutionary Leaps Toward a New Age of Aviation Reference: Alreshidi I, Yadav S, Moulitsas I & Jenkins KW (2023) A comprehensive analysis of machine learning and deep	**Published**

	Imbalanced Physiological Data	learning models for identifying pilots' mental states from imbalanced physiological data. In: 2023 AIAA Aviation and Aeronautics Forum and Exposition (AIAA AVIATION Forum), San Diego, 12-16 June 2023. https://doi.org/10.2514/6.2023-4529	
5	Illuminating the neural landscape of pilot mental states: a convolutional neural network approach with SHapley Additive exPlanations interpretability	Sensors MDPI Reference: Alreshidi I, Bisandu D & Moulitsas I (2023) Illuminating the neural landscape of pilot mental states: a convolutional neural network approach with SHapley Additive exPlanations interpretability, Sensors, 23 (22) Article No. 9052. https://doi.org/10.3390/s23229052	**Published**

1.4 Adopted Research Methodology to Achieve the Stated Objectives

This section elucidates the research methodology adopted in this thesis, underpinning the systematic approach towards achieving the stated objectives. A cornerstone of this research is the comprehensive analysis of a specialised dataset, chosen for its relevance and depth in providing insights into the psychophysiological states of pilots. The following subsections detail the dataset employed and the methodological approaches tailored to each objective.

1.4.1 Dataset Description

The dataset pivotal to this research was obtained from NASA's open data portal and encompasses a unique compilation of experimental EEG and non-brain data collected from 18 pilots. The data collection was conducted in two distinct environments to simulate various mental states: the Langley Research Centre's (LaRC) Research Flight Deck and Cockpit Motion Facility for high-fidelity flight simulations, and a controlled non-flight environment for benchmark activities.

Cognitive State Definitions:

- **Channelized Attention (CA)**: This state occurs when a pilot's focus is intensely concentrated on a single task or piece of information, potentially leading to the neglect of other relevant cues.
- **Diverted Attention (DA)**: This state involves the pilot's attention being split or shifted between multiple tasks, leading to a potential decrease in the performance of these tasks.
- **Startle/Surprise (SS)**: This state is characterized by a sudden, unexpected event that disrupts the pilot's situational awareness and can lead to rapid physiological and cognitive responses.

In the flight simulator, pilots participated in Line-Oriented Flight Training (LOFT) scenarios, designed to induce specific cognitive states (Stephens et al., 2017). These included CA, as demonstrated in the Hydraulic System/Anti-Skid Event and Trailing Edge Flap Asymmetry; DA, observed in the ATC Taxi Clearance Event; and SS, effectively elicited in events like the Wake Encounter and Runway Incursion. These scenarios were created to mirror realistic flight situations, engaging pilots in tasks that required intense focus, quick response to unexpected occurrences, and management of multiple tasks, thus simulating the corresponding cognitive states. In the non-flight environment, cognitive states were induced through various activities: puzzle-based video games for CA, display monitoring tasks interspersed with math problems for DA, and exposure to jump-scare movie clips for SS. These activities were designed to

replicate the cognitive demands and responses similar to those experienced in flight scenarios but in a more controlled setting.

Data acquisition employed the Advanced Brain Monitoring X24 EEG and the Mind Media B.V. Nexus Mark II systems. The dataset provides four sets of data for each pilot: EEG, ECG, Resp., and GSR. While the majority of these datasets were captured in the non-flight environment, a crucial set was obtained from the flight simulator. This high-fidelity flight data set, featuring approximately one hour of labelled benchmark data, comprises 25 columns including a time stamp, 20 EEG channels, channels for ECG, respiration, and GSR, along with an event label. This provides a comprehensive view of the pilots' physiological states during the simulation. The EEG signals were captured using 20 electrodes placed according to the standard 10-20 system, plus an additional POz electrode. The channels included Fz, Cz, Pz, F3, F4, C3, C4, P3, P4, O1, O2, T5, T3, F7, Fp1, Fp2, F8, T4, and T6 with Linked Mastoids, all recorded at a sampling rate of 256 Hz. The dataset is segmented into one-second epochs without overlap, resulting in a combined dataset of 89,198 samples. These samples predominantly belonged to the Normal Event (NE) class, followed by CA, DA, and SS classes, with the majority (80%) being from the NE class, indicating a significant class imbalance.

1.4.2 Evaluation Metrics

Several indicators are being used to determine the reliability of the findings in each chapter of the thesis. The confusion matrix, sometimes referred to as the error matrix or contingency table, is vital for assessing the overall performance of the proposed models. It consists of four key elements in multiclass classification tasks: True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). TP represents instances accurately classified as positive for a specific class, while FP indicates instances wrongly classified as positive but belonging to another class. FN refers to instances incorrectly classified as negative for a specific class when they belong to it, and TN denotes instances accurately classified as negative for a specific class. These elements are employed to calculate evaluation metrics such as precision, recall,

and F1-score for each class, aiding in the assessment of a multiclass classification model's performance. The proposed model's classification performance on the testing set was evaluated using four confusion matrix-based metrics: accuracy, precision, recall, and F1-score. These performance measures are defined as follows:

- **Accuracy:** This metric measures the ratio of correctly classified instances to the total number of instances, essentially quantifying how many instances in the dataset are accurately classified by the model.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad \textbf{(1-1)}$$

- **Precision:** This metric gauges the ratio of true positive predictions to the total number of positive predictions, essentially determining how many instances classified as positive are genuinely positive.

$$Precision = \frac{TP}{TP + FP} \quad \textbf{(1-2)}$$

- **Recall:** This measure calculates the proportion of true positive predictions out of the total number of actual positive instances, indicating how many of the real positive instances are accurately identified as positive by the model.

$$Recall = \frac{TP}{TP + FN} \quad \textbf{(1-3)}$$

- **F1-score:** Serving as a balance between precision and recall, the F1-score is the harmonic mean of these two metrics and offers a single score that reflects the model's overall performance.

$$F1 - score = 2 * (Precision * Recall)/(Precision + Recall) \quad \textbf{(1-4)}$$

1.4.3 Methodological Approach for Each Objective

The literature review was done through an exhaustive systematic literature review following the guidelines (Kitchenham, 2007; Kitchenham et al., 2009; Moher et al., 2009). By leveraging key search terms pertinent to ML and flight

simulators, the researcher assembled, analysed, and synthesised literature from prominent databases.

To achieve **Objective 1**, this thesis undertook a comprehensive experimental procedure utilising pilots' EEG data from the AHPLS dataset across three scenarios: a non-flight environment, a flight simulator, and combined data from both contexts. These three data scenarios underwent three preprocessing conditions: unprocessed, filtered, and filtered with ocular artefacts excised through Independent Component Analysis (ICA). Each dataset post-preprocessing was subsequently deployed to train two discrete ML models: a Support Vector Machine (SVM) and an Artificial Neural Network (ANN).

In realising **Objective 2**, this thesis proposed an automated EEG preprocessing approach, coupled with a novel hybrid ensemble learning model. This preprocessing method incorporated filtering, Autoreject (Jas et al., 2017), and an MNE-Python function for automatic ocular-artifact removal (Gramfort et al., 2013), substantially mitigating non-cortical influences. Additionally, this thesis employed Riemannian geometry (Barachant et al., 2012) for feature extraction and a downsampling strategy, addressing challenges related to dataset imbalance and dimensionality. The extracted tangent space features from the EEG data were then employed to train several models including Random Forests (RF), Extremely Randomised Trees (ERT), Gradient Tree Boosting (GTB), Adaptive Boosting (AdaBoost), and a hybrid ensemble model that fused RF, ERT, and GTB, thereby successfully classifying four unique mental states.

For **Objective 3**, this thesis constructed a 1D-CNN coupled with an LSTM model trained on Power Spectral Density (PSD) features. These features were extracted from five frequency bands (delta, theta, alpha, beta, and gamma) of the EEG data, supplemented with filtered ECG, GSR, and Resp. data. To address data imbalance, this thesis utilised the Synthetic Minority Oversampling Technique (SMOTE) in one experiment, and a fusion of SMOTE and Cosine Similarity (CS) in another. Subsequent performance outcomes were compared to ascertain the model's efficacy, providing a thorough understanding of the developed model in contrast with conventional ML models like Feed-Forward

Neural Network (FFNN), eXtreme Gradient Boosting (XGBoost), AdaBoost, 1D-CNN, LSTM, and RF.

In **Objective 4**, this thesis architected an interpretable 1D-CNN model consisting of five layers, explicitly designed to discern the mental states of pilots using EEG data. The PSD features from five frequency bands: delta, theta, alpha, beta, and gamma, across all 20 EEG channels were meticulously extracted. To decode the model's decision-making, the SHAP technique was utilised to identify the top 10 influential features for CA, DA, SS, and Normal Event (NE) mental states, thereby enhancing model transparency and bolstering confidence in the generated results.

In accordance with the outlined objectives, the thesis has been structured to ensure a progressive elucidation of the methodologies and findings. Chapter 2 is a systematic review of the relevant literature is undertaken to explore the advancements in aviation safety achieved through ML and psychophysiological data. Chapter 3, which addresses Objective 1, presents a detailed experimental framework that investigates the influence of various EEG preprocessing techniques on ML models, using pilots' EEG data from the AHPLS dataset in diverse scenarios. This chapter lays the groundwork for understanding the impact of preprocessing on cognitive state prediction using ML models. Chapter 4, which addresses Objective 2, introduces an innovative approach for EEG data preprocessing and mental state detection in pilots, focusing on improving artefact handling and model performance in predictive analytics. Objective 3, encapsulated in Chapter 5, explores the application of different ML models, including 1D-CNN and LSTM, for analysing integrated physiological data (EEG, ECG, GSR, and Resp.) for mental state prediction. It also discusses methods to handle data imbalances, vital for model accuracy and reliability. Chapter 6, addressing Objective 4, presents the development of an interpretable 1D-CNN model for mental state prediction and the application of interpretative techniques to understand the decision-making processes of these models in the context of aviation safety. Building on the detailed exploration and analyses in the preceding chapters, Chapter 7, titled "Discussion", serves as a crucible for

synthesising the key findings, methodologies, and implications of the thesis. It meticulously dissects the collective insights gleaned from the experimental chapters, engaging in a rigorous discourse on the broader ramifications of the findings within the aviation safety domain. It also delves into a critical examination of the limitations and the potential avenues for future research. Chapter 8 concludes the thesis by stating the synopsis of research objectives, key findings, and the potential impact.

1.5 Appendices

1.5.1 Appendix A: Ethical Approval Letter

10 November 2020

Dear Mr Alreshidi ,

Reference: CURES/12413/2020
Title: Pilots mental state measurement in flight simulators using machine learning techniques

Thank you for your application to the Cranfield University Research Ethics System (CURES).

We are pleased to inform you your CURES application, reference CURES/12413/2020 **has been reviewed. You may now proceed with the research activities you have sought approval for.**

If you have any queries, please contact CURES Support.

We wish you every success with your project.

Regards,

CURES Team

Page 1 of 1

REFERENCES

Ahmad, I., Wang, X., Javeed, D., Kumar, P., Samuel, O. W., & Chen, S. (2023). A Hybrid Deep Learning Approach for Epileptic Seizure Detection in EEG signals. *IEEE J Biomed Health Inform*, *PP*. https://doi.org/10.1109/JBHI.2023.3265983

Ahn, S., Nguyen, T., Jang, H., Kim, J. G., & Jun, S. C. (2016). Exploring Neuro-Physiological Correlates of Drivers' Mental Fatigue Caused by Sleep Deprivation Using Simultaneous EEG, ECG, and fNIRS Data. *Front Hum Neurosci*, *10*, 219. https://doi.org/10.3389/fnhum.2016.00219

Anderson, J. J. D. (1997). *A History of Aerodynamics: And Its Impact on Flying Machines*. Cambridge University Press. https://doi.org/https://doi.org/10.1017/CBO9780511607158

Barachant, A., Bonnet, S., Congedo, M., & Jutten, C. (2012). Multiclass brain-computer interface classification by Riemannian geometry. *IEEE Trans Biomed Eng*, *59*(4), 920-928. https://doi.org/10.1109/TBME.2011.2172210

Bilstein, R. E. (2001). Flight in America: From the Wrights to the Astronauts. *Johns Hopkins University Press*. https://www.press.jhu.edu/books/title/1592/flight-america

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., & Babiloni, F. (2014). Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews*, *44*, 58-75.

Buckova, B., Brunovsky, M., Bares, M., & Hlinka, J. (2020). Predicting Sex From EEG: Validity and Generalizability of Deep-Learning-Based Interpretable Classifier. *Front Neurosci*, *14*, 589303. https://doi.org/10.3389/fnins.2020.589303

Charles, R. L., & Nixon, J. (2019). Measuring mental workload using physiological measures: A systematic review. *Applied ergonomics*, *74*, 221-232.

Chaudhuri, A., & Routray, A. (2020). Driver Fatigue Detection Through Chaotic Entropy Analysis of Cortical Sources Obtained From Scalp EEG Signals. *Ieee Transactions on Intelligent Transportation Systems*, *21*(1), 185-198. https://doi.org/10.1109/Tits.2018.2890332

Corsi, L., Liuzzi, P., Ballanti, S., Scarpino, M., Maiorelli, A., Sterpu, R., Macchi, C., Cecchi, F., Hakiki, B., Grippo, A., Lanatà, A., Carrozza, M. C., Bocchi, L., & Mannini, A. (2023). EEG asymmetry detection in patients with severe acquired brain injuries via machine learning methods. *Biomedical Signal Processing and Control*, *79*. https://doi.org/10.1016/j.bspc.2022.104260

Dehais, F., Dupres, A., Blum, S., Drougard, N., Scannella, S., Roy, R. N., & Lotte, F. (2019). Monitoring Pilot's Mental Workload Using ERPs and Spectral

Power with a Six-Dry-Electrode EEG System in Real Flight Conditions. *Sensors (Basel)*, *19*(6). https://doi.org/10.3390/s19061324

Dissanayake, T., Fernando, T., Denman, S., Sridharan, S., & Fookes, C. (2021). Deep Learning for Patient-Independent Epileptic Seizure Prediction Using Scalp EEG Signals. *Ieee Sensors Journal*, *21*(7), 9377-9388. https://doi.org/10.1109/Jsen.2021.3057076

Fitts, P., & Jones, R. (1947). *Analysis of factors contributing to 460 "pilot-error" experiences in operating aircraft controls*.

Fogel, A. L., & Kvedar, J. C. (2017). Benefits and risks of machine learning decision support systems. *Jama*, *318*(23), 2356-2356.

Gao, Z., Wang, X., Yang, Y., Mu, C., Cai, Q., Dang, W., & Zuo, S. (2019). EEG-Based Spatio-Temporal Convolutional Neural Network for Driver Fatigue Evaluation. *IEEE Trans Neural Netw Learn Syst*, *30*(9), 2755-2763. https://doi.org/10.1109/TNNLS.2018.2886414

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hamalainen, M. (2013). MEG and EEG data analysis with MNE-Python. *Front Neurosci*, *7*, 267. https://doi.org/10.3389/fnins.2013.00267

Han, S. Y., Kwak, N. S., Oh, T., & Lee, S. W. (2020). Classification of pilots' mental states using a multimodal deep learning network. *Biocybernetics and Biomedical Engineering*, *40*(1), 324-336. https://doi.org/10.1016/j.bbe.2019.12.002

Hankins, T. C., & Wilson, G. F. (1998). A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. *Aviat Space Environ Med*, *69*(4), 360-367. https://www.ncbi.nlm.nih.gov/pubmed/9561283

Harrivel, A. R., Liles, C., Stephens, C. L., Ellis, K. K., Prinzel, L. J., & Pope, A. T. (2016). Psychophysiological Sensing and State Classification for Attention Management in Commercial Aviation. AIAA Infotech @ Aerospace,

Harrivel, A. R., Stephens, C. L., Milletich, R. J., Heinich, C. M., Last, M. C., Napoli, N. J., Abraham, N., Prinzel, L. J., Motter, M. A., & Pope, A. T. (2017). Prediction of Cognitive States during Flight Simulation using Multimodal Psychophysiological Sensing. AIAA Information Systems-AIAA Infotech @ Aerospace,

Heppenheimer, T. A. (1995). *Turbulent skies : the history of commercial aviation*. J. Wiley & Sons.

Hogervorst, M. A., Brouwer, A. M., & van Erp, J. B. (2014). Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. *Front Neurosci*, *8*, 322. https://doi.org/10.3389/fnins.2014.00322

Holford, W. D. (2022). An ethical inquiry of the effect of cockpit automation on the responsibilities of airline pilots: Dissonance or meaningful control? *Journal of Business Ethics*, *176*(1), 141-157.

Jas, M., Engemann, D. A., Bekhti, Y., Raimondo, F., & Gramfort, A. (2017). Autoreject: Automated artifact rejection for MEG and EEG data. *Neuroimage*, *159*, 417-429. https://doi.org/10.1016/j.neuroimage.2017.06.030

Jiao, Z. C., Gao, X. B., Wang, Y., Li, J., & Xu, H. J. (2018). Deep Convolutional Neural Networks for mental load classification based on EEG data. *Pattern Recognition*, *76*, 582-595. https://doi.org/10.1016/j.patcog.2017.12.002

Kelly, D., & Efthymiou, M. (2019). An analysis of human factors in fifty controlled flight into terrain aviation accidents from 2007 to 2017. *Journal of Safety Research*, *69*, 155-165. https://doi.org/https://doi.org/10.1016/j.jsr.2019.03.009

Kitchenham, B. (2007). Guidelines for performing Systematic Literature Reviews in Software Engineering. *Technical report EBSE-2007-001, UK*.

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering - A systematic literature review. *Information and Software Technology*, *51*(1), 7-15. https://doi.org/10.1016/j.infsof.2008.09.009

Lan, K., Wang, D.-t., Fong, S., Liu, L.-s., Wong, K. K., & Dey, N. (2018). A survey of data mining and deep learning in bioinformatics. *Journal of medical systems*, *42*, 1-20.

Liu, Y., Ayaz, H., & Shewokis, P. A. (2017). Multisubject "Learning" for Mental Workload Classification Using Concurrent EEG, fNIRS, and Physiological Measures. *Front Hum Neurosci*, *11*, 389. https://doi.org/10.3389/fnhum.2017.00389

Loh, E., & Nguyen, T. (2022). Artificial intelligence for medical robotics. In *Endorobotics* (pp. 23-30). Elsevier. https://doi.org/10.1016/b978-0-12-821750-4.00002-5

Marinescu, A. C., Sharples, S., Ritchie, A. C., Sanchez Lopez, T., McDowell, M., & Morvan, H. P. (2018). Physiological parameter response to variation of mental workload. *Human Factors*, *60*(1), 31-56.

McKay, M. P., & Groff, L. (2016). 23 years of toxicology testing fatally injured pilots: Implications for aviation and other modes of transportation. *Accid Anal Prev*, *90*, 108-117. https://doi.org/10.1016/j.aap.2016.02.008

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group*, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine*, *151*(4), 264-269. https://doi.org/10.1136/bmj.b2535

Müller-Putz, G. R., Riedl, R., & C Wriessnegger, S. (2015). Electroencephalography (EEG) as a research tool in the information systems discipline: Foundations, measurement, and applications. *Communications of the Association for Information Systems*, *37*(1), 46.

NTSB. (2019). *Aircraft Accident Report Rapid Descent and Crash into Water*. N. T. S. Board.

Oehling, J., & Barry, D. J. (2019). Using machine learning methods in airline flight data monitoring to generate new operational safety knowledge from existing data. *Safety Science*, *114*, 89-104. https://doi.org/10.1016/j.ssci.2018.12.018

Peißl, S., Wickens, C. D., & Baruah, R. (2018). Eye-Tracking Measures in Aviation: A Selective Literature Review. *The International Journal of Aerospace Psychology*, *28*(3-4), 98-112. https://doi.org/10.1080/24721840.2018.1514978

Rosekind, M. R., Gander, P. H., Connell, L. J., & Co, E. L. (1999). *Crew Factors in Flight Operations X: Alertness Management in Flight Operations*. N.-N. T. R. Server. https://ntrs.nasa.gov/citations/19990116851

Salas, E., Maurino, D., & Curtis, M. (2010). Chapter 1 - Human Factors in Aviation: An Overview. In E. Salas & D. Maurino (Eds.), *Human Factors in Aviation (Second Edition)* (pp. 3-19). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-374518-7.00001-8

SKY_Brary. (2019). *SE211_ Airplane State Awareness - Training for Attention Management (R-D)*. https://skybrary.aero/articles/se211-airplane-state-awareness-training-attention-management-r-d

Sonnleitner, A., Treder, M. S., Simon, M., Willmann, S., Ewald, A., Buchner, A., & Schrauf, M. (2014). EEG alpha spindles and prolonged brake reaction times during auditory distraction in an on-road driving study. *Accid Anal Prev*, *62*, 110-118. https://doi.org/10.1016/j.aap.2013.08.026

Stephens, C. L., Harrivel, A., Prinzel, L. J., Comstock, R., Abraham, N., Pope, A. T., Wilkerson, J., & Kiggins, D. (2017). *Crew State Monitoring and Line-Oriented Flight Training for Attention Management* International Symposium on Aviation Psychology (ISAP 2017),

Terwilliger, P. S., Jack; Walker, Shannon; Harrivel, Angela. (2020). *A ResNet Autoencoder Approach for Time Series Classification of Cognitive State* MODSIM,

Van Weelden, E., Alimardani, M., Wiltshire, T. J., & Louwerse, M. M. (2022). Aviation and neurophysiology: A systematic review. *Applied ergonomics*, *105*, 103838.

Wang, H., Yu, Y., Li, S., & Wang, Q. (2022, 2-3 Dec. 2022). The Situation Perception of Pilots is Evaluated and Predicted Based on Neural Network.

2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC),

Wiegmann, D. A., Goh, J., & O'Hare, D. (2002). The Role of Situation Assessment and Flight Experience in Pilots' Decisions to Continue Visual Flight Rules Flight into Adverse Weather. *Human Factors*, *44*(2), 189-197. https://doi.org/10.1518/0018720024497871

Wiegmann, D. A., & Shappell, S. A. (2017). *A Human Error Approach to Aviation Accident Analysis: The Human Factors Analysis and Classification System*. CRC Press. https://books.google.co.uk/books?id=jitEDwAAQBAJ

Wu, E. Q., Peng, X. Y., Zhang, C. Z. Z., Lin, J. X., & Sheng, R. S. F. (2019). Pilots' Fatigue Status Recognition Using Deep Contractive Autoencoder Network. *Ieee Transactions on Instrumentation and Measurement*, *68*(10), 3907-3919. https://doi.org/10.1109/Tim.2018.2885608

Yen, J. R., Hsu, C. C., Yang, H., & Ho, H. (2009). An investigation of fatigue issues on different flight operations. *Journal of Air Transport Management*, *15*(5), 236-240. https://doi.org/10.1016/j.jairtraman.2009.01.001

Zaslona, J. L., O'Keeffe, K. M., Signal, T. L., & Gander, P. H. (2018). Shared responsibility for managing fatigue: Hearing the pilots. *PLoS One*, *13*(5), e0195530.

Zhang, P., Wang, X., Chen, J., & You, W. (2017). Feature Weight Driven Interactive Mutual Information Modeling for Heterogeneous Bio-Signal Fusion to Estimate Mental Workload. *Sensors (Basel)*, *17*(10). https://doi.org/10.3390/s17102315

Zhang, P., Wang, X., Zhang, W., & Chen, J. (2019). Learning Spatial-Spectral-Temporal EEG Features With Recurrent 3D Convolutional Neural Networks for Cross-Task Mental Workload Assessment. *IEEE Trans Neural Syst Rehabil Eng*, *27*(1), 31-42. https://doi.org/10.1109/TNSRE.2018.2884641

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

2.1 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 (ML) techniques, offering the potential to enhance safety by understanding pilot behaviour. This systematic literature review rigorously follows a widely accepted methodology, scrutinising 80 peer-reviewed studies out of 3352 studies from five key electronic databases. The paper focuses on behavioural aspects, data types, preprocessing techniques, ML models, and performance metrics used in existing studies. It reveals that the majority of research disproportionately concentrates on workload and fatigue, leaving behavioural aspects like emotional responses and attention dynamics less explored. ML models such as Tree-based and Support Vector Machines (SVM) are most commonly employed, but the utilisation of advanced techniques like Deep Learning (DL) 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 behaviour.

2.2 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 (Wiegmann et al., 2002). In light of the complex nature of contemporary aviation operations, a comprehensive understanding of pilot behaviour becomes paramount for enhancing aviation safety. 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 behaviour, including cognitive states and emotional responses. This paper serves as a systematic literature review, conducted in accordance with the rigorous methodological guidelines (Kitchenham, 2007; Kitchenham et al., 2009; Moher et al., 2009). It aims to offer an exhaustive synthesis of existing research on the application of ML techniques and psychophysiological data for understanding pilot behaviour.

2.2.1 Importance of Aviation Safety

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 (Oster Jr et al., 2013).

However, the achievement of optimal safety levels is a complex endeavour, influenced by a myriad of factors ranging from technological innovation to regulatory oversight (Henderson, 2022). Advances in technology have undeniably contributed to enhanced safety mechanisms, from state-of-the-art (SOTA) air traffic control systems to predictive maintenance algorithms that preempt mechanical failures. Nonetheless, the industry is not immune to challenges (Boyd, 2017; Li & Baker, 2007; Li et al., 2001; SKY_Brary, 2019). 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 (Aurino, 2000).

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 (Chung, 2017).

The importance of aviation safety is further underscored by empirical data on the occurrences that lead to accidents (SKY_Brary, 2019). As demonstrated in Figure 2-1, the predominant cause of accidents is 'Loss of Control In-Flight', accounting for a significant number of incidents. This category alone shows more than 2,000 accidents, a stark reminder of the complexity and inherent risks associated with flight operations. Of these, a notable proportion are fatal, indicating that loss of control during flight represents not just a prevalent challenge but also a critical area for safety enhancement. The data highlights that while system/component failures, both powerplant and non-powerplant, are also significant contributors to aviation incidents, it is the in-flight control loss that stands as the most consequential. This analysis provides a quantitative grounding for the qualitative understanding of aviation safety, emphasising the need for rigorous research into preventive measures, improved pilot training, and technological advancements that can mitigate such risks.

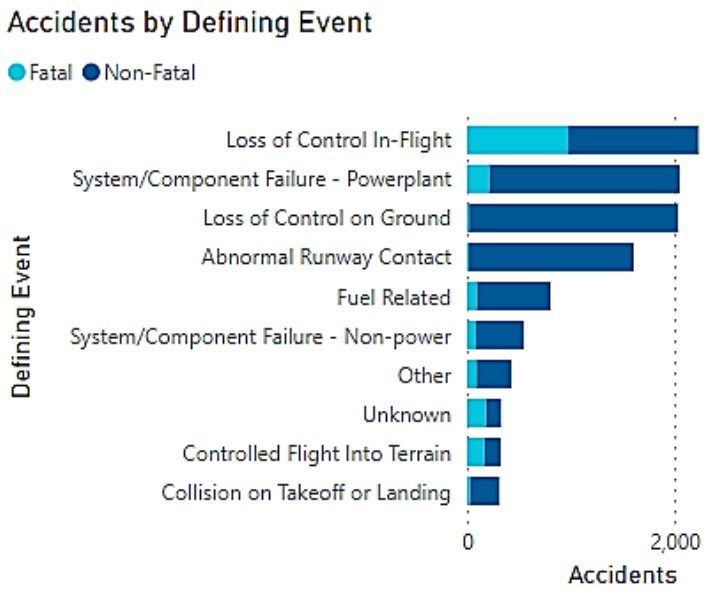

Figure 2-1 Bar chart of aviation accidents by defining event, contrasting fatal and non-fatal outcomes, with 'Loss of Control In-Flight' as the most prevalent cause (SKY_Brary, 2019).

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.

2.2.2 Role of Pilot Behaviour in Aviation

In the intricate system of aviation safety, the role of pilot behaviour 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 (Behrend & Dehais, 2020).

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 (Stanton et al., 2001). 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 (K et al., 2020; Marinescu et al., 2018).

Moreover, the role of pilot behaviour has systemic implications that ripple through the aviation safety ecosystem, influencing everything from regulatory frameworks to the design of new technologies (Sant'Anna & Hilal, 2021; Sarter et al., 2007; Stanton et al., 2019). 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 behaviour 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.

2.2.3 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 behaviour. These advanced computational techniques offer a comprehensive framework for analysing intricate, high-dimensional psychophysiological data sets like EEG, which are often beyond the scope of traditional statistical methods to interpret in a meaningful manner (Qin et al., 2021).

ML algorithms, encompassing a broad array of models such as tree-based, SVM, and various neural networks, have proven to be immensely effective in predicting and understanding multiple facets of pilot behaviour. 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 (Morgan Iii et al., 2007). 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 capitalises 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 (Chen et al., 2016). 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 (Jiang et al., 2023; Lee, Kim, & Choi, 2023; Li, Li, Wang, Chen, & Wen, 2022; Mohanavelu et al., 2022; Wu et al., 2022; Zhu et al., 2023).

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 behaviour. Following this introductory section, the paper delineates its systematic review methodology, presents a comprehensive synthesis of key findings, offers an extensive discussion contextualising these results within the broader landscape of aviation safety and pilot behaviour, and concludes by summarising the salient insights while identifying research gaps that offer promising avenues for future inquiry.

2.3 Related Work

This section examines the integration of ML and DL techniques with psychophysiological signal analysis to better understand human behaviour. It highlights research in emotion recognition and mental states detection, showcasing how ML and DL models are applied to interpret complex patterns of emotional and cognitive states through physiological signals. The focus here is on state-of-the-art approaches that have the potential to be adapted for the aviation industry, particularly in the analysis of pilot behaviour. We categorize our exploration into three main themes: a review of general human behaviour analysis with an emphasis on innovative approaches relevant to aviation, an overview of emotion recognition methods, and a detailed examination of mental states detection techniques.

2.3.1 A Review of General Human Behaviour Analysis

In recent years, the field of human behaviour analysis has seen significant advancements through the application of ML and DL techniques, particularly in the processing and interpretation of psychophysiological signals such as EEG. These advancements hold potential for various applications, including but not limited to detecting pilots' mental states in aviation contexts. This subsection explores state-of-the-art ML and DL approaches proposed in non-aviation contexts that could be instrumental in analysing pilots' mental states.

Deep learning, especially, has revolutionized EEG analysis, enabling robust automatic classification of signals which is crucial for practical applications in

neuroscience and beyond. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Deep Belief Networks (DBNs) have shown superior performance in classifying EEG signals across various tasks such as emotion recognition, motor imagery, and mental workload analysis (Craik et al., 2019). DL approaches not only automate the feature extraction process but also adapt to the high-dimensional nature of EEG data, managing to capture complex patterns related to cognitive states. Moreover, innovations in DL methodologies have facilitated end-to-end EEG analysis, eliminating the need for manual feature extraction and allowing for more nuanced interpretations of neural signals. Techniques such as Deep ConvNets have been applied successfully for EEG decoding and visualization, providing insights into the spatial and spectral features that contribute to various mental states (Schirrmeister et al., 2017). These advancements underscore the potential of DL in extracting meaningful information from EEG data, which can be particularly beneficial in the context of monitoring pilots' cognitive load and mental fatigue.

The integration of DL with novel computing platforms like IBM's neuromorphic TrueNorth chip offers a promising direction for deploying advanced EEG analysis techniques in low-power, minimal device footprints. This convergence could pave the way for real-time, efficient monitoring of cognitive states in high-stakes environments such as aviation (Nurse et al., 2016). Other approaches such as the use of Graph Neural Networks (GNNs) for analysing non-Euclidean data structures like graphs have been explored. GNNs could offer innovative ways to model complex relationships and interdependencies in EEG data, enhancing the understanding and classification of human behavioural states (Z. Wu et al., 2021). Moreover, the exploration of DL in EEG classification has also led to the development of specialized architectures and algorithms, such as Deep Extreme Learning Machines (DELMs), which combine the benefits of deep learning with the efficiency of Extreme Learning Machines (ELMs) for fast and effective EEG signal classification (Ding et al., 2015). These methodologies not only enhance classification accuracy but also reduce the computational burden, making them suitable for real-time applications.

The classification of brain signals for epilepsy detection using DL classification showcases the power of CNNs in analysing EEG signals. Remarkably, high classification accuracy can be achieved using only a single-channel EEG, indicating the potential for efficient and simplified analysis suitable for real-time monitoring of pilots' mental states (Liu & Woodson, 2019). Furthermore, DL algorithms, especially Long Short-Term Memory (LSTM) networks, have been implemented for predicting impending crises in patients with epilepsy and assessing prognostic risk factors in Parkinson's disease. These models significantly outperform traditional ML algorithms, demonstrating DL's potential to provide nuanced insights into complex mental states (Kannan et al., 2022).

Deep CNNs have revolutionized EEG analysis by automating the feature extraction and classification processes. This is particularly relevant for the real-time assessment of mental states, where CNNs can detect normal, preictal, and seizure classes with high accuracy, specificity, and sensitivity. Such an approach could be adapted to detect subtle changes in pilots' mental states, indicating attention shifts or the onset of surprise/startle reactions (Acharya et al., 2018). Deep metric learning is another innovative approach for epileptic seizure detection that addresses the challenges posed by small dataset sizes in EEG analysis. This methodology enhances the capability for real-time, accurate detection of seizures, which can be paralleled to the detection of sudden mental state changes in pilots, such as startle or surprise, that may impact flight safety (Duan et al., 2022). Moreover, the novel deep convolutional long short-term memory (C-LSTM) model demonstrates exceptional performance in detecting seizures and tumours from EEG signals. By predicting results in extremely short durations, this model underscores the potential for rapid and accurate monitoring of pilots' cognitive states, offering a promising avenue for exploring attentional dynamics (Liu et al., 2020).

This review highlights the potential of applying state-of-the-art ML and DL approaches, initially developed for epilepsy and general human behaviour analysis, to the aviation context. By leveraging these techniques, researchers and practitioners can improve the accuracy and efficiency of detecting and

analysing pilots' mental states, ultimately enhancing safety and performance in aviation operations.

2.3.2 An Overview of Emotion Recognition Methods

Emotion recognition stands as a pivotal area within the broader field of human behaviour analysis, particularly in its application of ML and DL techniques to interpret physiological signals. This subsection delves into the diverse methodologies employed in recent studies to classify emotions, showcasing the integration of both traditional ML techniques and advanced DL models. Through the examination of various research efforts, we explore the efficacy of these models in accurately classifying emotions from complex datasets like DEAP, which include EEG and peripheral physiological signals, among others. While highlighting the potential of these approaches, we also note their current limitations in terms of direct applicability to aviation-specific contexts, such as pilot mental state prediction, and the integration of additional physiological signals like ECG, GSR, and respiration. The following discussion provides a critical overview of the key studies.

Various studies have explored the use of both DL models and traditional ML techniques to classify emotions using physiological signals (Cecotti & Graser, 2011; Cesar Cavalcanti Roza & Adrian Postolache, 2019; Tripathi et al., 2017; Wei-Long & Bao-Liang, 2015). For example, Tripathi et al. (Tripathi et al., 2017) utilized a 1D-CNN+LSTM model for accurate emotion classification on the Dataset for Emotion Analysis using Physiological and Audiovisual Signals (DEAP) (Koelstra et al., 2012), which contains EEG and peripheral physiological signals. Similarly, Zheng and Lu (Wei-Long & Bao-Liang, 2015) investigated critical frequency bands and channels for EEG-based emotion recognition using a 1D-CNN+LSTM model. On the other hand, Bhardwaj et al. (Bhardwaj et al., 2015) employed Support Vector Machines (SVM) and Linear Discriminant Analysis (LDA) classifiers to classify human emotions from EEG signals. Although these studies employed DL and ML models for emotion recognition using EEG signals, they did not specifically focus on predicting pilots' mental states or incorporate other physiological signals such as ECG, GSR, and

respiration (Subasi, 2007). Roza et al. (Cesar Cavalcanti Roza & Adrian Postolache, 2019) employed a Multilayer Perceptron (MLP) model to identify emotions through analyzing physiological signals. The accuracy of the MLP model's performance varied between 55% and 100%, depending on the different sets of features used.

2.3.3 A Review of Mental States Detection Methods

This subsection reviews the efforts made to classify cognitive states by leveraging EEG signals alongside a variety of ML and DL approaches. It covers a range of studies, from early investigations employing statistical analyses of EEG data to identify fatigue levels and task complexity, to more recent advancements that incorporate DL techniques for more nuanced detection of mental workload and cognitive states without manual feature extraction. The discussion also extends to the integration of multimodal data sources, such as EEG, ECG, and fNIRS, demonstrating the potential for enhanced detection performance through the combination of various physiological signals.

Numerous attempts have been made to classify individuals' cognitive states by combining EEG signals with a variety of ML and DL approaches. Previous research by Lal et al. (Lal et al., 2003), Jap et al. (Jap et al., 2009), Kar et al. (Kar et al., 2010), and Trejo et al. (Trejo et al., 2015) investigated statistical alterations in EEG during driving simulation tasks to determine fatigue levels. Johnson et al. (Johnson et al., 2015) examined algorithms independent of probes for classifying three degrees of task complexity in an EEG-based flight simulator experiment. Binias et al. (Binias et al., 2018) implemented spatial pattern characteristics extracted from EEG signals and diverse ML methods to differentiate between specific brain activity states related to idle but focused visual cue anticipation and the following response. Sonnleitner et al. (Sonnleitner et al., 2014) applied regularized LDA to study the predictive power of EEG for detecting distraction in single-trial analyses. Chaudhuri et al. (Chaudhuri & Routray, 2020) focused on SVM classification of typical and fatigued states in a simulated setting using the source localization technique. Dehais et al. (Dehais et al., 2019) used frequency features derived from

shrinkage LDA to classify mental workload and typical states. Nevertheless, these investigations mainly depended on manually crafted EEG features for creating classifiers.

DL techniques have been increasingly adopted for identifying cognitive states without external support. For instance, Patel et al. (Patel et al., 2011) implemented a neural network to detect early signs of driver fatigue using ECG data. Bashivan et al. (Bashivan et al., 2016) proposed a deep recurrent CNN for identifying workload states from multi-channel EEG signals. Hajinoroozi et al. (Hajinoroozi et al., 2016) introduced a channel-wise CNN and a variation with restricted Boltzmann machine for determining suboptimal driver performance. Jiao et al. (Jiao et al., 2018) demonstrated a deep CNN method for detecting mental workload levels from EEG data, integrating a fusion strategy with a pointwise gated Boltzmann machine for various EEG inputs. Zhang et al. (Zhang et al., 2019) used a recurrent 3D CNN to learn spatial-spectral-temporal EEG features for assessing mental workload across tasks. Wu et al. (Wu et al., 2019) suggested a deep stacked contractive autoencoder network to learn fatigue-related features from raw EEG data for fatigue recognition. Gao et al. (Gao et al., 2019) developed an EEG-based spatial-temporal CNN for accurate fatigue state detection. However, it is essential to recognize that these studies focused solely on one type of signal for cognitive state detection.

Merging data from various biosignal sensors has proven to be a successful strategy for enhancing detection performance in comparison to single-sensor recognition. For instance, Hogervorst et al. (Hogervorst et al., 2014) explored combined features from physiological signals such as EEG, ECG, and eye blinks for mental workload assessment. Ahn et al. (Ahn et al., 2016) collected EEG, ECG, and Functional near-infrared spectroscopy (fNIRS) data simultaneously to examine the neurophysiological correlates of subjects' fatigue levels. Liu et al. (Liu et al., 2017) combined EEG, fNIRS, and physiological measures for workload classification, showcasing improved performance when fusing these modalities. Han et al. (Han et al., 2020) developed a multimodal

neural network architecture comprising CNN and LSTM models to identify distraction, workload, fatigue, and normal mental states.

2.4 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 (Kitchenham, 2007; Kitchenham et al., 2009; Moher et al., 2009), 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. Figure 2-2 presents a graphical description of the procedure.

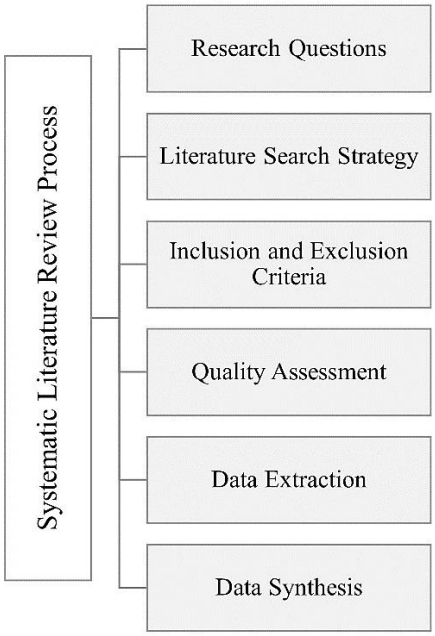

Figure 2-2 The adopted steps of the systematic review

2.4.1 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' behaviour?**
  - What behavioural and cognitive states are most studied?
- **RQ2: How are preprocessing, data types, and feature extraction approached in existing studies on psychophysiological data for pilot behaviour?**
  - Which psychophysiological data types are most used?
  - What artefacts 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 utilised to understand the pilot behaviour?**
  - Which evaluation mechanism and metrics were utilised to assess the models?
- **RQ4: What is the comparative performance of various ML and DL models in predicting pilot behaviour?**
  - 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?

2.4.2 Literature Search Strategy

The integrity of a systematic review is profoundly dependent on the comprehensiveness and rigour 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.

2.4.2.1 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 (Ian Goodfellow, 2016). Phrases such as “machine learning,” “psychophysiological data,” "EEG," and "pilot behaviour" were intricately woven together through Boolean operators like "AND" and "OR," fashioning a search net designed for both breadth and precision.

2.4.2.2 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.

2.4.2.3 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.

2.4.3 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 Behaviour:** Research specifically targeting pilot behaviour, either in real-world or simulated environments, was included.
3- **Machine Learning Models:** Studies employing ML or 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 behaviour, 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.

2.4.4 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 scrutinised.
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.

2.4.5 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.

2.4.5.1 Search Process

To synthesise 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 Figure 2-3.

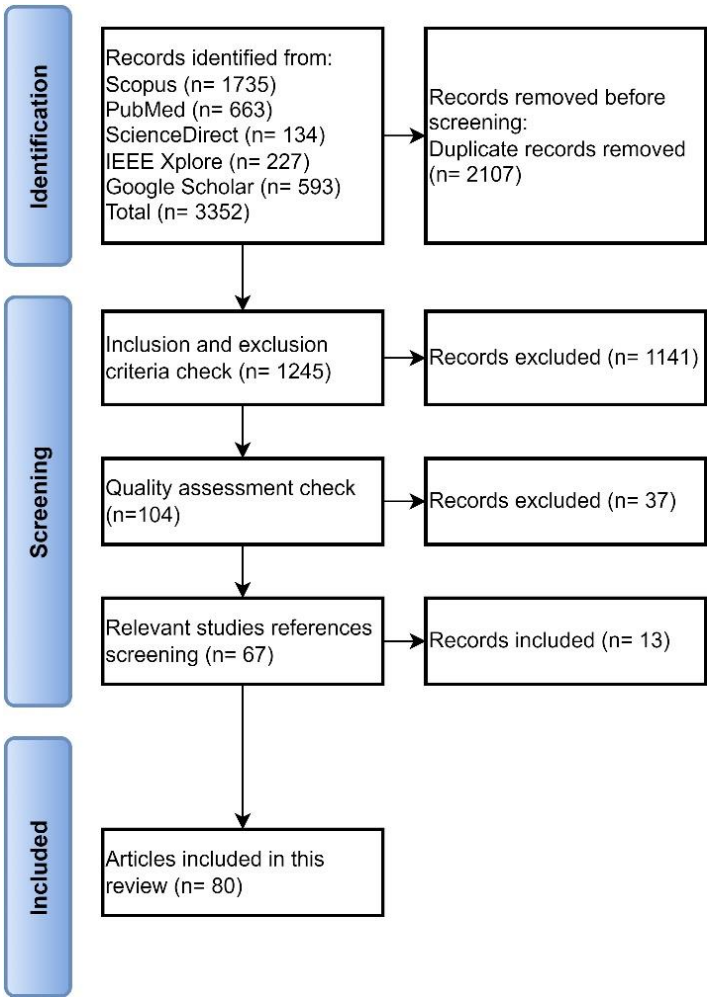

Figure 2-3 PRISMA flow diagram

2.4.5.2 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 meticulously 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. **Behavioural Aspects**: Specific behavioural 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 Machine Learning, Deep Learning, 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, ECG, and GSR were recorded to understand the range of data employed in assessing pilot behaviour.
9. **Preprocessing Techniques**: Methods used for preprocessing, such as Independent Component Analysis 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, wavelet coefficients, 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.

2.4.6 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, publication venues, and geographical distribution of studies. The second layer honed in on the behavioural aspects, where specific behavioural states like workload, fatigue, and attention, as well as emotional states, were analysed. 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 categorised and analysed 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 behaviour.

2.5 Results

The Results 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 synthesises the findings across multiple dimensions, including the types of ML models employed, their performance metrics, and the psychophysiological data types used for predicting pilot

behaviour. Furthermore, this section provides a granular breakdown of methodological choices in existing literature, including data preprocessing techniques, artefacts identified, and features extracted. The results presented herein aim to offer a comprehensive understanding of the current SOTA, serving as a foundational base for the subsequent Discussion section where these findings will be interpreted, contextualised, and evaluated.

2.5.1 Taxonomy of Pilot's Behavioural and Cognitive States

The taxonomy of behavioural and cognitive states in aviation-based empirical studies is visualised in Figure 2-4, 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.

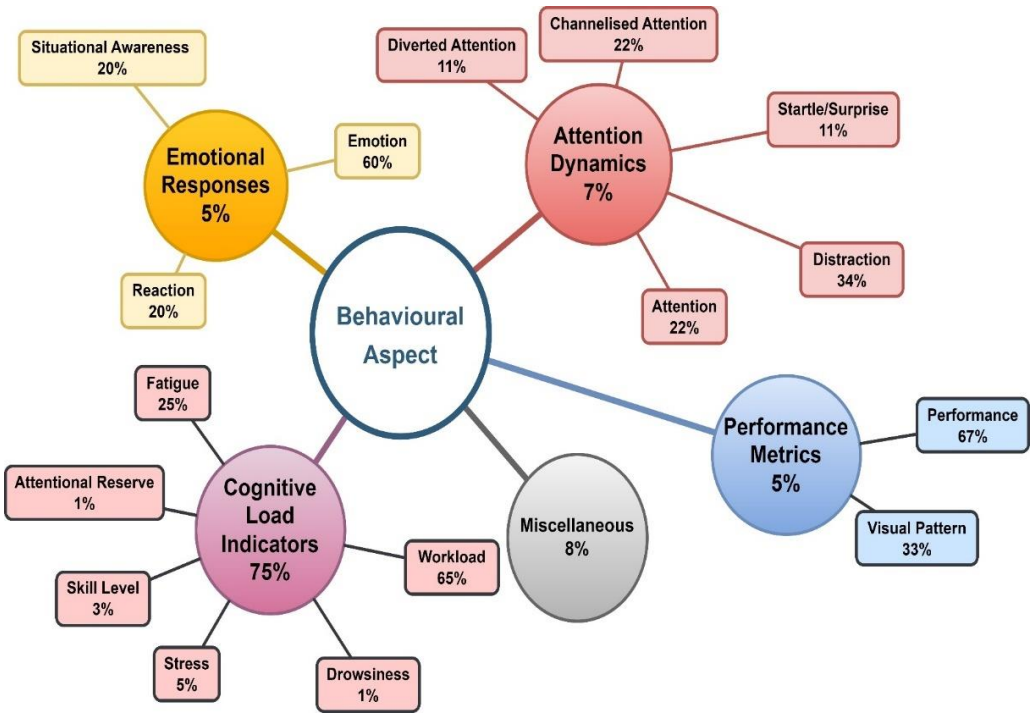

Figure 2-4 Conceptual map of behavioural aspects with associated percentage distributions, illustrating the interrelationships among Emotional Responses, Attention Dynamics, Cognitive Load Indicators, and Performance Metrics in pilot behaviour analysis.

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' 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 behavioural aspect was neither the central focus nor explicitly articulated.

2.5.2 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 artefacts identified, the methods used for data preprocessing, as well as the features extracted and their corresponding extraction techniques.

The distribution of psychophysiological data types used in research studies exhibits a notable range of diversity. As delineated in Figure 2-5, 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, Resp., Electrooculogram (EOG), and Electromyogram (EMG) data appear less frequently, each making up less than 5% of the studies.

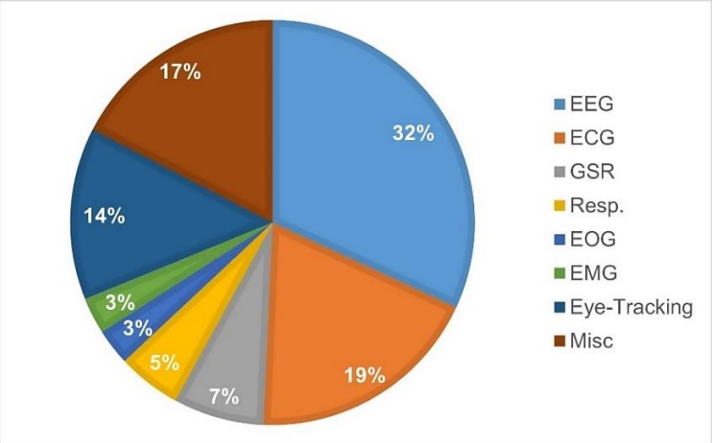

Figure 2-5 Comprehensive distribution of psychophysiological and other data types in existing literature on pilot behaviour

Turning to Table 2-1, a detailed inspection reveals a rich array of artefacts and their corresponding preprocessing methods, sorted by psychophysiological data type. EEG data, for instance, are predominantly subjected to preprocessing methods such as Independent Component Analysis (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 artefacts in the EEG data using ICA. For users of MATLAB, Artefact Subspace Reconstruction (ASR) is frequently employed. These techniques mitigate challenges posed by ocular and muscular artefacts 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 2-1 Summary of correlated behaviour aspects, artefacts and corresponding preprocessing methods. For cross-referencing of papers' IDs referred to as S1, S2, etc., please refer to Appendix A

Data type	Behavioural aspects	Artefact type	Preprocessing methods	Papers
EEG	• CLI • ER • AD • PM	• Eye movements • Muscle artefacts • Cardiac artefacts • 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	• CLI • ER • AD	• Electrode motion • Powerline interference	• Filtering • WT	S1, S2, S6, S8, S13, S16, S19, S23, S24, S26
GSR	• ER • AD • CLI	• Motion artefacts • Electrode artefacts	• Filtering	S2, S6, S8, S13, S16, S23, S24
Resp.	• CLI • AD	• Baseline drift • Motion artefacts	• Filtering	S2, S8, S16, S19, S23, S24
Eye-Tracking	• PM • CLI • AD	• Blink artefacts • Saccadic artefacts	• Filtering • Interpolation	S3, S19, S27, S36, S40, S42,

				S45, S46, S47
Misc.	• CLI • AD • ER	• Missing data • Outliers	• Data Smoothing • Interpolation	S1, S8, S13, S15, S23, S24
Artefact Subspace Reconstruction (ASR), Wavelet Transform (WT), Principal Component Analysis (PCA), Independent Component Analysis (ICA), Miscellaneous (Misc.), Cognitive Load Indicators (CLI), Emotional Responses (ER), Attention Dynamics (AD), Performance Metrics (PM)				

Complementing this, Table 2-2 offers a more nuanced examination of the features extracted from these psychophysiological data types. Within the domain of EEG, features such as power spectral density (PSD) and wavelet coefficients are frequently extracted, often employing Fourier and wavelet transforms. 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 (CSP). Furthermore, statistical tests and information-theoretic measures such as PCA, Analysis of Variance (ANOVA), Multivariate Analysis of Variance (MANOVA), Friedman tests, and mutual information coefficient were not uncommon. ML and DL methods also appeared as tools not only for classification but for feature extraction and selection as well.

Table 2-2 Summary of features extracted and extraction methods. For cross-referencing of papers' IDs referred to as S1, S2, etc., please refer to Appendix A

Data type	Extracted features	Extraction and selection methods	Papers
EEG	• Frequency bands	• FFT • WT	S2, S3, S4, S5, S6, S8, S14,

	• ERPs • PSD • Statistical features	• PCA • ICA • HHT • LSFT • WPD	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 • Friedman test • WT • 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), Mutual Information Coefficient (MIC), Power Spectral Density (PSD), Standard Deviation Abdominal (SDAbd), Skin Conductance Level (SCL), Skin Conductance Response (SCR), Standard Deviation Thoracic (SDThor), Wavelet Packet Decomposition (WPD)

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.

2.5.3 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 behaviour. Figure 2-6 shows a more nuanced analysis of the model’s types employed to identify the pilot's behaviour. A compelling trend that demands attention is the preeminent use of ML models, which constitute a significant 65% of the total models utilised. 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.

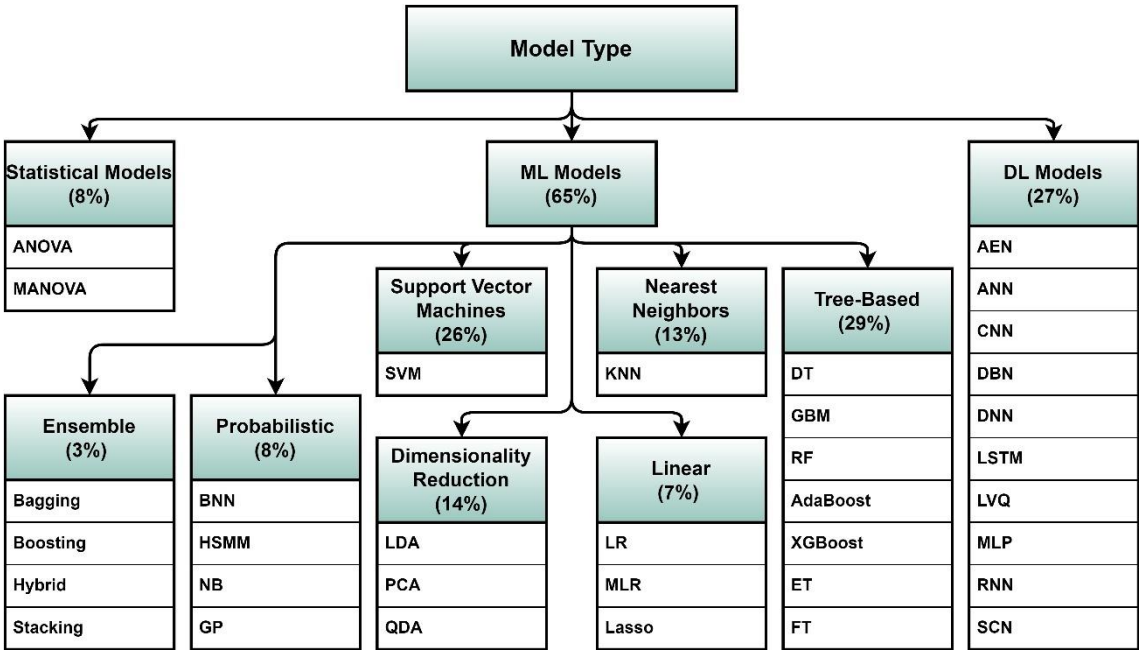

Figure 2-6 The model's types employed for identifying pilot's behaviour

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 Neighbours (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)

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 favoured 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.

In the sphere of DL models, the existing studies demonstrate a varied array of architectures that showcase the field's dynamic nature. DL techniques, contributing to 27% of the models applied, underscore their increasing relevance in this research area. Among these, traditional neural networks like ANN have been foundational, while architectures such as CNN and LSTM are distinguished for their ability to learn complex patterns in spatial and time-series data, respectively. LSTM models, in particular, are adept at addressing the challenge of temporal dependency in data, which is critical for behavioural analysis over time. Additionally, while less common, the inclusion of DBN and RNN in some studies points to a broadening of the methodological toolkit, allowing researchers to experiment with a range of neural network structures to find the most efficacious approaches for the task at hand. This diversity, including even more specialised architectures like SCN and LVQ, signifies a robust, evolving discipline that is constantly integrating new advancements to refine the analysis of psychophysiological data.

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.

On the metric front, as shown in Table 2-3, 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 specialised contexts. Notably, some studies adopt a multi-metric approach, indicating a comprehensive methodology for performance evaluation.

Table 2-3 The metrics used to evaluate the models' performance. For cross-referencing of papers' IDs referred to as S1, S2, etc., please refer to Appendix A

Performance Metric	Study ID
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)	

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.

2.5.4 Comparative Performance of Machine Learning and Deep Learning Models in Predicting Pilot Behaviour

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 visualised in a bar chart, denoted as Figure 2-7. Bear in mind that the results must be contextualised within the broader landscape of diverse experimental designs and behavioural aspects. The figure aggregates performance accuracies from a corpus of studies, which are not necessarily derived from identical experiments or behavioural metrics.

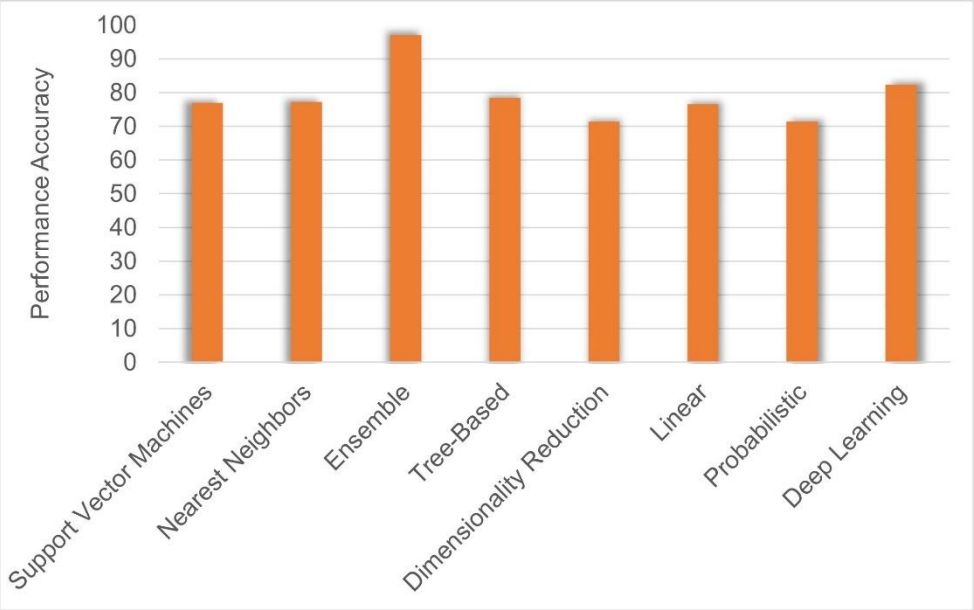

Figure 2-7 The performance accuracy of the models utilised in the literature

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 generalisability and robustness against overfitting.

Closely following Ensemble models, tree-based models exhibit an average performance rate of 78%. As illustrated in Figure 2-8, 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.

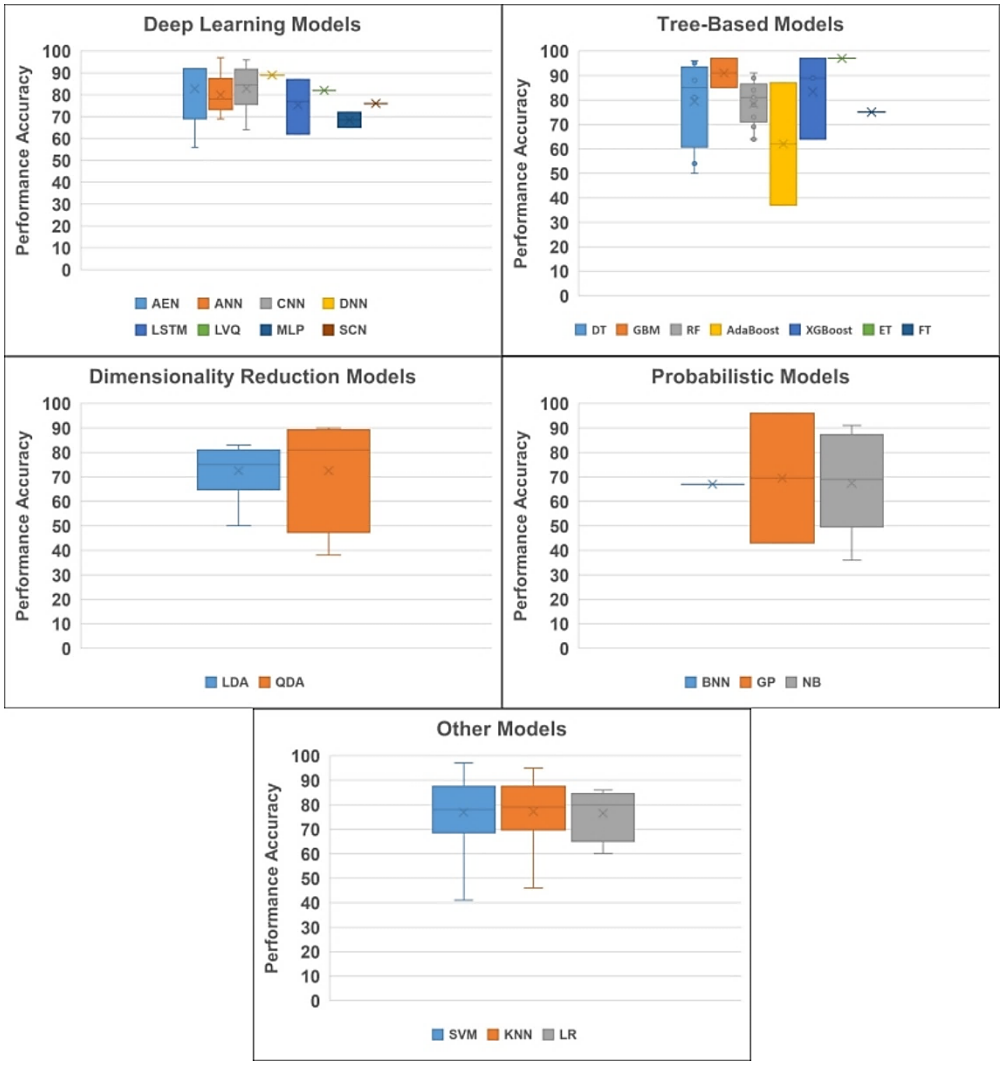

Figure 2-8 A box plot for each model type category

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 NN. 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.

By synthesising these average performance rates along with the detailed box plot statistics, this subsection furnishes an empirically substantiated framework for gauging the relative efficacy of various model categories. These insights do not merely serve as a performance benchmark but hold significant implications for future research directions. The data suggest avenues for methodological innovation and optimisation, thereby informing and guiding future research initiatives in the quest for more accurate and robust models for analysing pilot behaviour.

2.6 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 behaviour. It further contextualises these findings within the broader academic discourse and identifies both methodological limitations and avenues for future research.

2.6.1 Evaluation of Research Focus on Pilot's Behavioural and Cognitive States (RQ1)

The analysis encapsulated in subsection 2.5.1 offers a nuanced perspective on the existing body of research surrounding pilot behaviour. 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 'Channelised 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.

In considering the methodological underpinnings of the existing literature, we encounter two predominant approaches: multi-level and binary classifications. Multi-level classifications, commonly applied to analyse complex behavioural aspects like 'Workload,' offer a nuanced understanding but are challenged by issues of comparability and standardisation. This difficulty primarily arises from the absence of universally accepted metrics to define levels such as 'low,' 'medium,' or 'high.' Different studies may adopt varied criteria or thresholds for these classifications, leading to inconsistencies that can obscure the collective insights drawn from the literature. To address these challenges, future research could focus on developing uniform measurement scales and criteria, coupled with greater transparency in their application across studies. In contrast, binary classifications, often used for attributes like 'Fatigue,' provide a clear and interpretable framework. However, this method may not fully capture the continuum of behavioural states pilots experience, potentially resulting in a partial or skewed representation. Recognising these limitations, it is imperative to balance the simplicity of binary classifications with the detailed insights offered by multi-level classifications, striving for a methodology that encapsulates the full spectrum of pilot behaviour.

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 behavioural 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.

2.6.2 Interpreting Methodological Paradigms in Pilot Behaviour Research (RQ2)

The present analysis of existing studies offers a comprehensive perspective on the intricate methodologies adopted in the domain of pilot behaviour 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 standardisation, 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 behaviour, 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 behaviour, 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 artefacts. 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 standardisation and comparability of research outcomes. In the feature extraction stage, the landscape is equally diverse. While Fourier and wavelet transforms 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 behaviour from multiple psychophysiological perspectives. On the other hand, the lack of methodological standardisation hampers cross-study comparisons and meta-analyses, an issue that warrants attention in future research. The existing literature on pilot behaviour 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.

2.6.3 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 utilising psychophysiological data to identify pilot behaviour. One of the most salient aspects is the predominant deployment of ML models, which constitute 65% of the total models utilised. 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 standardisation 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 standardisation and comparative evaluation.

2.6.4 Interpretative Analysis Based on Model Performance (RQ4)

The comparative performance analysis presented in subsection 2.5.4 aggregates data from a broad array of studies, each with its unique experimental design and behavioural aspect focus. It offers rich insights into the relative performance of various ML and DL models in the domain of pilot behaviour prediction. It is important to recognize that these studies are not directly comparable due to variations in methodologies, pilot tasks, and psychophysiological measures. Consequently, the performance accuracies depicted should be understood as illustrative rather than definitive, offering a

general perspective on model effectiveness across a heterogeneous set of conditions. This aggregation serves to highlight potential trends and points of interest in the landscape of predictive modelling for pilot behaviour, rather than to establish a benchmark for model performance. The high average accuracy reported for Ensemble models, for instance, invites further investigation into the configurations of base learners and their synergistic effect under varied experimental settings. Similarly, the spread of performance across tree-based and DL models signals the need for more granular analysis to discern the influence of specific data features, model architectures, and task types on model effectiveness.

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 synthesise insights from multiple weak learners, thereby enhancing their generalisability 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 specialised nature of these architectures, optimised for specific kinds of data or tasks within the broader realm of pilot behaviour 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 behaviour. 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 behaviour analysis.

2.6.5 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 generalisable models.

Another limitation is the reliance on traditional preprocessing techniques. The complexity of psychophysiological data, fraught with various artefacts, calls for the exploration of advanced preprocessing methods. Incorporating such methods could potentially lead to more accurate and reliable models for understanding pilot behaviour, 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 generalisability 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 utilisation of 1D-CNN appears to be underexplored in the context of psychophysiological data analysis for pilot behaviour. 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 rigour in future studies.

The literature's focus on data from specific environmental settings constrains the generalisability of the findings. Future studies could benefit from collecting and analysing data from different environmental contexts, thereby enhancing the ecological validity of the research and providing a more comprehensive understanding of pilot behaviour 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.

2.7 Conclusion

This systematic literature review endeavours 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 behaviour 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 underutilisation of other types of data like ECG, GSR, and eye-tracking metrics.

Significantly, the review has identified a substantial gap in the behavioural 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 behaviour in a more holistic manner. The current methodological approaches often categorise these aspects into broader categories, thereby potentially missing nuanced interrelations between different behavioural and cognitive states. Therefore, a more balanced academic inquiry into these areas is warranted for a more comprehensive understanding of pilot behaviour.

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 artefacts, 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 behaviour through the lens of ML and psychophysiological data, there is ample room for methodological refinement and exploration. Addressing the identified gaps and under-researched areas will not only elevate the scientific rigour but also contribute to more nuanced, comprehensive, and practically applicable insights into pilot behaviour. 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.

2.8 Appendices

2.8.1 Appendix 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 categorising the studies. An initial enumeration of the studies is presented in Table 2-4, 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.

Table 2-4 List of the qualified studies

Study ID	Paper	Title
S1	(Nittala et al., 2018)	Pilot Skill Level and Workload Prediction for Sliding-Scale Autonomy
S2	(Han et al., 2020)	Classification of pilots' mental states using a multimodal deep learning network
S3	(Ziegler et al., 2016)	Sensing and Assessing Cognitive Workload Across Multiple Tasks
S4	(Dehais et al., 2019)	Monitoring Pilot's Mental Workload Using ERPs and Spectral Power with a Six-Dry-Electrode EEG System in Real Flight Conditions
S5	(Binias et al., 2018)	A Machine Learning Approach to the Detection of Pilot's Reaction to Unexpected Events Based on EEG Signals

S6	(Cesar Cavalcanti Roza & Adrian Postolache, 2019)	Multimodal Approach for Emotion Recognition Based on Simulated Flight Experiments
S7	(Thomas et al., 2015)	Fatigue detection in commercial flight operations: Results using physiological measures
S8	(Harrivel et al., 2017)	Prediction of Cognitive States during Flight Simulation using Multimodal Psychophysiological Sensing
S9	(Johnson et al., 2015)	Probe-Independent EEG Assessment of Mental Workload in Pilots
S10	(Oh et al., 2015)	A Composite Cognitive Workload Assessment System in Pilots Under Various Task Demands Using Ensemble Learning
S11	(Hannula et al., 2008)	Comparison between artificial neural network and multilinear regression models in an evaluation of cognitive workload in a flight simulator
S12	(Harrivel et al., 2016)	Psychophysiological Sensing and State Classification for Attention Management in Commercial Aviation
S13	(Roza et al., 2019)	Emotions Assessment on Simulated Flights
S14	(Wu et al., 2019)	Pilots' Fatigue Status Recognition Using Deep Contractive Autoencoder Network
S15	(Jaquess et al., 2017)	Empirical evidence for the relationship between cognitive workload and attentional reserve

S16	(Besson et al., 2013)	Effectiveness of Physiological and Psychological Features to Estimate Helicopter Pilots' Workload: A Bayesian Network Approach
S17	(Blanco et al., 2018)	Quantifying cognitive workload in simulated flight using passive, dry EEG measurements
S18	(Yingxue & Qi, 2019)	Pilots' Brain Cognitive State Inference Based on Remaining Life HSMM
S19	(Bargiotas et al., 2019)	The Complementary Role of Activity Context in the Mental Workload Evaluation of Helicopter Pilots: A Multi-tasking Learning Approach
S20	(Kacer et al., 2018)	Measurement and modelling of the behaviour of military pilots
S21	(Cai et al., 2016)	Cognitive state recognition using wavelet singular entropy and ARMA entropy with AFPA optimized GP classification
S22	(Masse et al., 2022)	Classification of Electrophysiological Signatures With Explainable Artificial Intelligence: The Case of Alarm Detection in Flight Simulator
S23	(Li, Li, Wang, Chen, & Wen, 2022)	Pilot Behaviour Recognition Based on Multi-Modality Fusion Technology Using Physiological Characteristics
S24	(Ding et al., 2020)	Measurement and identification of mental workload during simulated computer tasks with multimodal methods and machine learning

S25	(Jiang et al., 2023)	EEG-based analysis for pilots' at-risk cognitive competency identification using RF-CNN algorithm
S26	(Pan et al., 2021)	Identification of Pilots' Fatigue Status Based on Electrocardiogram Signals
S27	(Scannella et al., 2018)	Assessment of Ocular and Physiological Metrics to Discriminate Flight Phases in Real Light Aircraft
S28	(Taheri Gorji et al., 2023)	Using machine learning methods and EEG to discriminate aircraft pilot cognitive workload during flight
S29	(Friedman et al., 2019)	EEG-Based Prediction of Cognitive Load in Intelligence Tests
S30	(E. Q. Wu et al., 2021)	Detecting Fatigue Status of Pilots Based on Deep Learning Network Using EEG Signals
S31	(Zhu et al., 2023)	Recognition of Pilot Mental workload in the Simulation Operation of Carrier-based Aircraft Using the Portable EEG
S32	(Yiu et al., 2022)	Towards safe and collaborative aerodrome operations: Assessing shared situational awareness for adverse weather detection with EEG-enabled Bayesian neural networks
S33	(Yang et al., 2019)	Assessing cognitive mental workload via EEG signals and an ensemble deep learning classifier based on denoising autoencoders

S34	(Wang et al., 2020)	Cognitive Load Identification of Pilots Based on Physiological-Psychological Characteristics in Complex Environments
S35	(Snider et al., 2022)	Predicting hypoxic hypoxia using machine learning and wearable sensors
S36	(Peysakhovich et al., 2022)	Classification of flight phases based on pilots' visual scanning strategies
S37	(Pang et al., 2021)	Subject-specific mental workload classification using EEG and stochastic configuration network
S38	(Mohanavelu et al., 2022)	Machine learning-based approach for identifying mental workload of pilots
S39	(Li et al., 2023)	Securing air transportation safety through identifying pilot's risky VFR flying behaviours: An EEG-based neurophysiological modelling using machine learning algorithms
S40	(Ke et al., 2023)	Pilot Selection in the Era of Virtual Reality: Algorithms for Accurate and Interpretable Machine Learning Models
S41	(Hernández-Sabaté et al., 2022)	Recognition of the Mental Workloads of Pilots in the Cockpit Using EEG Signals
S42	(Gomolka et al., 2022)	Use of a DNN in Recording and Analysis of Operator Attention in Advanced HMI Systems
S43	(Lorenz et al., 2019)	Assessing Control Devices for the Supervisory Control of Autonomous Wingmen

S44	(Chen et al., 2022)	Real-time evaluation method of flight mission load based on sensitivity analysis of physiological factors
S45	(Berthelot et al., 2019)	Self-Affinity of an Aircraft Pilot's Gaze Direction as a Marker of Visual Tunneling
S46	(Xi et al., 2020)	Predicting Student Flight Performance with Multimodal Features
S47	(Huang & Wang, 2022)	Fatigue Detection and Analysis Based on Multi-channel Data Fusion
S48	(Hajra et al., 2020)	A comparison of ECG and EEG metrics for in-flight monitoring of helicopter pilot workload
S49	(Ding et al., 2022)	A machine learning approach to reduce mental fatigue risk of pilots based on HRV data
S50	(Lee, Jeong, et al., 2023)	Autonomous System for EEG-Based Multiple Abnormal Mental States Classification Using Hybrid Deep Neural Networks Under Flight Environment
S51	(Jeong et al., 2019)	Classification of Drowsiness Levels Based on a Deep Spatio-Temporal Convolutional Bidirectional LSTM Network Using Electroencephalography Signals
S52	(Lee, Kim, & Choi, 2023)	Classification of Distraction Levels Using Hybrid Deep Neural Networks From EEG Signals

S53	(Samani et al., 2021)	Collaborative Communications Between A Human And A Resilient Safety Support System
S54	(Lee et al., 2020)	Continuous EEG Decoding of Pilots' Mental States Using Multiple Feature Block-Based Convolutional Neural Network
S55	(Murthy & Biswas, 2022)	Deep Learning-based Eye Gaze Estimation for Military Aviation
S56	(Socha et al., 2022)	Design of Wearable Eye Tracker with Automatic Cockpit Areas of Interest Recognition
S57	(Khan et al., 2017)	Detection and classification of pilots cognitive state using EEG
S58	(Zhuang et al., 2021)	EEG Based Eye Movements Multi-Classification Using Convolutional Neural Network
S59	(Samima & Sarma, 2019)	EEG-Based Mental Workload Estimation
S60	(Gateau et al., 2018)	In silico vs. Over the Clouds: On-the-Fly Mental State Estimation of Aircraft Pilots, Using a Functional Near Infrared Spectroscopy Based Passive-BCI
S61	(Benaroch et al., 2021)	Long-Term BCI Training of a Tetraplegic User: Adaptive Riemannian Classifiers and User Training

S62	(Borghini et al., 2017)	A New Perspective for the Training Assessment: Machine Learning-Based Neurometric for Augmented User's Evaluation
S63	(Qin et al., 2021)	Detection of mental fatigue state using heart rate variability and eye metrics during simulated flight
S64	(Sauvet et al., 2014)	In-Flight Automatic Detection of Vigilance States Using a Single EEG Channel
S65	(Jiang et al., 2022)	Mental Workload Artificial Intelligence Assessment of Pilots' EEG Based on Multi-Dimensional Data Fusion and LSTM with Attention Mechanism Model
S66	(Klyde et al., 2021)	A New Approach to Aircraft Handling Qualities Prediction
S67	(Lounis et al., 2021)	Visual scanning strategies in the cockpit are modulated by pilots' expertise: A flight simulator study
S68	(Shuang et al., 2017)	Recognition of fatigue status of pilots based on deep sparse auto-encoding network
S69	(Mishra et al., 2019)	Reducing Commercial Aviation Fatalities Using Support Vector Machines
S70	(Huang et al., 2022)	Modeling and analysis of fatigue detection with multi-channel data fusion
S71	(Tortora et al., 2022)	Neural correlates of user learning during long-term BCI training for the Cybathlon competition

S72	(Wu et al., 2022)	Self-Paced Dynamic Infinite Mixture Model for Fatigue Evaluation of Pilots' Brains
S73	(Kamrud et al., 2021)	Generalized Deep Learning EEG Models for Cross-Participant and Cross-Task Detection of the Vigilance Decrement in Sustained Attention Tasks
S74	(Li, Li, Wang, Li, et al., 2022)	Towards Safer Flights: A Multi-modality Fusion Technology-based Cognitive Load Recognition Framework
S75	(Wang et al., 2023)	A Method for Classification and Evaluation of Pilot's Mental States Based on CNN
S76	(Binias et al., 2020)	Prediction of Pilot's Reaction Time Based on EEG Signals
S77	(I. Alreshidi, I. Moulitsas, & B. Bisandu, 2023)	Illuminating the Neural Landscape of Pilot Mental States: A Convolutional Neural Network Approach with SHAP Interpretability
S78	(I. Alreshidi, I. Moulitsas, & K. W. Jenkins, 2023)	Multimodal Approach for Pilot Mental State Detection Based on EEG
S79	(Ibrahim Alreshidi et al., 2023)	A Comprehensive Analysis of Machine Learning and Deep Learning Models for Identifying Pilots' Mental States from Imbalanced Physiological Data
S80	(Alreshidi et al., 2022)	Miscellaneous EEG Preprocessing and Machine Learning for Pilots' Mental States Classification: Implications

Further granularity is achieved through Table 2-5, which organises the selected studies by their respective publication venues and types. This table not only

informs about the number of studies affiliated with specific journals and conference proceedings but also provides a broader view of the academic platforms that prominently feature in this domain. The information presented in Table 2-5 is pivotal for understanding the disciplinary reach and the diverse channels of scholarly communication in this field.

Table 2-5 Studies distribution across publication venues and types

Publication venue	Type	Number of studies
Sensors	Journal	6
Biomedical Signal Processing and Control	Journal	3
Frontiers in Human Neuroscience	Journal	3
Applied Sciences	Journal	2
Computers in Biology and Medicine	Journal	2
Frontiers in Neuroinformatics	Journal	2
Frontiers in Neuroscience	Journal	2
IEEE Transactions on Cognitive and Developmental Systems	Journal	2
Advanced Engineering Informatics	Journal	1
Aerospace	Journal	1
Biocybernetics and Biomedical Engineering	Journal	1
Biosensors	Journal	1
Brain Sciences	Journal	1
Chinese Journal of Aeronautics	Journal	1
Computational Intelligence and Neuroscience	Journal	1

Computer Systems Science and Engineering	Journal	1
Ergonomics	Journal	1
Human Factors	Journal	1
Human Factors and Ergonomics in Manufacturing & Service Industries	Journal	1
IEEE Access	Journal	1
IEEE Transactions on Biomedical Engineering	Journal	1
Reliability Engineering and System Safety	Journal	1
IEEE Transactions on Cybernetics	Journal	1
IEEE Transactions on Instrumentation and Measurement	Journal	1
IEEE Transactions on Intelligent Transportation Systems	Journal	1
IEEE Transactions on Systems, Man, and Cybernetics: Systems	Journal	1
International Journal of Pattern Recognition and Artificial Intelligence	Journal	1
International Journal of Psychophysiology	Journal	1
Journal of Advanced Transportation	Journal	1
Journal of NeuroEngineering and Rehabilitation	Journal	1
Neurocomputing	Journal	1
PLoS One	Journal	1
Reliability Engineering and System Safety	Journal	1

Scientific Reports - Nature	Journal	1
The International Journal of Advanced Manufacturing Technology	Journal	1
AIAA Scitech Forum	Conference	3
AIAA AVIATION Forum	Conference	1
Chinese Control Conference	Conference	2
Foundations of Augmented Cognition	Conference	2
Annual International IEEE EMBS Conference on Neural Engineering	Conference	1
Chinese Control And Decision Conference	Conference	1
Cyber Security Intelligence and Analytics	Conference	1
Eye Tracking Research and Applications	Conference	1
Human Mental Workload: Models and Applications	Conference	1
IEEE 4th International Conference on Civil Aviation Safety and Information Technology	Conference	1
IEEE Aerospace Conference	Conference	1
IEEE Engineering in Medicine and Biology Society	Conference	1
IEEE International Conference on Computational Intelligence and Applications	Conference	1
IEEE International Conference on Machine Learning and Applications	Conference	1
IEEE International Conference on Systems, Man, and Cybernetics	Conference	1

IEEE International Symposium on Medical Measurements and Applications	Conference	1
International Conference on Applied Human Factors and Ergonomics	Conference	1
International Conference on Advances in Artificial Intelligence	Conference	1
International Conference on Autonomous Systems	Conference	1
International Conference on Human Machine Interaction	Conference	1
International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering	Conference	1
International Conference on Smart Systems and Inventive Technology	Conference	1
International Winter Conference on Brain-Computer Interface	Conference	1
Modelling and Simulation for Autonomous Systems	Conference	1
New Trends in Civil Aviation	Conference	1
SAE Technical Paper Series	Conference	1
Social, Cultural, and Behavioral Modeling	Conference	1
Systems and Information Engineering Design	Conference	1

In addition to tabulated data, Figure 2-9 offers a temporal mapping of the studies, illustrating the number of publications per year. Upon examination of Figure 2-9, it is evident that there has been a notable surge in the number of studies published from 2015 onwards, signalling an increased research focus on the subject matter. This uptick may be ascribed to a confluence of factors.

Notably, around this period, there were significant technological breakthroughs such as the rise of big data analytics and ML algorithms, which have revolutionized data processing capabilities. Concurrently, the advent of affordable and sophisticated sensor technology has facilitated more intricate psychophysiological studies. Policy changes, such as the Open Access movement, have likely played a role, broadening the dissemination of scientific knowledge and fostering a more collaborative research environment. Furthermore, global events and the increasing complexity of human-machine interactions may have steered research priorities towards understanding cognitive and behavioural dynamics in high-stakes settings such as aviation. Specific to the domain of aviation safety, the introduction of new regulations and the industry's heightened commitment to mitigating human factors in accidents could have spurred more extensive investigations. The data may also reflect a response to identified gaps in research, with academic and industry partners seeking to develop predictive models that can better ensure pilot performance and flight safety.

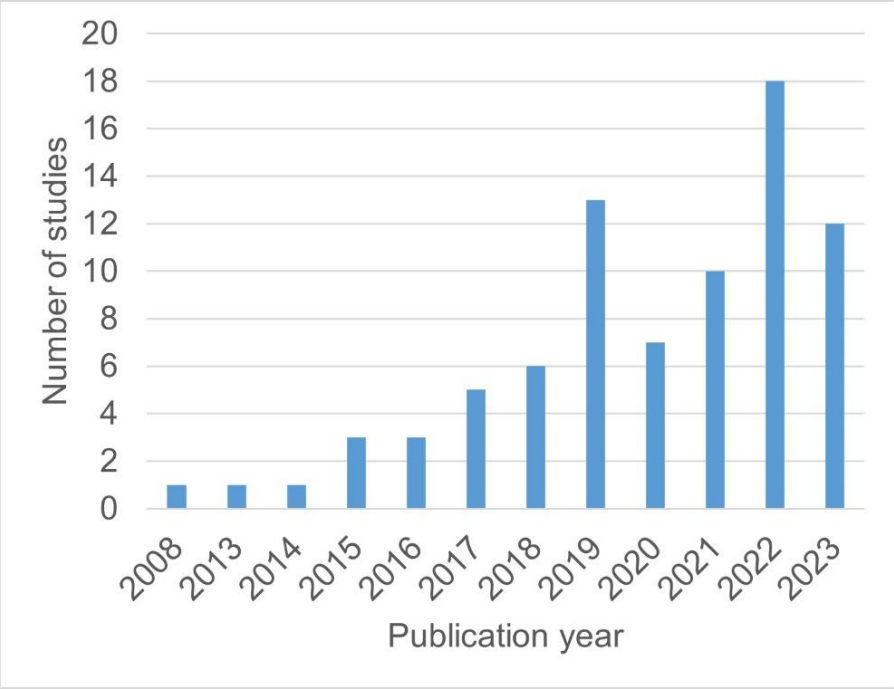

Figure 2-9 Study publication distribution using a yearly calendar.

REFERENCES

Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., & Adeli, H. (2018). Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. *Comput Biol Med*, *100*, 270-278. https://doi.org/10.1016/j.compbiomed.2017.09.017

Ahn, S., Nguyen, T., Jang, H., Kim, J. G., & Jun, S. C. (2016). Exploring Neuro-Physiological Correlates of Drivers' Mental Fatigue Caused by Sleep Deprivation Using Simultaneous EEG, ECG, and fNIRS Data. *Front Hum Neurosci*, *10*, 219. https://doi.org/10.3389/fnhum.2016.00219

Alreshidi, I., Moulitsas, I., & Bisandu, B. (2023). Illuminating the Neural Landscape of Pilot Mental States: A Convolutional Neural Network Approach with SHAP Interpretability. *Sensors (Basel)*.

Alreshidi, I., Moulitsas, I., & Jenkins, K. W. (2023). Multimodal Approach for Pilot Mental State Detection Based on EEG. *Sensors (Basel)*, *23*(17). https://doi.org/10.3390/s23177350

Alreshidi, I., Yadav, S., Moulitsas, I., & Jenkins, K. (2023). A Comprehensive Analysis of Machine Learning and Deep Learning Models for Identifying Pilots' Mental States from Imbalanced Physiological Data. AIAA AVIATION 2023 Forum,

Alreshidi, I. M., Moulitsas, I., & Jenkins, K. W. (2022). Miscellaneous EEG Preprocessing and Machine Learning for Pilots' Mental States Classification: Implications. 2022 The 6th International Conference on Advances in Artificial Intelligence,

Aurino, D. E. M. (2000). Human factors and aviation safety: What the industry has, what the industry needs. *Ergonomics*, *43*(7), 952-959. https://doi.org/10.1080/001401300409134

Bargiotas, I., Nicolaï, A., Vidal, P.-P., Labourdette, C., Vayatis, N., & Buffat, S. (2019). 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* (pp. 222-238). https://doi.org/10.1007/978-3-030-14273-5_13

Bashivan, P., Rish, I., Yeasin, M., & Codella, N. (2016). Learning representations from EEG with deep recurrent-convolutional neural networks. 4th International Conference on Learning Representations (ICLR),

Behrend, J., & Dehais, F. (2020). How role assignment impacts decision-making in high-risk environments: Evidence from eye-tracking in aviation. *Safety Science*, *127*, 104738. https://doi.org/10.1016/j.ssci.2020.104738

Benaroch, C., Sadatnejad, K., Roc, A., Appriou, A., Monseigne, T., Pramij, S., Mladenovic, J., Pillette, L., Jeunet, C., & Lotte, F. (2021). Long-Term BCI Training of a Tetraplegic User: Adaptive Riemannian Classifiers and User

Training. *Frontiers in Human Neuroscience*, *15*. https://doi.org/10.3389/fnhum.2021.635653

Berthelot, B., Mazoyer, P., Egea, S., André, J.-M., Grivel, É., & Legrand, P. (2019). Self-Affinity of an Aircraft Pilot's Gaze Direction as a Marker of Visual Tunneling. SAE Technical Paper Series,

Besson, P., Bourdin, C., Bringoux, L., Dousset, E., Maiano, C., Marqueste, T., Mestre, D. R., Gaetan, S., Baudry, J. P., & Vercher, J. L. (2013). Effectiveness of Physiological and Psychological Features to Estimate Helicopter Pilots' Workload: A Bayesian Network Approach. *Ieee Transactions on Intelligent Transportation Systems*, *14*(4), 1872-1881. https://doi.org/10.1109/Tits.2013.2269679

Bhardwaj, A., Gupta, A., Jain, P., Rani, A., & Yadav, J. (2015, 2015/02//). Classification of human emotions from EEG signals using SVM and LDA Classifiers.

Binias, B., Myszor, D., & Cyran, K. A. (2018). A Machine Learning Approach to the Detection of Pilot's Reaction to Unexpected Events Based on EEG Signals. *Comput Intell Neurosci*, *2018*, 2703513. https://doi.org/10.1155/2018/2703513

Binias, B., Myszor, D., Palus, H., & Cyran, K. A. (2020). Prediction of Pilot's Reaction Time Based on EEG Signals. *Frontiers in Neuroinformatics*, *14*. https://doi.org/10.3389/fninf.2020.00006

Blanco, J. A., Johnson, M. K., Jaquess, K. J., Oh, H., Lo, L.-C., Gentili, R. J., & Hatfield, B. D. (2018). Quantifying Cognitive Workload in Simulated Flight Using Passive, Dry EEG Measurements. *IEEE Transactions on Cognitive and Developmental Systems*, *10*(2), 373-383. https://doi.org/10.1109/tcds.2016.2628702

Borghini, G., Arico, P., Di Flumeri, G., Sciaraffa, N., Colosimo, A., Herrero, M. T., Bezerianos, A., Thakor, N. V., & Babiloni, F. (2017). A New Perspective for the Training Assessment: Machine Learning-Based Neurometric for Augmented User's Evaluation. *Front Neurosci*, *11*, 325. https://doi.org/https://doi.org/10.3389/fnins.2017.00325

Boyd, D. D. (2017). A Review of General Aviation Safety (1984-2017). *Aerosp Med Hum Perform*, *88*(7), 657-664. https://doi.org/10.3357/AMHP.4862.2017

Cai, Z. X., Wu, Q., Huang, D., Ding, L., Yu, B. T., Law, R., Huang, J. Y., & Fu, S. (2016). Cognitive state recognition using wavelet singular entropy and ARMA entropy with AFPA optimized GP classification. *Neurocomputing*, *197*, 29-44. https://doi.org/10.1016/j.neucom.2016.01.054

Cecotti, H., & Graser, A. (2011). Convolutional neural networks for P300 detection with application to brain-computer interfaces. *IEEE Trans Pattern Anal Mach Intell*, *33*(3), 433-445. https://doi.org/10.1109/TPAMI.2010.125

Cesar Cavalcanti Roza, V., & Adrian Postolache, O. (2019). Multimodal Approach for Emotion Recognition Based on Simulated Flight Experiments. *Sensors (Basel)*, *19*(24). https://doi.org/10.3390/s19245516

Chaudhuri, A., & Routray, A. (2020). Driver Fatigue Detection Through Chaotic Entropy Analysis of Cortical Sources Obtained From Scalp EEG Signals. *Ieee Transactions on Intelligent Transportation Systems*, *21*(1), 185-198. https://doi.org/10.1109/Tits.2018.2890332

Chen, F., Zhou, J., Wang, Y., Yu, K., Arshad, S. Z., Khawaji, A., & Conway, D. (2016). *Robust multimodal cognitive load measurement*. Springer.

Chen, J., Xue, L., Rong, J., & Gao, X. (2022). Real-time evaluation method of flight mission load based on sensitivity analysis of physiological factors. *Chinese Journal of Aeronautics*, *35*(3), 450-463. https://doi.org/10.1016/j.cja.2021.11.010

Chung, M. C. (2017). The Psychological Impact of Aircraft Disasters. *Passenger Behaviour*.

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

Dehais, F., Dupres, A., Blum, S., Drougard, N., Scannella, S., Roy, R. N., & Lotte, F. (2019). Monitoring Pilot's Mental Workload Using ERPs and Spectral Power with a Six-Dry-Electrode EEG System in Real Flight Conditions. *Sensors (Basel)*, *19*(6). https://doi.org/10.3390/s19061324

Ding, S., Pan, X., Han, R., Zeng, X., Li, Y., & Zheng, X. (2022, 27-30 July 2022). A machine learning approach to reduce mental fatigue risk of pilots based on HRV data. 12th International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering (QR2MSE 2022),

Ding, S., Zhang, N., Xu, X., Guo, L., & Zhang, J. (2015). Deep Extreme Learning Machine and Its Application in EEG Classification. *Mathematical Problems in Engineering*, *2015*, 1-11. https://doi.org/10.1155/2015/129021

Ding, Y., Cao, Y., Duffy, V. G., Wang, Y., & Zhang, X. (2020). Measurement and identification of mental workload during simulated computer tasks with multimodal methods and machine learning. *Ergonomics*, *63*(7), 896-908. https://doi.org/10.1080/00140139.2020.1759699

Duan, L., Wang, Z., Qiao, Y., Wang, Y., Huang, Z., & Zhang, B. (2022). An Automatic Method for Epileptic Seizure Detection Based on Deep Metric Learning. *IEEE J Biomed Health Inform*, *26*(5), 2147-2157. https://doi.org/10.1109/JBHI.2021.3138852

Friedman, N., Fekete, T., Gal, K., & Shriki, O. (2019). EEG-Based Prediction of Cognitive Load in Intelligence Tests. *Front Hum Neurosci*, *13*, 191. https://doi.org/10.3389/fnhum.2019.00191

Gao, Z., Wang, X., Yang, Y., Mu, C., Cai, Q., Dang, W., & Zuo, S. (2019). EEG-Based Spatio-Temporal Convolutional Neural Network for Driver Fatigue Evaluation. *IEEE Trans Neural Netw Learn Syst*, *30*(9), 2755-2763. https://doi.org/10.1109/TNNLS.2018.2886414

Gateau, T., Ayaz, H., & Dehais, F. (2018). 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 in Human Neuroscience*, *12*. https://doi.org/10.3389/fnhum.2018.00187

Gomolka, Z., Zeslawska, E., Twarog, B., Kordos, D., & Rzucidlo, P. (2022). Use of a DNN in Recording and Analysis of Operator Attention in Advanced HMI Systems. *Applied Sciences*, *12*(22). https://doi.org/10.3390/app122211431

Hajinoroozi, M., Mao, Z. J., Jung, T. P., Lin, C. T., & Huang, Y. F. (2016). EEG-based prediction of driver's cognitive performance by deep convolutional neural network. *Signal Processing-Image Communication*, *47*, 549-555. https://doi.org/10.1016/j.image.2016.05.018

Hajra, S. G., Xi, P., & Law, A. (2020, 11-14 Oct. 2020). A comparison of ECG and EEG metrics for in-flight monitoring of helicopter pilot workload. 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC),

Han, S. Y., Kwak, N. S., Oh, T., & Lee, S. W. (2020). Classification of pilots' mental states using a multimodal deep learning network. *Biocybernetics and Biomedical Engineering*, *40*(1), 324-336. https://doi.org/10.1016/j.bbe.2019.12.002

Hannula, M., Huttunen, K., Koskelo, J., Laitinen, T., & Leino, T. (2008). Comparison between artificial neural network and multilinear regression models in an evaluation of cognitive workload in a flight simulator. *Comput Biol Med*, *38*(11-12), 1163-1170. https://doi.org/10.1016/j.compbiomed.2008.09.007

Harrivel, A. R., Liles, C., Stephens, C. L., Ellis, K. K., Prinzel, L. J., & Pope, A. T. (2016). Psychophysiological Sensing and State Classification for Attention Management in Commercial Aviation. AIAA Infotech @ Aerospace,

Harrivel, A. R., Stephens, C. L., Milletich, R. J., Heinich, C. M., Last, M. C., Napoli, N. J., Abraham, N., Prinzel, L. J., Motter, M. A., & Pope, A. T. (2017). Prediction of Cognitive States during Flight Simulation using Multimodal Psychophysiological Sensing. AIAA Information Systems-AIAA Infotech @ Aerospace,

Henderson, I. L. (2022). Aviation safety regulations for unmanned aircraft operations: Perspectives from users. *Transport Policy*, *125*, 192-206.

Hernández-Sabaté, A., Yauri, J., Folch, P., Piera, M. À., & Gil, D. (2022). Recognition of the Mental Workloads of Pilots in the Cockpit Using EEG Signals. *Applied Sciences*, *12*(5). https://doi.org/10.3390/app12052298

Hogervorst, M. A., Brouwer, A. M., & van Erp, J. B. (2014). Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. *Front Neurosci*, *8*, 322. https://doi.org/10.3389/fnins.2014.00322

Huang, W., & Wang, C. (2022). Fatigue Detection and Analysis Based on Multi-channel Data Fusion. In *Cyber Security Intelligence and Analytics* (pp. 650-656). https://doi.org/10.1007/978-3-030-97874-7_85

Huang, W., Wang, C., Jia, H.-b., Xue, P., & Wang, L. (2022). Modeling and analysis of fatigue detection with multi-channel data fusion. *The International Journal of Advanced Manufacturing Technology*, *122*(1), 291-301. https://doi.org/10.1007/s00170-022-09364-0

Ian Goodfellow, Y. B. a. A. C. (2016). *Deep Learning*. The MIT Press.

Jap, B. T., Lal, S., Fischer, P., & Bekiaris, E. (2009). Using EEG spectral components to assess algorithms for detecting fatigue. *Expert Systems with Applications*, *36*(2), 2352-2359. https://doi.org/10.1016/j.eswa.2007.12.043

Jaquess, K. J., Gentili, R. J., Lo, L. C., Oh, H., Zhang, J., Rietschel, J. C., Miller, M. W., Tan, Y. Y., & Hatfield, B. D. (2017). Empirical evidence for the relationship between cognitive workload and attentional reserve. *Int J Psychophysiol*, *121*, 46-55. https://doi.org/10.1016/j.ijpsycho.2017.09.007

Jeong, J. H., Yu, B. W., Lee, D. H., & Lee, S. W. (2019). Classification of Drowsiness Levels Based on a Deep Spatio-Temporal Convolutional Bidirectional LSTM Network Using Electroencephalography Signals. *Brain Sci*, *9*(12). https://doi.org/10.3390/brainsci9120348

Jiang, G., Chen, H., Wang, C., & Xue, P. (2022). Mental Workload Artificial Intelligence Assessment of Pilots' EEG Based on Multi-Dimensional Data Fusion and LSTM with Attention Mechanism Model. *International Journal of Pattern Recognition and Artificial Intelligence*, *36*(11), 2259035. https://doi.org/10.1142/S0218001422590352

Jiang, S., Chen, W., Ren, Z., & Zhu, H. (2023). EEG-based analysis for pilots' at-risk cognitive competency identification using RF-CNN algorithm. *Front Neurosci*, *17*, 1172103. https://doi.org/10.3389/fnins.2023.1172103

Jiao, Z. C., Gao, X. B., Wang, Y., Li, J., & Xu, H. J. (2018). Deep Convolutional Neural Networks for mental load classification based on EEG data. *Pattern Recognition*, *76*, 582-595. https://doi.org/10.1016/j.patcog.2017.12.002

Johnson, M. K., Blanco, J. A., Gentili, R. J., Jaquess, K. J., Oh, H., & Hatfield, B. D. (2015). Probe-Independent EEG Assessment of Mental Workload in Pilots. 7th Annual International IEEE EMBS Conference on Neural Engineering,

K, M., S, P., K, A., D, R., Chinnadurai, V., S, V., K, R., & Jayaraman, S. (2020). Dynamic cognitive workload assessment for fighter pilots in simulated

fighter aircraft environment using EEG. *Biomedical Signal Processing and Control*, *61*, 102018. https://doi.org/https://doi.org/10.1016/j.bspc.2020.102018

Kacer, J., Kutilek, P., Krivanek, V., Doskocil, R., Smrcka, P., & Krupka, Z. (2018). Measurement and Modelling of the Behavior of Military Pilots. Modelling and Simulation for Autonomous Systems, Cham.

Kamrud, A., Borghetti, B., Schubert Kabban, C., & Miller, M. (2021). Generalized Deep Learning EEG Models for Cross-Participant and Cross-Task Detection of the Vigilance Decrement in Sustained Attention Tasks. *Sensors (Basel)*, *21*(16). https://doi.org/10.3390/s21165617

Kannan, S., Premalatha, G., Jamuna Rani, M., Jayakumar, D., Senthil, P., Palanivelrajan, S., Devi, S., & Sahile, K. (2022). Effective Evaluation of Medical Images Using Artificial Intelligence Techniques. *Comput Intell Neurosci*, *2022*, 8419308. https://doi.org/10.1155/2022/8419308

Kar, S., Bhagat, M., & Routray, A. (2010). EEG signal analysis for the assessment and quantification of driver's fatigue. *Transportation Research Part F-Traffic Psychology and Behaviour*, *13*(5), 297-306. https://doi.org/10.1016/j.trf.2010.06.006

Ke, L., Zhang, G., He, J., Li, Y., Li, Y., Liu, X., & Fang, P. (2023). Pilot Selection in the Era of Virtual Reality: Algorithms for Accurate and Interpretable Machine Learning Models. *Aerospace*, *10*(5). https://doi.org/10.3390/aerospace10050394

Khan, Q. A., Hassan, A., Rehman, S., & Riaz, F. (2017, 8-11 Sept. 2017). Detection and classification of pilots cognitive state using EEG. 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA),

Kitchenham, B. (2007). Guidelines for performing Systematic Literature Reviews in Software Engineering. *Technical report EBSE-2007-001, UK*.

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering - A systematic literature review. *Information and Software Technology*, *51*(1), 7-15. https://doi.org/10.1016/j.infsof.2008.09.009

Klyde, D. H., Lampton, A. K., Mitchell, D. G., Berka, C., & Rhinehart, M. (2021). A New Approach to Aircraft Handling Qualities Prediction. In *AIAA Scitech 2021 Forum*. American Institute of Aeronautics and Astronautics. https://doi.org/10.2514/6.2021-0178

Koelstra, S., Muhl, C., Soleymani, M., Jong-Seok, L., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., & Patras, I. (2012). DEAP: A Database for Emotion Analysis ;Using Physiological Signals. *IEEE Transactions on Affective Computing*, *3*(1), 18-31. https://doi.org/10.1109/t-affc.2011.15

Lal, S. K., Craig, A., Boord, P., Kirkup, L., & Nguyen, H. (2003). Development of an algorithm for an EEG-based driver fatigue countermeasure. *J Safety Res*, *34*(3), 321-328. https://doi.org/10.1016/s0022-4375(03)00027-6

Lee, D. H., Jeong, J. H., Kim, K., Yu, B. W., & Lee, S. W. (2020). Continuous EEG Decoding of Pilots' Mental States Using Multiple Feature Block-Based Convolutional Neural Network. *IEEE Access*, *8*, 121929-121941. https://doi.org/10.1109/ACCESS.2020.3006907

Lee, D. H., Jeong, J. H., Yu, B. W., Kam, T. E., & Lee, S. W. (2023). Autonomous System for EEG-Based Multiple Abnormal Mental States Classification Using Hybrid Deep Neural Networks Under Flight Environment. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *53*(10), 6426-6437. https://doi.org/10.1109/TSMC.2023.3282635

Lee, D. H., Kim, S. J., & Choi, Y. W. (2023, 20-22 Feb. 2023). Classification of Distraction Levels Using Hybrid Deep Neural Networks From EEG Signals. 2023 11th International Winter Conference on Brain-Computer Interface (BCI),

Li, G., & Baker, S. P. (2007). Crash Risk in General Aviation. *Jama*, *297*(14), 1596-1598. https://doi.org/10.1001/jama.297.14.1596

Li, G., Baker, S. P., Grabowski, J. G., & Rebok, G. W. (2001). Factors associated with pilot error in aviation crashes. *Aviation Space and Environmental Medicine*, *72*(1), 52-58.

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

Li, Y., Li, K., Wang, S., Chen, X., & Wen, D. (2022). Pilot Behaviour Recognition Based on Multi-Modality Fusion Technology Using Physiological Characteristics. *Biosensors (Basel)*, *12*(6). https://doi.org/10.3390/bios12060404

Li, Y., Li, K., Wang, S., Li, Y., Chen, J., & Wen, D. (2022). Towards Safer Flights: A Multi-modality Fusion Technology-based Cognitive Load Recognition Framework. 2022 IEEE 4th International Conference on Civil Aviation Safety and Information Technology (ICCASIT),

Liu, J., & Woodson, B. (2019). *Deep Learning Classification for Epilepsy Detection Using a Single Channel Electroencephalography (EEG)* Proceedings of the 2019 3rd International Conference on Deep Learning Technologies,

Liu, Y., Ayaz, H., & Shewokis, P. A. (2017). Multisubject "Learning" for Mental Workload Classification Using Concurrent EEG, fNIRS, and Physiological Measures. *Front Hum Neurosci*, *11*, 389. https://doi.org/10.3389/fnhum.2017.00389

Liu, Y., Huang, Y.-X., Zhang, X., Qi, W., Guo, J., Hu, Y., Zhang, L., & Su, H. (2020). Deep C-LSTM Neural Network for Epileptic Seizure and Tumor Detection Using High-Dimension EEG Signals. *IEEE Access*, *8*, 37495-37504. https://doi.org/10.1109/access.2020.2976156

Lorenz, G. T., Ehrenstrom, J. S., Ullmann, T. B., Palmer, R. C., Tenhundfeld, N. L., Visser, E. J. d., Donadio, B. T., & Tossell, C. C. (2019). Assessing Control Devices for the Supervisory Control of Autonomous Wingmen. 2019 Systems and Information Engineering Design Symposium (SIEDS),

Lounis, C., Peysakhovich, V., & Causse, M. (2021). Visual scanning strategies in the cockpit are modulated by pilots' expertise: A flight simulator study. *PLoS One*, *16*(2), e0247061. https://doi.org/10.1371/journal.pone.0247061

Marinescu, A. C., Sharples, S., Ritchie, A. C., Sanchez Lopez, T., McDowell, M., & Morvan, H. P. (2018). Physiological parameter response to variation of mental workload. *Human Factors*, *60*(1), 31-56.

Masse, E., Bartheye, O., & Fabre, L. (2022). Classification of Electrophysiological Signatures With Explainable Artificial Intelligence: The Case of Alarm Detection in Flight Simulator. *Front Neuroinform*, *16*, 904301. https://doi.org/10.3389/fninf.2022.904301

Mishra, A., Shrivastava, K. K., A. A, B., & Quadir, N. A. (2019). Reducing Commercial Aviation Fatalities Using Support Vector Machines. 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT),

Mohanavelu, K., Poonguzhali, S., Janani, A., & Vinutha, S. (2022). Machine learning-based approach for identifying mental workload of pilots. *Biomedical Signal Processing and Control*, *75*. https://doi.org/10.1016/j.bspc.2022.103623

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group*, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine*, *151*(4), 264-269. https://doi.org/10.1136/bmj.b2535

Morgan Iii, C. A., Aikins, D. E., Steffian, G., Coric, V., & Southwick, S. (2007). Relation between cardiac vagal tone and performance in male military personnel exposed to high stress: Three prospective studies. *Psychophysiology*, *44*(1), 120-127. https://doi.org/https://doi.org/10.1111/j.1469-8986.2006.00475.x

Murthy, L. R. D., & Biswas, P. (2022). Deep Learning-based Eye Gaze Estimation for Military Aviation. 2022 IEEE Aerospace Conference (AERO),

Nittala, S. K. R., Elkin, C. P., Kiker, J. M., Meyer, R., Curro, J., Reiter, A. K., Xu, K. S., & Devabhaktuni, V. K. (2018). *Pilot Skill Level and Workload Prediction for Sliding-Scale Autonomy* 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA),

Nurse, E., Mashford, B. S., Yepes, A. J., Kiral-Kornek, I., Harrer, S., & Freestone, D. R. (2016). *Decoding EEG and LFP signals using deep learning* Proceedings of the ACM International Conference on Computing Frontiers,

Oh, H., Hatfield, B. D., Jaquess, K. J., Lo, L.-C., Tan, Y. Y., Prevost, M. C., Mohler, J. M., Postlethwaite, H., Rietschel, J. C., Miller, M. W., Blanco, J. A., Chen, S., & Gentili, R. J. (2015). A Composite Cognitive Workload Assessment System in Pilots Under Various Task Demands Using Ensemble Learning. Foundations of Augmented Cognition,

Oster Jr, C. V., Strong, J. S., & Zorn, C. K. (2013). Analyzing aviation safety: Problems, challenges, opportunities. *Research in transportation economics*, *43*(1), 148-164. https://doi.org/10.1016/j.retrec.2012.12.001

Pan, T., Wang, H., Si, H., Li, Y., & Shang, L. (2021). Identification of Pilots' Fatigue Status Based on Electrocardiogram Signals. *Sensors (Basel)*, *21*(9). https://doi.org/10.3390/s21093003

Pang, L., Guo, L., Zhang, J., Wanyan, X., Qu, H., & Wang, X. (2021). Subject-specific mental workload classification using EEG and stochastic configuration network (SCN). *Biomedical Signal Processing and Control*, *68*. https://doi.org/10.1016/j.bspc.2021.102711

Patel, M., Lal, S. K. L., Kavanagh, D., & Rossiter, P. (2011). Applying neural network analysis on heart rate variability data to assess driver fatigue. *Expert Systems with Applications*, *38*(6), 7235-7242. https://doi.org/10.1016/j.eswa.2010.12.028

Peysakhovich, V., Ledegang, W., Houben, M., & Groen, E. (2022). *Classification of flight phases based on pilots' visual scanning strategies* 2022 Symposium on Eye Tracking Research and Applications,

Qin, H., Zhou, X., Ou, X., Liu, Y., & Xue, C. (2021). Detection of mental fatigue state using heart rate variability and eye metrics during simulated flight. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *31*(6), 637-651. https://doi.org/https://doi.org/10.1002/hfm.20927

Roza, V. C., Postolache, O., Groza, V., & Pereira, J. M. D. (2019, 26-28 June 2019). Emotions Assessment on Simulated Flights. 2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA),

Samani, S., Jessop, R., & Harrivel, A. (2021, 11-13 Aug. 2021). Collaborative Communications Between A Human And A Resilient Safety Support System. 2021 IEEE International Conference on Autonomous Systems (ICAS),

Samima, S., & Sarma, M. (2019, 23-27 July 2019). EEG-Based Mental Workload Estimation. 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),

Sant'Anna, D. A. L. M. d., & Hilal, A. V. G. d. (2021). The impact of human factors on pilots' safety behavior in offshore aviation companies: A brazilian case. *Safety Science*, *140*, 105272. https://doi.org/https://doi.org/10.1016/j.ssci.2021.105272

Sarter, N. B., Mumaw, R. J., & Wickens, C. D. (2007). Pilots' Monitoring Strategies and Performance on Automated Flight Decks: An Empirical Study Combining Behavioral and Eye-Tracking Data. *Human Factors*, *49*(3), 347-357. https://doi.org/https://doi.org/10.1518/001872007X196685

Sauvet, F., Bougard, C., Coroenne, M., Lely, L., Beers, P. V., Elbaz, M., Guillard, M., Léger, D., & Chennaoui, M. (2014). In-Flight Automatic Detection of Vigilance States Using a Single EEG Channel. *IEEE Transactions on Biomedical Engineering*, *61*(12), 2840-2847. https://doi.org/10.1109/TBME.2014.2331189

Scannella, S., Peysakhovich, V., Ehrig, F., Lepron, E., & Dehais, F. (2018). Assessment of Ocular and Physiological Metrics to Discriminate Flight Phases in Real Light Aircraft. *Hum Factors*, *60*(7), 922-935. https://doi.org/10.1177/0018720818787135

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

Shuang, H., Chuanfeng, W., & Qi, W. (2017, 26-28 July 2017). Recognition of fatigue status of pilots based on deep sparse auto-encoding network. 2017 36th Chinese Control Conference (CCC),

SKY_Brary. (2019). *SE211_ Airplane State Awareness - Training for Attention Management (R-D)*. https://skybrary.aero/articles/se211-airplane-state-awareness-training-attention-management-r-d

Snider, D. H., Linnville, S. E., Phillips, J. B., & Rice, G. M. (2022). Predicting hypoxic hypoxia using machine learning and wearable sensors. *Biomedical Signal Processing and Control*, *71*. https://doi.org/10.1016/j.bspc.2021.103110

Socha, V., Vidensky, J., Kusmirek, S., Hanakova, L., & Valenta, V. (2022). Design of Wearable Eye Tracker with Automatic Cockpit Areas of Interest Recognition. 2022 New Trends in Civil Aviation (NTCA),

Sonnleitner, A., Treder, M. S., Simon, M., Willmann, S., Ewald, A., Buchner, A., & Schrauf, M. (2014). EEG alpha spindles and prolonged brake reaction times during auditory distraction in an on-road driving study. *Accid Anal Prev*, *62*, 110-118. https://doi.org/10.1016/j.aap.2013.08.026

Stanton, N. A., Chambers, P. R. G., & Piggott, J. (2001). Situational awareness and safety. *Safety Science*, *39*(3), 189-204. https://doi.org/https://doi.org/10.1016/S0925-7535(01)00010-8

Stanton, N. A., Plant, K. L., Revell, K. M. A., Griffin, T. G. C., Moffat, S., & Stanton, M. (2019). Distributed cognition in aviation operations: a gate-to-gate study with implications for distributed crewing. *Ergonomics*, *62*(2), 138-155. https://doi.org/10.1080/00140139.2018.1520917

Subasi, A. (2007). EEG signal classification using wavelet feature extraction and a mixture of expert model. *Expert Systems with Applications*, *32*(4), 1084-1093. https://doi.org/10.1016/j.eswa.2006.02.005

Taheri Gorji, H., Wilson, N., VanBree, J., Hoffmann, B., Petros, T., & Tavakolian, K. (2023). Using machine learning methods and EEG to discriminate aircraft pilot cognitive workload during flight. *Sci Rep*, *13*(1), 2507. https://doi.org/10.1038/s41598-023-29647-0

Thomas, L. C., Gast, C., Grube, R., & Craig, K. (2015). Fatigue detection in commercial flight operations: Results using physiological measures. *6th International Conference on Applied Human Factors and Ergonomics (Ahfe 2015) and the Affiliated Conferences, Ahfe 2015*, *3*, 2357-2364. https://doi.org/10.1016/j.promfg.2015.07.383

Tortora, S., Beraldo, G., Bettella, F., Formaggio, E., Rubega, M., Del Felice, A., Masiero, S., Carli, R., Petrone, N., Menegatti, E., & Tonin, L. (2022). Neural correlates of user learning during long-term BCI training for the Cybathlon competition. *Journal of NeuroEngineering and Rehabilitation*, *19*(1), 69. https://doi.org/10.1186/s12984-022-01047-x

Trejo, L. J., Kubitz, K., Rosipal, R., Kochavi, R. L., & Montgomery, L. D. (2015). EEG-Based Estimation and Classification of Mental Fatigue. *Psychology*, *06*(05), 572-589. https://doi.org/10.4236/psych.2015.65055

Tripathi, S., Acharya, S., Sharma, R., Mittal, S., & Bhattacharya, S. (2017). Using Deep and Convolutional Neural Networks for Accurate Emotion Classification on DEAP Data. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(2), 4746-4752. https://doi.org/10.1609/aaai.v31i2.19105

Wang, H., Jiang, N., Pan, T., Si, H., Li, Y., Zou, W., & Du, Y. (2020). Cognitive Load Identification of Pilots Based on Physiological-Psychological Characteristics in Complex Environments. *Journal of Advanced Transportation*, *2020*, 1-16. https://doi.org/10.1155/2020/5640784

Wang, Q., Wang, Z., Xiong, R., Liao, X., & Tan, X. (2023). A Method for Classification and Evaluation of Pilot's Mental States Based on CNN. *Computer Systems Science and Engineering*, *46*(2). https://doi.org/10.32604/csse.2023.034183

Wei-Long, Z., & Bao-Liang, L. (2015). Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural

Networks. *IEEE Transactions on Autonomous Mental Development*, *7*(3), 162-175. https://doi.org/10.1109/tamd.2015.2431497

Wiegmann, D. A., Goh, J., & O'Hare, D. (2002). The Role of Situation Assessment and Flight Experience in Pilots' Decisions to Continue Visual Flight Rules Flight into Adverse Weather. *Human Factors*, *44*(2), 189-197. https://doi.org/10.1518/0018720024497871

Wu, E. Q., Deng, P. Y., Qiu, X. Y., Tang, Z. R., Zhang, W. M., Zhu, L. M., Ren, H., Zhou, G. R., & Sheng, R. S. F. (2021). Detecting Fatigue Status of Pilots Based on Deep Learning Network Using EEG Signals. *IEEE Transactions on Cognitive and Developmental Systems*, *13*(3), 575-585. https://doi.org/10.1109/Tcds.2019.2963476

Wu, E. Q., Peng, X. Y., Zhang, C. Z. Z., Lin, J. X., & Sheng, R. S. F. (2019). Pilots' Fatigue Status Recognition Using Deep Contractive Autoencoder Network. *Ieee Transactions on Instrumentation and Measurement*, *68*(10), 3907-3919. https://doi.org/10.1109/Tim.2018.2885608

Wu, E. Q., Zhou, M., Hu, D., Zhu, L., Tang, Z., Qiu, X. Y., Deng, P. Y., Zhu, L. M., & Ren, H. (2022). Self-Paced Dynamic Infinite Mixture Model for Fatigue Evaluation of Pilots' Brains. *IEEE Transactions on Cybernetics*, *52*(7), 5623-5638. https://doi.org/10.1109/TCYB.2020.3033005

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). A Comprehensive Survey on Graph Neural Networks. *IEEE Trans Neural Netw Learn Syst*, *32*(1), 4-24. https://doi.org/10.1109/TNNLS.2020.2978386

Xi, Z., Newton, O., McGowin, G., Sukthankar, G., Fiore, S., & Oden, K. (2020). Predicting Student Flight Performance with Multimodal Features. In *Social, Cultural, and Behavioral Modeling* (pp. 277-287). https://doi.org/10.1007/978-3-030-61255-9_27

Yang, S., Yin, Z., Wang, Y., Zhang, W., Wang, Y., & Zhang, J. (2019). Assessing cognitive mental workload via EEG signals and an ensemble deep learning classifier based on denoising autoencoders. *Comput Biol Med*, *109*, 159-170. https://doi.org/10.1016/j.compbiomed.2019.04.034

Yingxue, L., & Qi, W. (2019, 3-5 June 2019). Pilots' Brain Cognitive State Inference Based on Remaining Life HSMM. 2019 Chinese Control And Decision Conference (CCDC),

Yiu, C. Y., Ng, K. K. H., Li, X., Zhang, X., Li, Q., Lam, H. S., & Chong, M. H. (2022). Towards safe and collaborative aerodrome operations: Assessing shared situational awareness for adverse weather detection with EEG-enabled Bayesian neural networks. *Advanced Engineering Informatics*, *53*. https://doi.org/10.1016/j.aei.2022.101698

Zhang, P., Wang, X., Chen, J., You, W., & Zhang, W. (2019). Spectral and Temporal Feature Learning With Two-Stream Neural Networks for Mental

Workload Assessment. *IEEE Trans Neural Syst Rehabil Eng*, *27*(6), 1149-1159. https://doi.org/10.1109/TNSRE.2019.2913400

Zhu, W., Zhang, C., Liu, C., Yuan, J., Li, X., Wang, Y., & Jiang, C. (2023). *Recognition of Pilot Mental workload in the Simulation Operation of Carrier-based Aircraft Using the Portable EEG* Proceedings of the 2023 3rd International Conference on Human Machine Interaction,

Zhuang, H., Yang, B., Li, B., Zan, P., Ma, B., & Meng, X. (2021, 26-28 July 2021). EEG Based Eye Movements Multi-Classification Using Convolutional Neural Network. 2021 40th Chinese Control Conference (CCC),

Ziegler, M. D., Kraft, A., Krein, M., Lo, L.-C., Hatfield, B., Casebeer, W., & Russell, B. (2016). Sensing and Assessing Cognitive Workload Across Multiple Tasks.*Lecture Notes in Computer Science* Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience,

3 Miscellaneous EEG Preprocessing and Machine Learning for Pilots' Mental States Classification: Implications

3.1 Abstract

Higher cognitive process efforts may result in mental exhaustion, poor performance, and long-term health issues. An EEG-based methods for detecting a pilot's mental state have recently been created utilizing machine learning algorithms. EEG signals include a significant noise component, and these approaches either ignore this or use a random mix of preprocessing techniques to reduce noise. In the absence of uniform preprocessing procedures for cleaning, it would be impossible to compare the efficacy of machine learning models across research, even if they employ data obtained from the same experiment. In this study, we intend to evaluate how preprocessing approaches affect the performance of machine learning models. To do this, we concentrated on fundamental preprocessing techniques, such as a band-pass filter and independent component analysis. Using a publicly accessible actual physiological dataset gathered from pilots who was exposed to a variety of mental events, we explore the influence of these preprocessing strategies on two machine learning models, SVMs and ANNs. Our findings reveal that the application of band-pass filtering and ICA preprocessing techniques can indeed influence the performance of these ML models, though the extent and nature of this impact vary across different datasets. Moreover, our findings indicate that the models were able to anticipate the mental states from merged data collected in two environments. These findings demonstrate the necessity for a standardized methodological framework for the application of machine learning models to EEG inputs.

3.2 Introduction

Detecting the pilot's mental state is critical to ensuring the safety of the plane's flight path (Oehling & Barry, 2019). Performance measures, questionnaires, and

neurophysiological methods such as brain activity have all been shown by researchers to be effective in detecting people's mental states (Reid & Nygren, 1988). In particular, brain activity measures using electroencephalography (EEG) have been shown to be the most effective method of identifying the mental state of pilots. This is due to the fact that EEG signals can capture brain activity with great temporal resolution (Arico et al., 2015). Thus, improved machine learning (ML) models have been built to reliably diagnose mental states by capturing variance properties in EEG data. Support Vector Machines (SVMs) and Artificial Neural Networks (ANNs) are examples of such efforts (Raduntz et al., 2017).

However, neuroscientists have demonstrated that EEG signals are susceptible to noise. Numerous efforts have been made in the past to improve data preprocessing procedures to reduce noise, and standardized protocols for cleaning EEG data have been constructed. The ML algorithms that utilise EEG signals do not, however, adhere to a typical data decontamination approach. Due to these anomalies in their data preparation, it is impossible to quantify the true impact of ML models. Second, even when utilizing the same experiment's data, it is impossible to compare the outcomes of various experiments. In light of these facts, there is still work to be done on the standardised artefact removal procedure to be used in ML (Bigdely-Shamlo et al., 2015).

Consequently, the purpose of this work is to address the following research question: "What are the impacts of various preprocessing procedures on the performance of ML models that use EEG data to classify the pilot's mental states?" We examine the influence of applying various preprocessing approaches to EEG data using SVM and ANN algorithms. The SVM and ANN models were trained with data collected from pilots in a non-flight environment, a flight environment, and merged flight and non-flight environment data using a 5-fold cross-validation method. The training data of the SVM and ANN models consists of unprocessed data, filtered data, and data that has been filtered and eye blinks were removed from it.

The paper is organized as follows: Section 2 provides an overview of mental states and relevant research studies. Section 3 describes the methods used.

Section 4 presents the findings and discussion, and Section 5 offers a summary and conclusion.

3.3 Background

Mental workload, particularly within the demanding environment of aviation, plays a crucial role in influencing a pilot's performance. Recognized as either overload or underload, both extremes can detrimentally affect performance, impacting the efficiency and quality of operations within complex working systems. An underload of mental workload occurs when the cognitive demands placed on an individual are too low, leading to issues such as decreased vigilance, boredom, and a reduction in task engagement, which can be just as hazardous as overload in critical occupations like piloting (Longo et al., 2012). The influence of mental underload is particularly concerning as it can lead to decreased situational awareness and a slower response to unexpected events, compromising safety (Young et al., 2015).

To ensure optimal human-system interaction, accurately monitoring and detecting mental states has become imperative. Traditional methods like self-report questionnaires or performance assessments on secondary tasks have been utilized to gauge mental states, requiring individuals to be well-versed with the reporting tools, which paradoxically may add to their workload (Wiebe et al., 2010). Consequently, there has been a shift towards utilizing EEG recordings to classify mental states, given their direct measure of brain activity. However, the reliability of EEG signals is contingent upon the minimization of noise interference, known as artifacts, which can stem from both internal sources like ocular, muscle, and cardiovascular activities, and external sources such as instrument or body movement noise (Kim, 2018). Artifacts can significantly degrade the quality of EEG data, thus undermining the accuracy of the analytical models employed. Therefore, artifact removal has become a critical prerequisite in the preprocessing of EEG data to ensure the validity of mental state classifications (Uriguen & Garcia-Zapirain, 2015).

Although neuroscientists propose various EEG preprocessing guidelines (Bigdely-Shamlo et al., 2015; Gabard-Durnam et al., 2018; Makeig et al., 2012), they are rather general and not universally accepted. The process for some existing pipelines includes a visual inspection and hand labelling (Mognon et al., 2011). While these techniques can be highly beneficial for reducing signal noise, they have three limitations: they are time-intensive, especially when working with large datasets; they can introduce bias into the analysis (Vaid et al., 2015); and they restrict the use of such pipelines in automated procedures.

3.3.1 Standard and Problem-Dependent EEG Pre-processing Techniques

For epilepsy detection and other applications, EEG pre-processing involves several methodologies that are crucial for enhancing signal quality and facilitating accurate analysis. While specific pre-processing techniques can vary depending on the application, certain methods such as bandpass filtering and Independent Component Analysis (ICA) for artifact removal are widely adopted across various fields. These standard practices reflect a consensus on effective strategies for minimizing noise and extracting meaningful data from EEG signals.

In the domain of epilepsy detection, various pre-processing methods are used to enhance EEG signal quality before epileptic seizure detection, including normalization, filtering (to remove noise and artifacts), and signal segmentation. For instance, a study proposed a one-dimensional convolutional neural network-long short-term memory (1D-CNN-LSTM) model for epileptic seizure recognition, where raw EEG signals were first pre-processed and normalized before feature extraction and classification (Xu et al., 2020). This suggests a tailored approach to pre-processing that is aligned with the objectives of specific applications. Similarly, a study on depression diagnosis using EEG signals employed variational mode decomposition (VMD) and standardized low-resolution brain electromagnetic tomography (sLORETA) for EEG source localization (Kaur et al., 2019). This approach demonstrated effectiveness in

handling EEG signals for depression patients, highlighting the importance of accurate localization in EEG pre-processing.

In the context of brain-computer interface (BCI) systems, a review focused on EEG-based BCI systems emphasized the significance of artifact detection and removal, considering artifacts from internal and external factors as major challenges in EEG signal quality (Maswanganyi et al., 2018). The study presented various efficient techniques for artifact detection and elimination, underscoring the critical role of signal pre-processing in BCI applications. Furthermore, the use of empirical mode decomposition (EMD) and wavelet transform (WT) in EEG pre-processing has been analysed for improving brain-source reconstruction accuracy (Munoz-Gutierrez et al., 2018). These techniques, by offering good frequency and temporal resolution, have shown promise in applications such as epilepsy, ADHD, and evoked-related potentials.

Despite these commonalities, the selection of pre-processing techniques is informed by the specific needs of the application and the characteristics of the EEG data. For instance, a review covering recent works in EEG pre-processing and feature extraction highlighted the importance of these steps for applications like drowsiness detection and assistive technologies (Sarma et al., 2016). It discussed various pre-processing techniques that prepare EEG signals for further analysis or classification by capturing essential signal details.

These studies collectively indicate that while the specific pre-processing methods may vary depending on the application, there are common techniques employed across different EEG-based applications. These include signal decomposition, artifact removal, and feature extraction, which are crucial for improving the signal quality and reliability of EEG data for further analysis. The selection of pre-processing techniques is guided by the nature of the EEG data, the specific requirements of the application, and the type of analysis or classification to be performed. This problem-dependent approach ensures that the pre-processing stage effectively enhances signal quality and facilitates the accurate interpretation of EEG data, thereby supporting the diverse objectives

of epilepsy detection, clinical diagnosis, BCI systems, and other EEG-based applications.

3.3.2 Pilot's Mental State Classification through EEG Signal Processing and ML Applications

EEG signals could efficiently indicate the pilot's mental state. Due to its ability to gather accurate data representations of features, ML has recently been successfully applied to EEG analysis (Roy et al., 2019). For EEG preprocessing, some studies applied noise reduction methodologies; however, the efficacy of each strategy on ML models for mental state classification has not been evaluated. Such models cannot be compared due to the lack of a consistent preprocessing framework.

In a simulation environment, Chaudhuri et al. (Chaudhuri & Routray, 2020) applied a band-pass filter to the EEG data to remove extraneous signals and the SVM algorithm to identify normal and fatigue states. As a result, their classification accuracy has increased by an average of 86%. The band-pass filter was also utilized in the investigation by Han et al. (Han et al., 2020). Nonetheless, the frequency range has been adjusted to a different value. In this work, the sampling frequency was adjusted between 0.1 and 50 Hz, and the ICA components pertaining to eye blinks and movements were eliminated. Using SVMs, k-Nearest Neighbours (k-NN), Logistic Regression (LR), Random Forest (RF), shrinkage Linear Discriminant Analysis (sLDA), and deep Convolutional Neural Network (CNN) classifiers, the preprocessed data has been utilised to detect four distinct mental states induced by four benchmark activities. The obtained classification accuracy ranged between 64% and 83%. A notch filter is an additional type of filter that has been applied. In their investigation, Binias et al. (Binias et al., 2018) used the EPOC+ headset, which contains an integrated digital 5th-order Sinc filter, notch filters at 50Hz and 60Hz, and a band-pass filter between 0.16 and 43 Hz. LDA, k-NN, SVMs, RF, and ANNs have categorised two mental states that were induced to distinguish between states of brain activity associated with idle but concentrated anticipation of a visual cue and a reaction to it. The average accuracy of the

proposed models ranges from 67 to 78%. The Butterworth band-pass filter with a high-pass cut-off frequency of 0.5 Hz and a low-pass cut-off frequency of 50 Hz was used in (Yang et al., 2019), but the authors left ocular artefacts in their data. The filtered data was then incorporated into a two-stream neural network (TSNN) model for a three-class mental workload classification task (Zhang et al., 2019). The model's average degree of accuracy is 91.9 percent. According to published studies, using preprocessing methods developed by neurobiologists, ML researchers have attempted to reduce noise from their EEG data. However, there is no standard preprocessing approach that everyone follows. In particular, the band-pass filter approach, which appears to be the most popular tool, has been characterized in a variety of ways.

In addition, event participants classify the same dataset using various preprocessing methods. Using their dataset, Harrivel et al. (Harrivel et al., 2016; Harrivel et al., 2017) performed attention-related human performance-limiting states (AHPLS) classification. The authors examined frequency domain components between 0 and 40 Hz using the A Lomb-Scargle frequency transform during the artefact removal phase. This method of spectral analysis takes sample rate abnormalities into consideration. The authors have achieved 82% AHPLS classification accuracy with a Deep Neural Network (DNN) model. Harrivel et al. (Harrivel et al., 2017) generated 40 power spectral density (PSD) features per channel to represent the EEG frequency bands between 1 and 40 Hz. Using gradient boosting, RF, and SVM classifiers, they achieved 50 to 78% classification accuracy. The AHPLS dataset was also utilised for AHPLS classification in the Terwilliger et al. (Terwilliger, 2020). However, no preprocessing approaches to eliminate signal artefacts were utilised. The proposed ResNet Autoencoder model has been fed with the original data for AHPLS detection. They have undertaken an analysis to determine whether or not there is an event. In their study, the proposed model showed a low rate of false positives and false negatives.

The ML articles for EEG analysis do not take into account the effect of preprocessing processes, hence findings cannot be compared across studies.

In order to determine the effect of preprocessing strategies on ML models, we conducted nine experiments employing three experimental preprocessing scenarios on EGG data collected from pilots in two distinct settings. Using a 5-fold cross-validation method, we developed and trained SVM and ANN models with non-flight environment data, flight environment data, and merged flight and non-flight environment data. The training data of the SVM and ANN models comprises unfiltered data, filtered data, and data from which eye blinks have been removed. The following section presents the used cases as well as a framework for removing EEG artefacts for ML model evaluation.

3.4 Methodology

3.4.1 Data Acquisition

The dataset was acquired from the website of NASA's open data portal. It contains experimental EEG data gathered from 18 pilots who were required to complete tasks in two environments. Included in the tasks were resting tasks, benchmark tasks meant to elicit AHPLS, and experimental flight situations. LaRC's Research Flight Deck and Cockpit Motion Facility were utilized for the collecting of data. A psychophysiological sensor was utilized to measure EEG signals using an Advanced Brain Monitoring X24 EEG System. The EEG system consists of 20 electrodes in the standard 10-20 format + POz (Fz, Cz, Pz, F3, F4, C3, C4, P3, P4, O1, O2, T5, T3, F7, Fp1, Fp2, F8, T4, and T6 with Linked Mastoids) with sampling rate of 256 Hz. The signals were captured in two distinct contexts, a motion-based flight simulator and a non-flying environment. The pilot did a complete flying simulation using the motion-based flight simulator (take off, flight and landing). The pilot conducted three benchmark activities outside of the flight simulation environment. The situations that the pilot encountered during the experiments were designed to elicit one of the three cognitive states listed below:

Channelized Attention (CA): the state of concentrating on a single task. The benchmarking process is triggered by having the pilot play a puzzle-based video game.

Diverted Attention (DA): the state of having one's attention diverted by decisions-related behaviours or thinking processes. This is accomplished by having pilots conduct a display monitoring task while math problems appear intermittently and must be solved before returning to the monitoring task.
Startle/Surprise (SS): it is induced by showing the pilot jump-scare movie clips.

The EEG data was downloaded in CSV format. To perform fundamental and sophisticated preprocessing approaches, we used an open-source library (MNE-Python) and generated an appropriate object for continuous EEG data's core data structures (i.e., raw object) (Gramfort et al., 2013). The core data structures object is initialized with the necessary fields of information, including a list of channel names and types, the standard montage naming schemes, and the sampling frequency. The EEG data were then segmented into trials of one second with no overlap.

3.4.2 EEG Preprocessing

EEG preprocessing involves multiple strategies. Some can reduce data noise automatically, while others must be performed manually. The purpose of this work is to examine the impact of preprocessing procedures that require manual execution, as opposed to those that function autonomously, i.e., devoid of human interaction. The advantage of an automated analysis lies in its capacity to circumvent the issue of subjective identification of artefacts through visual inspection (Vaid et al., 2015). Consequently, we study the impact of the two most prevalent preprocessing approaches, a band-pass filter and ICA. Contrary to the automatic nature of the band-pass filter, ICA in this context is employed for the manual removal of eye-related artefacts. As a result, we have three experimental cases for each type of environmental data: non-flight environment data, flight environment data, and flight and non-flight environment data combined. Following is a description of the three experimental cases:
Case 1: Unprocessed data. The data have not been preprocessed. Figure 3-1 illustrates a 10 second window of the unprocessed EEG data collected from the pilot in a non-flight environment.

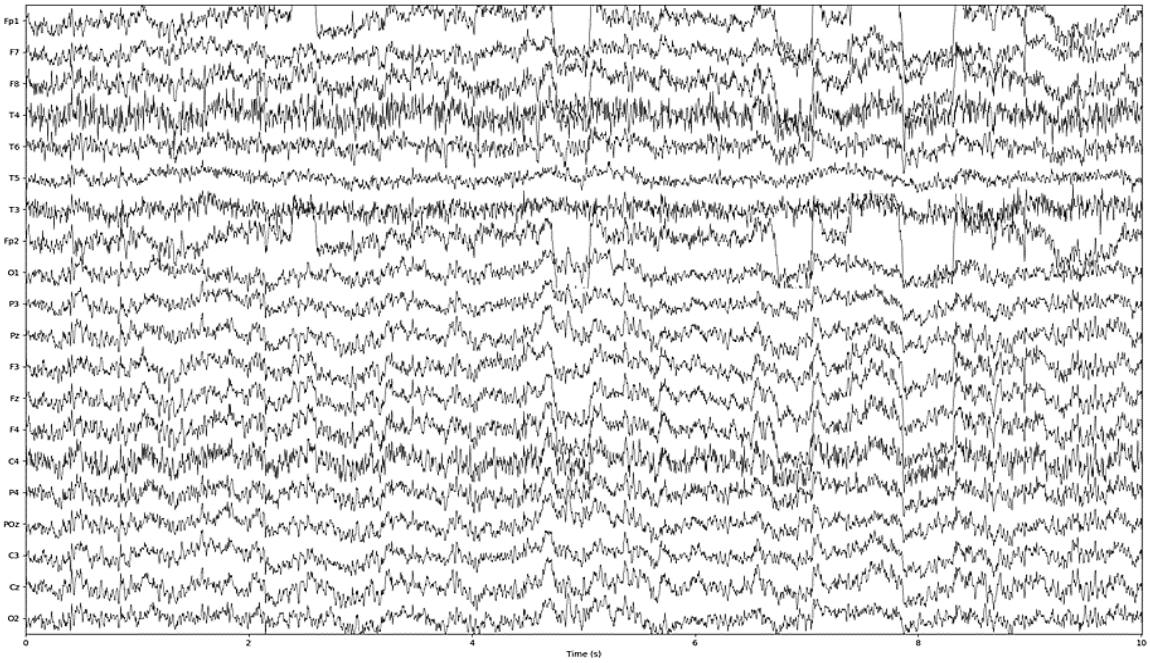

Figure 3-1 The unprocessed EEG signal

Case 2: Band-pass filtering data. In this instance, a band-pass filter (finite impulse response (FIR), 1-50 Hz) was utilized to minimize artefacts and increase signal-to-noise ratio (SNR) appears in Figure 3-1. Figure 3-2 illustrates the filtered EEG data.

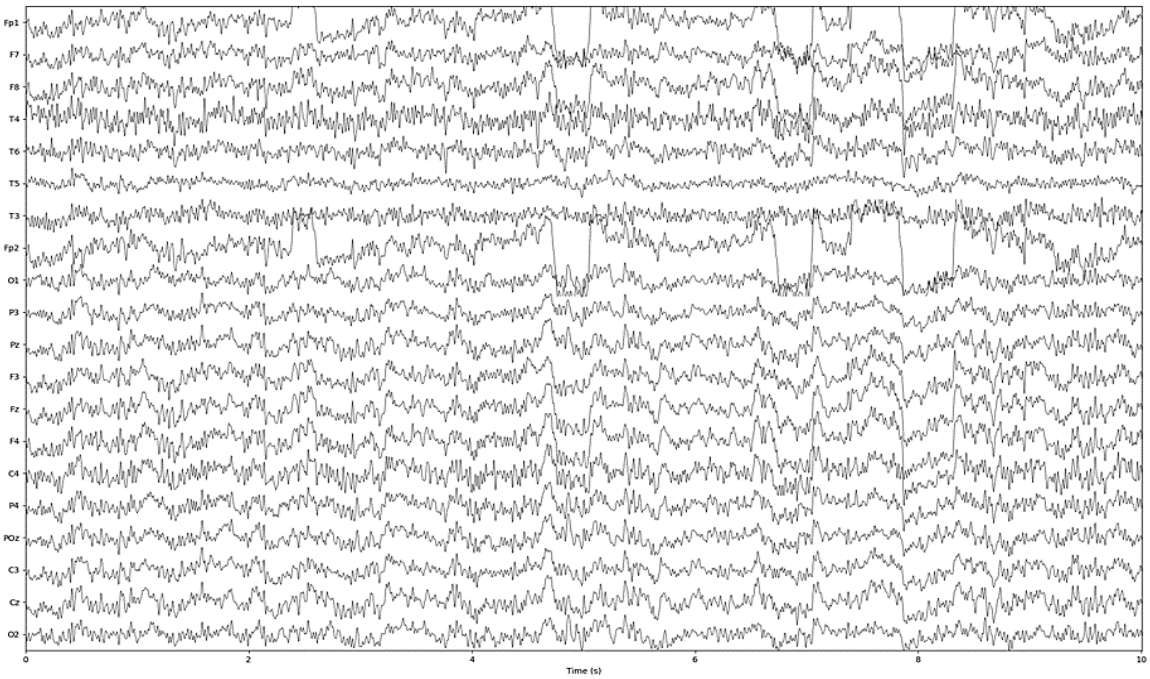

Figure 3-2: The filtered EEG signal

Case 3: Bandpass filtering + ICA data. To extract the eye-related artefact components from the EEG signals in this instance, ICA was applied to the filtered signals from the previous phase (Case 2) using the Fastica approach (Uriguen & Garcia-Zapirain, 2015). After partitioning the multichannel EEG into ICs, eye blinks, as it can be seen from Figure 3-2 at the fifth, seventh and eighth seconds in channel Fp1 and Fp2 which are the channels near the eyes, were automatically recognized and rectified. As shown in Figure 3-3, eye blinks are removed from the EEG signals.

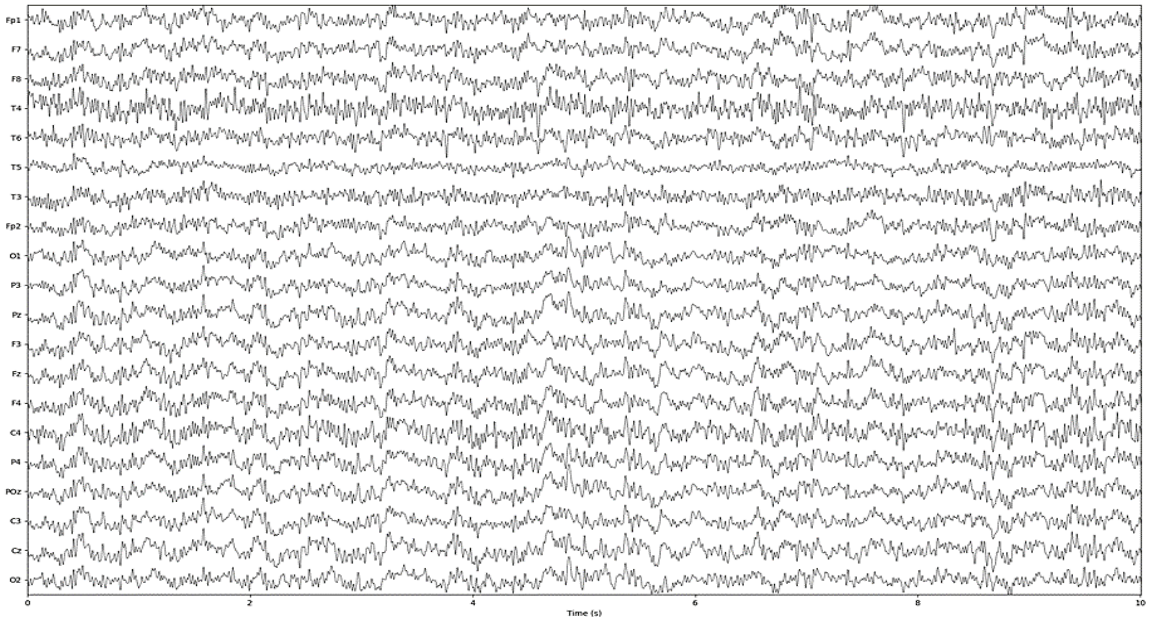

Figure 3-3: EEG signal after filtering and removing eye-related artefacts.

Table 3-1 is a summary of our three preprocessing cases.

Table 3-1 Preprocessing cases

Case	Preprocessing procedure
1	None (EEG Raw data)
2	Band-pass filtering
3	Band-pass filtering and ICA

3.4.3 EEG Feature Extraction/Engineering

The objective of feature extraction in EEG analysis is to discern informative characteristics within the EEG signals. The EEG data underwent spatial filtering using the Xdawn technique, and subsequently, Tangent space projection was applied.

Xdawn Technique: The Xdawn algorithm, typically known for enhancing event related potentials (ERP) components, was here applied to the EEG data. It computes spatial filters that maximize the variance of the signal within each epoch (Hajinoroozi et al., 2016). This approach is effective in extracting features that are representative of the entire EEG signal, not just ERPs. The spatial filtering process can be represented as:

$$V = \sum_{signal}(\sum_{signal} + \sum_{noise})^{-1} \quad \textbf{(3-1)}$$

where $\sum_{signal}$ is the covariance matrix of the EEG signal, and $\sum_{noise}$ is the covariance matrix of background EEG activity.

Tangent Space Projection: Post spatial filtering, Tangent space projection transforms the covariance matrices into a Euclidean space. This transformation is crucial for applying advanced classifiers to EEG data, which often exhibits complex, high-dimensional structures (Barachant et al., 2012). It involves projecting a covariance matrix C onto a tangent space at a reference point, typically the mean covariance matrix C_m, expressed as:

$$T(C) = \log(C_m^{-1/2} \,.\, C \,.\, C_m^{-1/2}) \quad \textbf{(3-2)}$$

Here, log denotes the matrix logarithm, and the operation transforms the manifold-structured data into a linear space, suitable for conventional ML models.

The choice of spatial filtering (i.e., Xdawn technique) and Tangent space projection was driven by their suitability for the specific classification tasks in this chapter. While there are many features extraction and dimension reduction techniques available, these methods were selected for their effectiveness in

enhancing and transforming EEG data into a format that is compatible with ML models.

3.4.4 Classification Models

In EEG data analysis, the choice of ML models is critical for discerning complex neural patterns. These models need to effectively generalize from training data to new, unseen instances, which is particularly challenging due to the high variability and complexity inherent in EEG signals. This study focuses on SVMs and ANNs, specifically Multilayer Perceptron (MLPs), to assess the impact of different preprocessing techniques on EEG data classification. The models' performances in classifying EEG signals were evaluated using accuracy, precision, recall, and F1-score metrics, providing a comprehensive understanding of its classification capabilities.

SVMs. SVMs are powerful classes of supervised learning models used for classification and regression. In the context of EEG data analysis, SVMs offer several advantages. First, they are effective in high-dimensional spaces, typical of EEG datasets. SVMs function by finding the optimal hyperplane that maximizes the margin between different classes. This is achieved through the use of kernels, which transform the original data into a higher-dimensional space where it becomes linearly separable. The Radial Basis Function (RBF) kernel is particularly suited for EEG data due to its ability to handle non-linear separations (Bishop, 2006). In this study, the RBF kernel's capacity to manage the complexity and non-linearity of EEG signals was leveraged to enhance classification accuracy.

ANNs. In this research, the specific type of ANN employed is the MLP, a class of feedforward ANN. The MLP architecture includes an input layer, one hidden layer with 100 neurons, and an output layer. Each neuron uses a ReLU activation function, chosen for its effectiveness in processing tangent space vectors derived from EEG data. The ReLU function is particularly well-suited for handling the properties of these tangent space vectors, which often encapsulate the high-dimensional and non-linear characteristics of EEG signals.

The training of the MLP involves back-propagation, a widely used method for training ANNs where the network adjusts its weights based on the errors it makes. The learning process was configured with a learning rate of 0.001, employing a stochastic gradient-based optimizer (i.e., Adam). We chose to train the model over 100 iterations based on a blend of factors including convergence analysis and experimental findings. This number of iterations was found to be optimal for ensuring sufficient learning of the complex patterns in EEG data. Convergence analysis indicated that beyond 100 iterations, improvements in model performance were marginal, suggesting that the model effectively captures the underlying patterns in the data within this timeframe.

The choice of SVMs and MLPs in this study is driven by their respective strengths in modelling complex data. SVMs, with their structured approach to maximizing class separability, are well-suited for pattern recognition tasks in high-dimensional spaces like EEG data. On the other hand, MLPs offer a dynamic and flexible architecture capable of capturing non-linear relationships within the data, which is critical for EEG signal analysis. This comparison provides a comprehensive understanding of how different ML paradigms perform in the context of EEG preprocessing and classification, thereby contributing valuable insights into their applicability and effectiveness in EEG data analysis.

3.5 Results and discussion

In this section, we present a detailed examination of the results derived from a comprehensive analysis of EEG preprocessing techniques and their impact on ML models in aviation contexts. The study utilises datasets from 18 pilots, encompassing both non-flight and flight simulator environments, in addition to a merged dataset integrating these two scenarios. The primary focus is on assessing the efficacy of these preprocessing methods in enhancing the performance of SVM and ANN models for the detection of four key mental states, namely CA, DA, SS, and NE. This exploration is critical for understanding how different environmental contexts and preprocessing techniques influence the accuracy and reliability of ML models in identifying

cognitive states in pilots. The analyses encompass a broad spectrum of conditions, providing insights into the generalisability and robustness of these techniques across varied scenarios.

The results are methodically outlined, offering a comprehensive view of the impact of each preprocessing technique on the performance of the ML models. Table 3-2 summarises the classification results obtained from the non-flight environment data, the flight environment data, and the combined dataset, thereby facilitating a comparative understanding of the outcomes across different experimental setups.

Table 3-2 Classification results

Environment type	Data type	Model type	Evaluation Metric			
			Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Non-flight	Unprocessed	SVM	86.11	67.52	46.46	46.26
		ANN	85.83	62.21	56.28	57.90
	Filtered	SVM	85.96	69.39	45.79	46.04
		ANN	85.31	59.17	52.55	55.11
	Filtered + ICA	SVM	93.00	66.42	52.06	52.95
		ANN	92.22	67.87	60.24	63.36
Flight	Unprocessed	SVM	90.42	46.64	25.03	25.40
		ANN	87.40	36.51	33.31	33.58
	Filtered	SVM	90.22	46.66	25.06	24.14
		ANN	85.71	36.90	33.37	34.33
	Filtered + ICA	SVM	89.82	43.08	25.04	24.44
		ANN	86.44	37.59	33.43	35.37
Merged	Unprocessed	SVM	87.06	44.62	35.87	38.23

		ANN	83.25	43.99	41.92	43.42
	Filtered	SVM	87.07	44.82	35.11	37.80
		ANN	82.64	45.10	42.57	44.14
	Filtered + ICA	SVM	90.09	46.98	40.10	41.75
		ANN	86.14	52.03	47.66	50.48

3.5.1 Performance Evaluation of ML Models with Unprocessed EEG Data (Case 1)

In this subsection, we explore the performance of SVM and ANN models using unprocessed EEG data. The focus is on understanding how these models behave with raw data across different settings: non-flight, flight, and a combined dataset.

Non-flight Data: The results reveal that both the SVM and ANN models can effectively capture relevant information and classify mental states with notable model performance scores. Specifically, the SVM model achieved an accuracy of 86%, with a precision of 68%, and the ANN model displayed a similar accuracy level, alongside a precision of 62%. The recall and F1-score further refine our understanding of the models' predictive power, with the SVM and ANN models recording values of 46% and 58% for the F1-score, respectively. The confusion matrix, as visualised in Figure 3-4 , illustrates the distribution of predictions across the various mental states. It is observed that the SVM model exhibits challenges in the accurate classification of the SS state, which is predominantly confused with the NE state. This is also apparent in the ANN model, though to a lesser extent, indicating a potential area for improvement in the detection of subtle mental state nuances. The preponderance of misclassified samples as normal states suggests that the unprocessed data may suffer from a low SNR, which could detrimentally impact the learning algorithms' performance. Such noise within the data has the propensity to unnecessarily complicate model architectures and protract the learning process,

thereby warranting further exploration into more advanced preprocessing strategies to bolster classification accuracy.

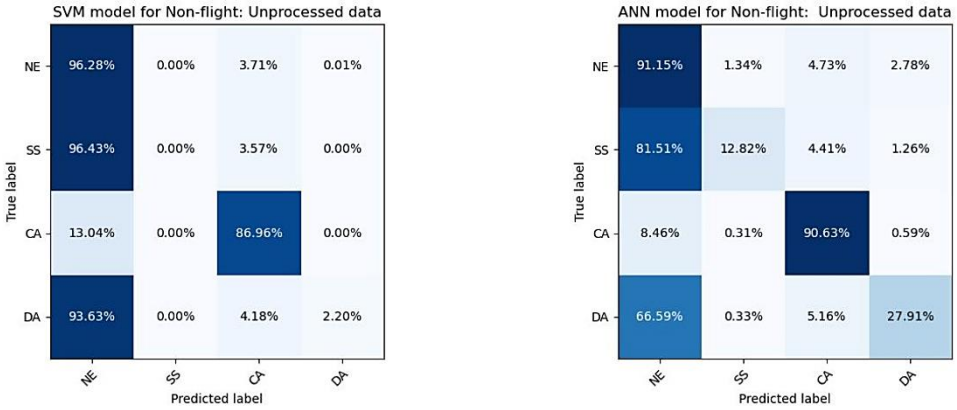

Figure 3-4 Confusion matrices for SVM and ANN models using unprocessed non-flight data

Flight Data: Examination of the unprocessed flight data presents an intriguing perspective on the performance of both the SVM and ANN models in classifying mental states. The SVM model demonstrated a high level of accuracy at 90%, yet it is notable that this was accompanied by a relatively low precision of 47%. The recall and F1-score, both standing at 25%, suggest a model that, while accurate in general, may be overly conservative in its predictions, leading to a substantial number of false negatives. This is further elucidated by the confusion matrix, depicted in Figure 3-5 , where the SVM model shows an overwhelming tendency to classify states as NE, with negligible recognition of other mental states such as SS, CA, and DA. In contrast, the ANN model, while slightly less accurate at 87%, exhibited a more balanced performance across the metrics of precision (37%), recall (33%), and F1-score (34%). The corresponding confusion matrix reveals a more evenly distributed set of predictions across the four mental states. Notably, the ANN model demonstrates a discernible aptitude in identifying the CA state, with a correct classification rate of 20.84%, which markedly surpasses the SVM's performance in this regard. These results suggest that while the SVM model excels in overall accuracy, it does so at the cost of a nuanced understanding of diverse mental states, predominantly categorising most instances as NE. The ANN model,

conversely, shows a more equitable distribution in its predictive capability, albeit with a slight reduction in overall accuracy. This disparity in model performance underscores the potential challenges in employing unprocessed flight data for mental state classification. The high accuracy yet low diversity in SVM's predictions might indicate an inclination towards overfitting to the most common state, while the more balanced ANN outcomes suggest a model that is potentially more adaptable, but less precise. This dichotomy highlights the necessity for refined preprocessing methods that can enhance the SNR and facilitate more accurate and diverse classifications by ML algorithms.

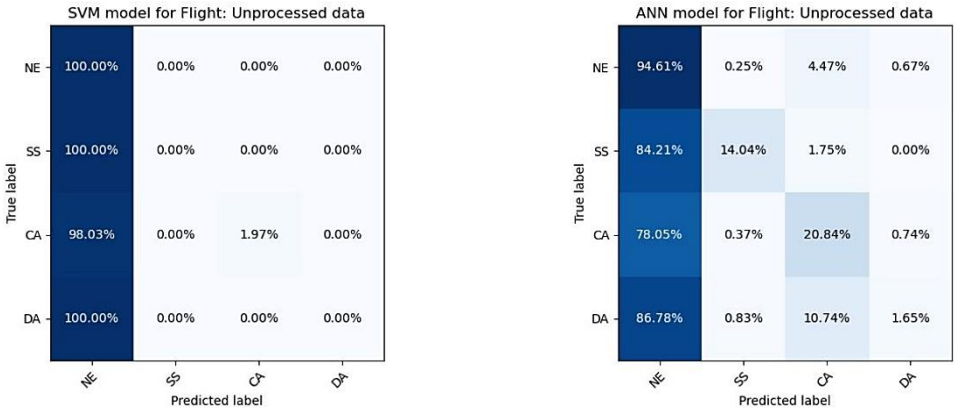

Figure 3-5 Confusion matrices for SVM and ANN models using unprocessed flight data

Merged Non-Flight and Flight Data: The fusion of unprocessed non-flight and flight data offers a comprehensive insight into the functionality of ML models in a broader aviation context. The SVM model, when applied to this combined dataset, achieved an accuracy of 87%, yet its precision (45%), recall (36%), and F1-score (38%) indicate a nuanced complexity in its classification capability. As illustrated in Figure 3-6, the confusion matrix for the SVM model reveals a pronounced bias towards predicting the NE state, with exceptionally high accuracy. However, this comes at the expense of other mental states, particularly CA and DA, where the model only correctly identifies 43.21% and 1.65% of cases, respectively. In comparison, the ANN model, though slightly less accurate overall with an 83% accuracy rate, demonstrates a more balanced performance. The precision, recall, and F1-score are all relatively aligned, standing at 44%, 42%, and 43% respectively. The ANN's confusion

matrix indicates a more equitable distribution of predictive success across the different mental states. Notably, the model shows a markedly improved ability to classify the CA state correctly in 58.22% of cases, a significant improvement over the SVM's performance. These results highlight the challenges and potential benefits of utilizing a merged dataset encompassing both non-flight and flight environments. The high degree of accuracy in predicting the NE state by the SVM model, while impressive, suggests a possible overfitting to the most dominant state in the dataset. Conversely, the more evenly distributed predictions of the ANN model suggest a greater adaptability to varied mental states but with a compromise in overall accuracy. This comparative analysis underscores the importance of selecting and fine-tuning ML models according to the specific characteristics of the dataset. Moreover, it raises critical considerations about the suitability of unprocessed data in such complex classification tasks, pointing towards the potential need for more advanced preprocessing techniques to enhance the models' ability to discern and accurately classify diverse mental states in aviation scenarios.

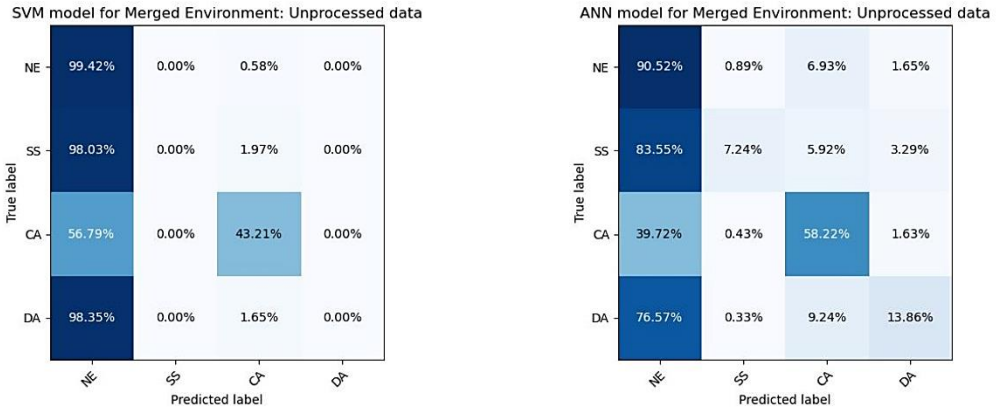

Figure 3-6 Confusion matrices for SVM and ANN models using unprocessed merged flight and non-flight data

In summary, the evaluation of ML models using unprocessed EEG data across various aviation-related scenarios has yielded interesting findings. The performance metrics of the SVM and ANN models, as applied to non-flight, flight, and combined datasets, have highlighted distinct characteristics and limitations inherent in each model when dealing with raw EEG data.

3.5.2 Performance Evaluation of ML Models with Filtered EEG Data (Case 2)

In this subsection, we explore the performance of SVM and ANN models using filtered EEG data. The focus is on understanding how these models behave with filtered data across different settings: non-flight, flight, and a combined dataset.

Filtered Non-Flight Data: The analysis of the bandpass filtered non-flight data through SVM and ANN models reveals significant insights into the classification of mental states. The SVM model demonstrated an accuracy of 86%, with a precision of 69%, indicating its effectiveness in correctly identifying true positives. The recall and F1-score, both at 46%, suggest moderate sensitivity in detecting all relevant instances. According to the confusion matrix shown in Figure 3-7, the SVM model is highly effective in identifying the NE state, with a 95.87% accuracy. However, it exhibits a notable limitation in distinguishing the SS state, incorrectly classifying 96.29% of SS instances as NE. The model performs remarkably well in identifying CA with an 86.49% accuracy but shows a similar tendency to misclassify DA as NE. The ANN model, on the other hand, shows a slightly lower overall accuracy of 85% but a more balanced performance across precision (59%), recall (53%), and F1-score (55%). The confusion matrix for the ANN model indicates a more equitable classification capability across the mental states. It correctly identifies 88.95% of CA instances and shows a significant improvement in classifying DA, with 24.40% accuracy. However, like the SVM model, the ANN also predominantly misclassifies SS as NE, though to a lesser extent. These results demonstrate the impact of bandpass filtering in enhancing the classification accuracy of both models, particularly in the identification of CA and DA states. However, the high misclassification rate of SS as NE by both models indicates a challenge in distinguishing SS from NE states using filtered data. This suggests a potential area for improvement in the preprocessing or modelling techniques to better capture the nuances of SS state. The findings highlight the importance of

tailored preprocessing strategies in EEG data analysis for accurate mental state classification.

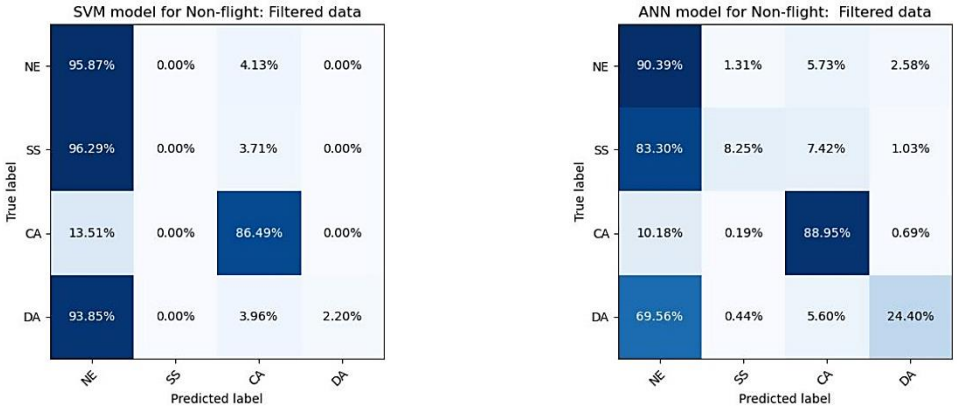

Figure 3-7 Confusion matrices for SVM and ANN models using filtered non-flight data

Filtered Flight Data: The analysis of the filtered flight data using the SVM and ANN models presents an intriguing perspective on the classification of mental states in an aviation setting. The SVM model achieved a high accuracy of 90%, but this figure belies certain limitations in its classification capability. The precision of 47%, along with a notably lower recall of 25% and an F1-score of 24%, indicates a model that is highly accurate overall but has considerable difficulties in correctly identifying positive instances across diverse mental states. The confusion matrix, as presented in Figure 3-8, starkly illustrates this issue. The SVM model uniformly classifies almost all instances as the NE state, with negligible recognition of the SS, CA, and DA states. Conversely, the ANN model exhibits a more balanced, albeit slightly less accurate, performance with an overall accuracy of 86%. Its precision, recall, and F1-score are more evenly distributed at 37%, 33%, and 34% respectively. The corresponding confusion matrix for the ANN model suggests a more diverse classification capability, with more evenly distributed predictions across the different mental states. While it still shows a predominant classification of NE, there is a noticeable improvement in identifying SS, CA, and DA states compared to the SVM model, indicating a better grasp of the nuances in mental state variations. These findings underscore the challenges inherent in employing filtered flight data for

mental state classification. The high accuracy yet low diversity in the SVM's predictions might point towards a tendency for overfitting to the most common state, possibly due to the filtering process overly simplifying the data's complexity. On the other hand, the more balanced ANN outcomes suggest a model that, while less precise, is potentially more adaptable to the intricacies of different mental states. This comparison highlights the importance of considering both the nature of the dataset and the characteristics of the chosen ML models. It also emphasises the need for careful consideration of preprocessing techniques in EEG data analysis, as they can significantly influence the models' ability to accurately and comprehensively classify complex mental states in a high-stakes environment like aviation.

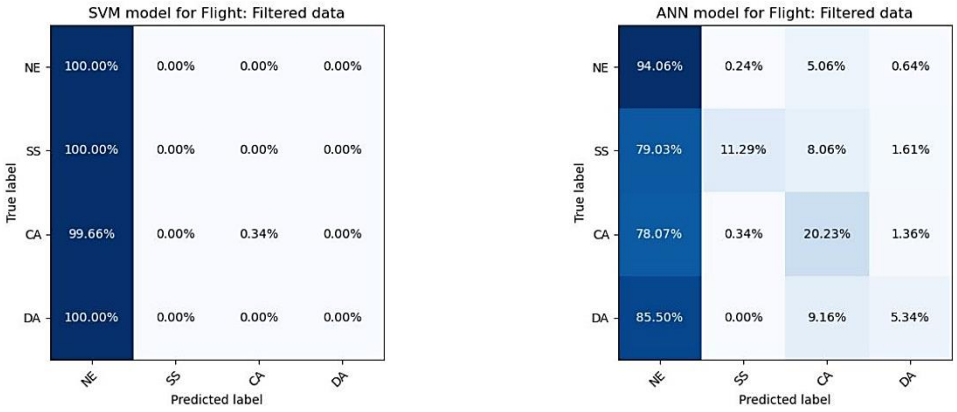

Figure 3-8 Confusion matrices for SVM and ANN models using filtered flight data

Filtered Merged Flight and Non-Flight Data: The application of bandpass filtering to the merged dataset of flight and non-flight data offers a comprehensive view of the efficacy of SVM and ANN models in classifying various mental states. The SVM model, achieving an accuracy of 87%, demonstrates a commendable overall classification capability. However, the precision of 45%, coupled with a recall of 35% and an F1-score of 38%, suggests certain limitations in its ability to distinguish between different mental states accurately. The confusion matrix, as shown in Figure 3-9, underscores this limitation, revealing the model's overwhelming tendency to categorise the majority of instances as the NE state, with substantial misclassification of SS and DA states as NE. In contrast, the ANN model exhibits a slightly lower

overall accuracy of 83% but presents a more balanced performance in terms of precision (45%), recall (43%), and F1-score (44%). The corresponding confusion matrix for the ANN model indicates a more evenly distributed predictive capability across the mental states. While it still shows a high classification rate for the NE state, the model demonstrates an improved ability to identify CA (53.95%) and DA (13.42%) states correctly. This suggests a greater sensitivity of the ANN model to the nuances of different mental states, likely attributed to its inherent capacity for handling the complexities in the merged dataset. These results highlight the impact and potential benefits of preprocessing techniques like bandpass filtering in enhancing the classification performance of ML models, particularly in complex datasets that combine different environmental contexts. The high accuracy yet low diversity in the SVM's predictions might be indicative of a tendency towards overgeneralisation, potentially due to the preprocessing approach simplifying the dataset's complexity. Meanwhile, the more evenly distributed predictions of the ANN model suggest a capacity to better navigate the intricacies of diverse mental states, albeit with a slight reduction in overall accuracy. This comparative analysis emphasises the significance of selecting suitable ML models and preprocessing methods that align with the specific challenges and requirements of mental state classification, especially in diverse and complex aviation environments.

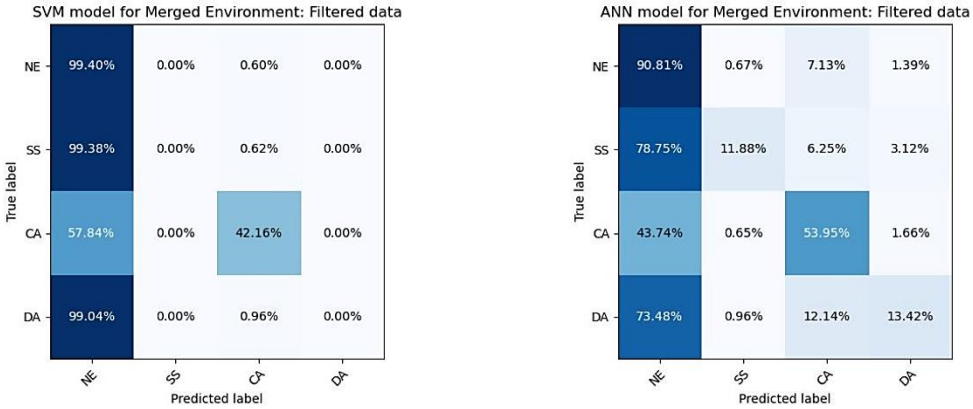

Figure 3-9 Confusion matrices for SVM and ANN models using filtered merged flight and non-flight data

Given the classification results for both unprocessed and filtered data across non-flight, flight, and merged datasets, we observe the following impacts of filtering EEG data on the performance of the SVM and ANN models:

1. **Non-Flight Data**: The introduction of filtering in the non-flight data shows a marginal impact on the performance of both models. For the SVM, accuracy remains constant at 86%, with a slight improvement in precision (from 68% to 69%) but no change in recall and F1-score. The ANN model shows a minor decrease in accuracy (from 86% to 85%), precision (from 62% to 59%), recall (from 56% to 53%), and F1-score (from 58% to 55%). This suggests that filtering had a minimal effect on enhancing the models' ability to distinguish between different mental states in the non-flight environment.

2. **Flight Data**: In the flight data, filtering does not appear to significantly alter the performance metrics of both models. The accuracy, precision, recall, and F1-score for the SVM remain unchanged, indicating that the model's performance is not substantially influenced by the filtering process. Similarly, the ANN model's performance metrics remain consistent before and after filtering, suggesting that filtering EEG data did not provide a notable advantage in this context.

3. **Merged Flight and Non-Flight Data**: For the merged datasets, filtering shows a negligible impact on the SVM model's performance, with accuracy, precision, and F1-score remaining the same, and a slight decrease in recall (from 36% to 35%). The ANN model, however, exhibits a slight improvement in precision (from 44% to 45%) and F1-score (from 43% to 44%), with accuracy and recall remaining consistent. This indicates a modest enhancement in the ANN model's precision in classifying mental states post-filtering.

Overall, the filtering process does not significantly enhance the performance of the SVM and ANN models across the different datasets. The improvements are either marginal or non-existent, suggesting that the filtering technique used may not be optimally aligned with the specific characteristics of the EEG data and the mental states being classified. This highlights the importance of considering

the nature of the data and the specific requirements of the classification task when selecting and applying preprocessing techniques in EEG-based studies.

3.5.3 Performance Evaluation of ML Models with Filtered and ICA EEG Data (Case 3)

In this subsection, we explore the performance of SVM and ANN models using EEG data that are filtered and eye-related artefacts were removed using ICA. The focus is on understanding how these models behave with filtered + ICA data across different settings: non-flight, flight, and a combined dataset.

Filtered + ICA Non-Flight Data: The analysis of the non-flight data, subjected to bandpass filtering and ICA for eye-related artefact removal, through SVM and ANN models, provides a nuanced view of mental state classification. The SVM model achieves a high accuracy of 93%, with a precision of 66%, suggesting effective identification of true positives. However, its recall and F1-score, at 52% and 53% respectively, indicate a limitation in capturing all instances of the mental states. The confusion matrix, detailed in Figure 3-10, shows that the SVM model is exceptionally accurate in classifying the NE state with 99.42% accuracy and CA with 99.38% accuracy. However, it notably misclassifies all instances of SS as NE, indicating a significant limitation in differentiating these specific states. The model also shows varied performance in identifying DA, correctly classifying 10.99% of these instances. The ANN model, displaying an accuracy of 92%, presents a more balanced performance across precision (68%), recall (60%), and F1-score (63%). Its confusion matrix reveals a more evenly distributed classification capability, accurately identifying 99.39% of CA instances and showing an improved ability to classify DA, with 31.43% accuracy. However, similar to the SVM model, it demonstrates a challenge in classifying SS, albeit with a better performance than the SVM, correctly identifying 13.40% of SS instances. These results underscore the complexity of mental state classification using EEG data, especially in the context of differentiating between states such as SS and NE. While the preprocessing techniques improve the overall signal quality, enabling high accuracy in certain states, they also highlight the challenges in distinguishing states with potentially

subtle EEG differences. This analysis emphasises the need for further refinement in ML models and possibly more advanced preprocessing approaches to enhance the sensitivity and specificity of mental state classification in non-flight aviation settings.

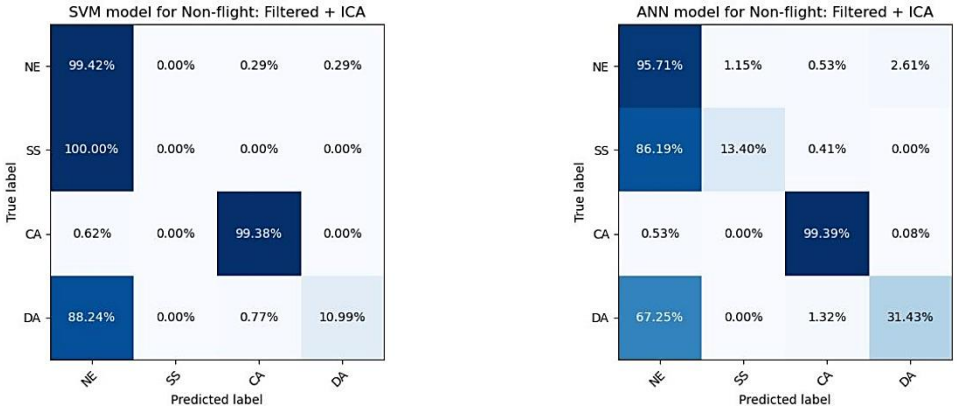

Figure 3-10 Confusion matrices for SVM and ANN models using filtered + ICA non-flight data

Filtered + ICA Flight Data: The implementation of bandpass filtering coupled with ICA to remove eye-related artefacts in the flight data provides a nuanced understanding of the capabilities of SVM and ANN models in mental state classification. The SVM model exhibits a high accuracy of 90%, which, while impressive, is accompanied by a precision of 43%, a recall of 25%, and an F1-score of 24%. This indicates a model that is predominantly accurate in a general sense but struggles with the detailed classification of various mental states. The confusion matrix, depicted in Figure 3-11, reinforces this observation, revealing an overwhelming tendency of the SVM to classify nearly all instances as the NE state, with minimal identification of other states such as SS, CA, and DA. In contrast, the ANN model, showing an accuracy of 86%, presents a more evenly balanced performance with a precision of 38%, a recall of 33%, and an F1-score of 35%. Its confusion matrix indicates a more diversified approach in classifying the different mental states. While the model still predominantly classifies instances as NE, it demonstrates an improved capacity to identify SS, CA, and DA, with particularly notable performance in recognising CA (22.05%) state. These results highlight the potential benefits

and limitations of employing advanced preprocessing techniques, such as bandpass filtering combined with ICA, in EEG data analysis for aviation settings. The high overall accuracy of the SVM model suggests that the preprocessing effectively enhances the signal's clarity, aiding in general classification tasks. However, the model's limited success in distinguishing between the more nuanced mental states points to potential overfitting issues or a lack of sensitivity to subtle EEG variations. On the other hand, the ANN model's more balanced classification across different states suggests that it is better equipped to handle the complexities introduced by the filtering and ICA processes, albeit with a slight compromise in overall accuracy.

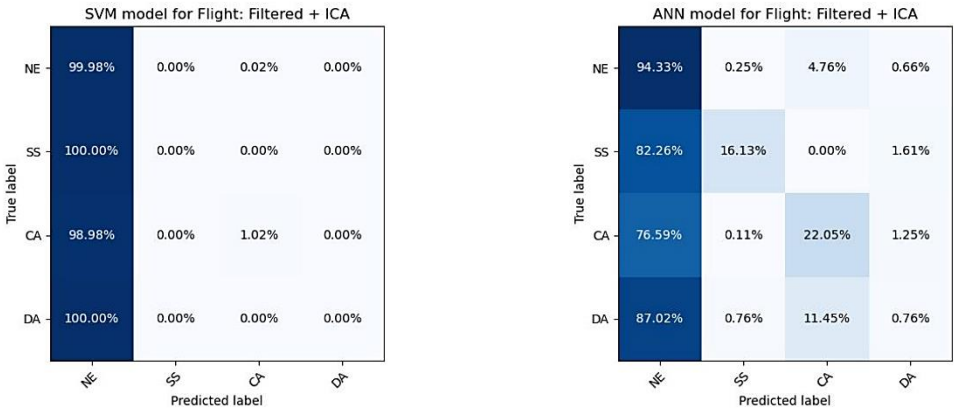

Figure 3-11 Confusion matrices for SVM and ANN models using filtered + ICA flight data

Filtered + ICA Merged Flight and Non-Flight Data: The utilisation of bandpass filtering and ICA on the combined flight and non-flight data offers insightful revelations when analysed through SVM and ANN models. The SVM model shows a high accuracy of 90%, complemented by a precision of 47%, indicating its efficacy in identifying true positive instances. However, the recall and F1-score, standing at 40% and 42% respectively, suggest a moderate level of sensitivity and precision in predicting different mental states. The confusion matrix, as presented in Figure 3-12, reveals a significant inclination of the SVM model to classify the majority of instances as the NE state with an accuracy of 99.89%. It completely classifies all SS instances as NE, indicating a profound challenge in distinguishing this state. The model performs better in identifying

CA, with a correct classification rate of 58.90%, but it similarly misclassifies most DA instances as NE. In contrast, the ANN model, with an accuracy of 86%, demonstrates a more balanced performance across precision (52%), recall (48%), and F1-score (50%). The corresponding confusion matrix for the ANN model indicates a more equitable distribution of predictive success across different mental states. While it also predominantly classifies instances as NE, there is a noticeable improvement in the identification of other states. The model correctly identifies 65.74% of CA instances and demonstrates a significantly better ability to classify DA, with 19.81% of DA instances accurately identified. Additionally, it shows an enhanced capacity to identify SS, with a correct classification rate of 12.50%, though still with room for improvement. These results highlight the nuanced challenges in mental state classification using EEG data, especially when merging datasets from diverse environments like flight and non-flight contexts. The preprocessing techniques, while enhancing overall signal quality, also reveal the models' limitations in distinguishing certain mental states, particularly SS. This analysis emphasises the need for further refinement in both ML models and preprocessing approaches to improve the accuracy and diversity of mental state classifications in complex aviation environments.

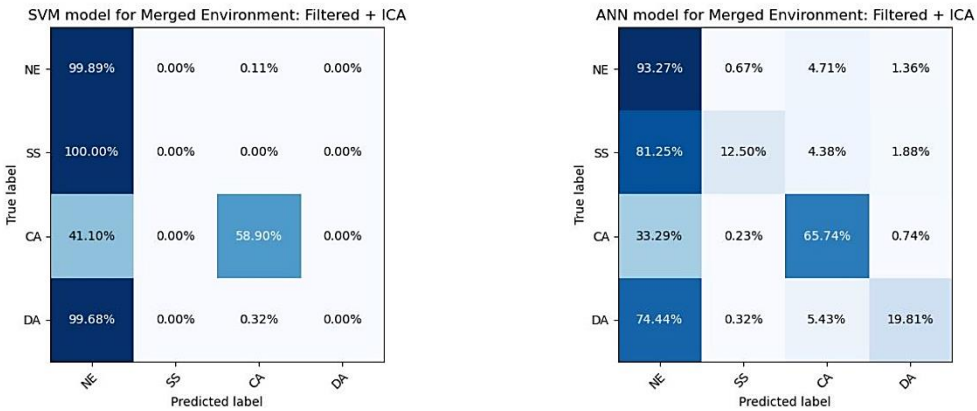

Figure 3-12 Confusion matrices for SVM and ANN models using filtered + ICA merged flight and non-flight data

The integration of filtering with ICA for the removal of eye-related artefacts from EEG data presents distinct impacts on the performance of SVM and ANN models across different datasets (non-flight, flight, and merged).

1. **Non-Flight Data (Filtered + ICA)**: The SVM model exhibits a significant improvement in the non-flight data, with accuracy increasing from 86% to 93%, precision from 68% to 66%, recall from 46% to 52%, and F1-score from 46% to 53%. This suggests that the combined preprocessing techniques substantially enhance the model's ability to accurately classify mental states. The ANN model also shows considerable improvement, with accuracy rising from 86% to 92%, precision from 62% to 68%, recall from 56% to 60%, and F1-score from 58% to 63%. These improvements indicate that the filtering and ICA effectively enhance data quality, leading to better model performance.

2. **Flight Data (Filtered + ICA)**: For the flight data, the SVM model's performance metrics remain relatively unchanged post-filtering and ICA, suggesting that these preprocessing techniques do not significantly alter its ability to classify mental states in this context. The ANN model shows a slight decrease in accuracy from 87% to 86%, but an improvement in precision from 37% to 38%, and F1-score from 34% to 35%.

3. **Merged Flight and Non-Flight Data (Filtered + ICA)**: In the combined dataset, the SVM model shows an increase in accuracy from 87% to 90%, precision from 45% to 47%, recall from 36% to 40%, and F1-score from 38% to 42%. This indicates that filtering and ICA preprocessing contribute positively to the model's performance in a more complex dataset. The ANN model also demonstrates improved performance, with accuracy increasing from 83% to 86%, precision from 44% to 52%, recall from 42% to 48%, and F1-score from 43% to 50%.

The high accuracy yet low precision, recall, and F1-scores observed in Table 3-2 primarily stem from the imbalanced nature of your dataset, where a significant majority (around 80%) of the samples represent the NE state. In such scenarios, even simple models that predict the majority class for all instances

can achieve high accuracy, simply because they are correct for the majority class most of the time. However, this approach fails to capture the essence of the minority classes effectively, leading to poor precision and recall. Precision measures the proportion of true positive predictions in all positive predictions, and recall measures the proportion of true positive predictions out of all actual positives. When a model predominantly classifies instances as the NE state, it may rarely predict minority classes (i.e., CA, DA, and SS states). Consequently, even a few misclassifications can drastically affect precision and recall, leading to their low values. The F1-score, which is the harmonic mean of precision and recall, is also low as it reflects the imbalance in the performance metrics, emphasizing the model's inability to correctly identify and classify the minority classes.

The reliance on accuracy as a performance metric in the presence of imbalanced datasets can be misleading, as observed in your experiments. Accuracy calculates the proportion of true positive and true negative predictions among all predictions, which, in the case of imbalanced datasets, favours the majority class. The confusion matrices in Figures 3-4 to 3-12, where most samples are classified as NE state, especially with SVM for flight data, illustrate this issue. Since the majority of the dataset comprises the NE state, classifying most samples as NE inflates the accuracy metric, giving the illusion of a well-performing model. However, this does not reflect the model's performance on minority classes, where it is crucial to detect nuanced mental states in pilots accurately.

In summary, the application of filtering and ICA preprocessing techniques generally enhances the performance of both SVM and ANN models, particularly in the non-flight and merged datasets. These improvements are most notable in terms of accuracy, precision, recall, and F1-score, indicating that the combined preprocessing methods are effective in improving the signal quality of EEG data and, consequently, the ability of the models to accurately classify different mental states. The impact is more pronounced in the non-flight data, suggesting that these techniques are particularly effective in contexts with less complex

signal characteristics. However, the relatively modest improvements in the flight data imply that the effectiveness of preprocessing may vary depending on the specific characteristics of the dataset and the inherent complexities of the mental states being classified.

3.6 Conclusion

In this chapter, we investigated the effect of neuroscientist-defined preprocessing approaches on the efficacy of ML models in classifying pilot's mental states using EEG data. Specifically, we focused on the application of band-pass filtering and ICA as our primary preprocessing techniques. Utilising a publicly available dataset of EEG signals, we examined the impact of these preprocessing strategies on enhancing the performance of two widely used ML models: SVMs and ANNs.

Our analysis aligns with existing literature, which substantiates the positive influence of conventional EEG preprocessing methods, such as band-pass filtering and ICA, on the performance of ML models. Studies across various applications beyond pilot mental state classification have demonstrated similar improvements. For example, (Sundaram et al., 2022) found that employing band-pass filtering and ICA significantly enhanced the classification accuracies of SVM and ANN models in diagnosing neurological disorders through EEG signal analysis. This evidence corroborates our findings, suggesting that the effectiveness of these preprocessing techniques is not limited to specific datasets or mental state classifications but extends across diverse EEG analysis applications. In our experiment with non-flight data, the incorporation of filtering and ICA preprocessing markedly improved accuracy, precision, recall, and F1-score for both SVM and ANN models, highlighting these methods' utility in environments with simpler signal characteristics. However, the enhancements observed in flight data were more moderate, underscoring that the efficacy of such preprocessing methods may be dependent on the dataset's inherent complexities and the specific mental states under examination. Moreover, our results demonstrate the adaptability of SVM and ANN models to effectively process combined datasets from different environments, such as

flight and non-flight settings, opening new avenues for future research aimed at developing robust models capable of generalizing across varied environmental contexts. The challenge of extracting brain activity linked to specific cognitive tasks in EEG data analysis necessitates refined strategies for discerning relevant information within raw data. Our exploration into the effects of band-pass filtering and the ICA algorithm is supported by the broader literature and lays the groundwork for further studies into diverse artifact removal strategies, especially in the context of complex cognitive tasks and heterogeneous EEG datasets.

Future research should extend to evaluating the impact of these preprocessing strategies on a broader spectrum of ML and DL models. Moreover, this study underscores the potential benefits of developing and implementing advanced automated preprocessing pipelines. Such pipelines could integrate advanced tools like Autoreject algorithm to dynamically identify and correct various artefacts and inconsistencies in EEG data. This integrated approach would not only streamline the preprocessing phase but also significantly enhance the reliability and accuracy of EEG data analysis, contributing substantially to advancements in neuroscientific research and its applications.

3.7 Appendices

3.7.1 Appendix A: Data and Reproducibility Code

In the interest of promoting transparency and reproducibility, the data utilised in this chapter, along with the associated code for analyses, have been made publicly accessible. The dataset and the code for replicating the analyses can be found under the Digital Object Identifier (DOI): https://doi.org/10.17862/cranfield.rd.24156249

REFERENCES

Arico, P., Borghini, G., Di Flumeri, G., Colosimo, A., Graziani, I., Imbert, J. P., Granger, G., Benhacene, R., Terenzi, M., Pozzi, S., & Babiloni, F. (2015). Reliability over time of EEG-based mental workload evaluation during Air Traffic Management (ATM) tasks. *Annu Int Conf IEEE Eng Med Biol Soc*, *2015*, 7242-7245. https://doi.org/10.1109/EMBC.2015.7320063

Barachant, A., Bonnet, S., Congedo, M., & Jutten, C. (2012). Multiclass brain-computer interface classification by Riemannian geometry. *IEEE Trans Biomed Eng*, *59*(4), 920-928. https://doi.org/10.1109/TBME.2011.2172210

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K. M., & Robbins, K. A. (2015). The PREP pipeline: standardized preprocessing for large-scale EEG analysis. *Front Neuroinform*, *9*, 16. https://doi.org/10.3389/fninf.2015.00016

Binias, B., Myszor, D., & Cyran, K. A. (2018). A Machine Learning Approach to the Detection of Pilot's Reaction to Unexpected Events Based on EEG Signals. *Comput Intell Neurosci*, *2018*, 2703513. https://doi.org/10.1155/2018/2703513

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer New York, NY.

Chaudhuri, A., & Routray, A. (2020). Driver Fatigue Detection Through Chaotic Entropy Analysis of Cortical Sources Obtained From Scalp EEG Signals. *Ieee Transactions on Intelligent Transportation Systems*, *21*(1), 185-198. https://doi.org/10.1109/Tits.2018.2890332

Gabard-Durnam, L. J., Mendez Leal, A. S., Wilkinson, C. L., & Levin, A. R. (2018). The Harvard Automated Processing Pipeline for Electroencephalography (HAPPE): Standardized Processing Software for Developmental and High-Artifact Data. *Front Neurosci*, *12*, 97. https://doi.org/10.3389/fnins.2018.00097

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hamalainen, M. (2013). MEG and EEG data analysis with MNE-Python. *Front Neurosci*, *7*, 267. https://doi.org/10.3389/fnins.2013.00267

Hajinoroozi, M., Mao, Z. J., Jung, T. P., Lin, C. T., & Huang, Y. F. (2016). EEG-based prediction of driver's cognitive performance by deep convolutional neural network. *Signal Processing-Image Communication*, *47*, 549-555. https://doi.org/10.1016/j.image.2016.05.018

Han, S. Y., Kwak, N. S., Oh, T., & Lee, S. W. (2020). Classification of pilots' mental states using a multimodal deep learning network. *Biocybernetics and Biomedical Engineering*, *40*(1), 324-336. https://doi.org/10.1016/j.bbe.2019.12.002

Harrivel, A. R., Liles, C., Stephens, C. L., Ellis, K. K., Prinzel, L. J., & Pope, A. T. (2016). Psychophysiological Sensing and State Classification for Attention Management in Commercial Aviation. AIAA Infotech @ Aerospace,

Harrivel, A. R., Stephens, C. L., Milletich, R. J., Heinich, C. M., Last, M. C., Napoli, N. J., Abraham, N., Prinzel, L. J., Motter, M. A., & Pope, A. T. (2017). Prediction of Cognitive States during Flight Simulation using Multimodal Psychophysiological Sensing. AIAA Information Systems-AIAA Infotech @ Aerospace,

Kaur, H., Pannu, H. S., & Malhi, A. K. (2019). A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. *ACM Comput. Surv.*, *52*(4), Article 79. https://doi.org/10.1145/3343440

Kim, S.-P. (2018). Preprocessing of EEG. In *Computational EEG Analysis* (pp. 15-33). https://doi.org/10.1007/978-981-13-0908-3_2

Longo, L., Rusconi, F., & Noce, L. (2012). *The Importance of Human Mental Workload in Web Design* Proceedings of the 8th International Conference on Web Information Systems and Technologies,

Makeig, S., Kothe, C., Mullen, T., Bigdely-Shamlo, N., Zhang, Z. L., & Kreutz-Delgado, K. (2012). Evolving Signal Processing for Brain-Computer Interfaces. *Proceedings of the Ieee*, *100*(Special Centennial Issue), 1567-1584. https://doi.org/10.1109/Jproc.2012.2185009

Maswanganyi, C., Tu, C., Owolawi, P., & Du, S. (2018, 27-29 June 2018). Overview of Artifacts Detection and Elimination Methods for BCI Using EEG. 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC),

Mognon, A., Jovicich, J., Bruzzone, L., & Buiatti, M. (2011). ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features. *Psychophysiology*, *48*(2), 229-240. https://doi.org/10.1111/j.1469-8986.2010.01061.x

Munoz-Gutierrez, P. A., Giraldo, E., Bueno-Lopez, M., & Molinas, M. (2018). Localization of Active Brain Sources From EEG Signals Using Empirical Mode Decomposition: A Comparative Study. *Front Integr Neurosci*, *12*, 55. https://doi.org/10.3389/fnint.2018.00055

Oehling, J., & Barry, D. J. (2019). Using machine learning methods in airline flight data monitoring to generate new operational safety knowledge from existing data. *Safety Science*, *114*, 89-104. https://doi.org/10.1016/j.ssci.2018.12.018

Raduntz, T., Scouten, J., Hochmuth, O., & Meffert, B. (2017). Automated EEG artifact elimination by applying machine learning algorithms to ICA-based features. *J Neural Eng*, *14*(4), 046004. https://doi.org/10.1088/1741-2552/aa69d1

Reid, G. B., & Nygren, T. E. (1988). The Subjective Workload Assessment Technique: A Scaling Procedure for Measuring Mental Workload. In *Human Mental Workload* (pp. 185-218). https://doi.org/10.1016/s0166-4115(08)62387-0

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., & Faubert, J. (2019). Deep learning-based electroencephalography analysis: a systematic review. *J Neural Eng*, *16*(5), 051001. https://doi.org/10.1088/1741-2552/ab260c

Sarma, P., Tripathi, P., Sarma, M. P., & Sarma, K. K. (2016). Pre-processing and Feature Extraction Techniques for EEGBCI Applications- A Review of Recent Research. *ADBU Journal of Engineering Technology (AJET)*, *5*.

Sundaram, C. K., Gayathri, S., & Soundarya, P. (2022). Discrete wavelet transform based on EEG signal analysis for diagnosing neurological disorder. *International journal of health sciences*, 9556-9566. https://doi.org/10.53730/ijhs.v6nS1.7213

Terwilliger, P. S., Jack; Walker, Shannon; Harrivel, Angela. (2020). *A ResNet Autoencoder Approach for Time Series Classification of Cognitive State* MODSIM,

Uriguen, J. A., & Garcia-Zapirain, B. (2015). EEG artifact removal-state-of-the-art and guidelines. *J Neural Eng*, *12*(3), 031001. https://doi.org/10.1088/1741-2560/12/3/031001

Vaid, S., Singh, P., & Kaur, C. (2015). *EEG Signal Analysis for BCI Interface: A Review* 2015 Fifth International Conference on Advanced Computing & Communication Technologies,

Wiebe, E. N., Roberts, E., & Behrend, T. S. (2010). An examination of two mental workload measurement approaches to understanding multimedia learning. *Computers in Human Behavior*, *26*(3), 474-481. https://doi.org/10.1016/j.chb.2009.12.006

Xu, G., Ren, T., Chen, Y., & Che, W. (2020). A One-Dimensional CNN-LSTM Model for Epileptic Seizure Recognition Using EEG Signal Analysis. *Front Neurosci*, *14*, 578126. https://doi.org/10.3389/fnins.2020.578126

Yang, S., Yin, Z., Wang, Y., Zhang, W., Wang, Y., & Zhang, J. (2019). Assessing cognitive mental workload via EEG signals and an ensemble deep learning classifier based on denoising autoencoders. *Comput Biol Med*, *109*, 159-170. https://doi.org/10.1016/j.compbiomed.2019.04.034

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: mental workload in ergonomics. *Ergonomics*, *58*(1), 1-17. https://doi.org/10.1080/00140139.2014.956151

Zhang, P., Wang, X., Chen, J., You, W., & Zhang, W. (2019). Spectral and Temporal Feature Learning With Two-Stream Neural Networks for Mental

Workload Assessment. *IEEE Trans Neural Syst Rehabil Eng*, *27*(6), 1149-1159. https://doi.org/10.1109/TNSRE.2019.2913400

4 Multifaceted Approach for Pilot Mental State Detection Based on EEG

4.1 Abstract

The safety of flight operations depends on the cognitive abilities of pilots. In recent years, there has been growing concern about potential accidents caused by the declining mental states of pilots. We have developed a novel multifaceted approach for mental state detection in pilots using Electroencephalography (EEG) signals. Our approach includes an advanced automated preprocessing pipeline to remove artefacts from the EEG data, a feature extraction method based on Riemannian geometry analysis of the cleaned EEG data and a hybrid ensemble learning technique that combines the results of several machine learning classifiers. The proposed approach provides improved accuracy compared to existing methods, achieving an accuracy of 86% when tested on cleaned EEG data. The EEG dataset was collected from 18 pilots who participated in flight experiments and publicly released at NASA's open portal. This study presents a reliable and efficient solution for detecting mental states in pilots and highlights the potential of EEG signals and ensemble learning algorithms in developing cognitive cockpit systems. The use of an automated preprocessing pipeline, feature extraction method based on Riemannian geometry analysis, and hybrid ensemble learning technique sets this work apart from previous efforts in the field and demonstrates the innovative nature of the proposed approach.

4.2 Introduction

The evolution of the aviation industry is heavily dependent on maintaining the highest standards of safety. Advances in aircraft design, endurance, and safety have led to a decrease in the number of aircraft accidents worldwide since the 1960s (Kelly & Efthymiou, 2019). However, operator reliability remains a crucial factor in maintaining flight safety, as flight crews are responsible for a wide range of tasks, including receiving instructions from air traffic control,

interpreting onboard instrument data, making course corrections, briefing cabin crew and passengers, and responding to unexpected events. Operating an airplane requires a high level of mental acuity, and these responsibilities can compromise flight safety (Boksem & Tops, 2008; Hankins & Wilson, 1998; Yen et al., 2009). According to data analyzed by the International Air Transport Association (IATA), there were 45 plane crashes caused by pilots losing control of the aircraft, resulting in 1,645 fatalities between 2012 and 2021 (Association, 2019; *International_Air_Transport_Association*, 2022). Furthermore, the Commercial Aviation Safety Team (CAST) investigated 18 aircraft accidents in which pilots lost control and found that deficiencies in flight crew attention were involved in 16 of the 18 incidents (SKY_Brary, 2019). As a result, CAST recommended that the aviation community, which includes government, business, and academic institutions, conduct research to detect and assess attention-related pilot performance deficiencies (APPD), specifically focusing on channelized attention (CA), diverted attention (DA), and startle/surprise (SS) mental states. CA is a state where pilots engage in a puzzle-based video game called Tetris while remaining focused entirely on the game without paying attention to other tasks. DA is a state in which pilots solve math problems that periodically appear while performing display monitoring tasks. Pilots who are in the SS mental state experience unexpected inversions of the primary flight display in the simulator.

To achieve this goal, researchers from both academia and industry have investigated a variety of approaches based on physiological signals and machine learning (ML) methods. In terms of physiological signals, quantitative sensors, both singular and multiple, have been employed to capture biological signals from the human body in both field studies and near-realistic laboratory environments. The electroencephalography (EEG) sensor is widely regarded as the most crucial physiological signal for analyzing mental states due to its ability to detect transient alterations in brain activity that may be indicative of pilots' attention deficits (Hamann & Carstengerdes, 2022; Hernández-Sabaté et al., 2022; Jiang et al., 2022). It seems to provide the most accurate data for distinguishing mental states. It is also preferable to other methods of brain

monitoring since it is safe, adaptable, non-invasive, and an utterly passive recording technique. Despite its advantages, EEG is notorious for picking up artefacts from environmental factors and physiological phenomena, such as muscle activity, ocular movements, line noise, and heartbeats, which compromise the quality of the signals. Therefore, isolating the neural signal relative to the cognitive processes that reflect brain activity only from the recorded artefacts is crucial.

The presence of artefacts in EEG data can negatively impact the performance of ML models used to detect different mental states of pilots. To address this issue, researchers have employed various signal processing and feature extraction techniques. One approach is to record and combine EEG with non-brain physiological signals, such as functional near-infrared spectroscopy, electrocardiogram (ECG), galvanic skin response (GSR), and respiration (Resp.) simultaneously (Hogervorst et al., 2014; Liu et al., 2017). However, the fusion of features derived from EEG and non-brain physiological signals may not always improve the performance of ML models. Another approach is to utilize traditional preprocessing techniques to handle contaminated EEG data. Visual inspection and rejection, filtering, and Independent Component Analysis (ICA) are examples of such conventional denoising procedures. In Chapter 3, we observed that the implementation of preprocessing techniques such as filtering and ICA to remove eye-related artefacts improved the classification performance of these models in certain cases, compared to using unprocessed, raw EEG data. This highlights the critical role of preprocessing in enhancing data quality for accurate model training. However, these methods, while effective, have several downsides, including the need for manual implementation, being slow and inefficient for longer recording sessions, and being difficult for beginners to execute (Bigdely-Shamlo et al., 2015; Flo et al., 2022).

To overcome these limitations, the development of an automated preprocessing method is crucial. Automated preprocessing refers to the use of algorithms and techniques that can process EEG data with minimal human intervention (Santos

& Ferreira, 2023). The importance of such a method lies in its ability to consistently and efficiently process large volumes of EEG data, reducing the time and expertise required for manual preprocessing. This automation is particularly beneficial for handling long recording sessions, common in pilot monitoring studies. Automated methods can include algorithms for artefact detection and removal, adaptive filtering, and automatic component analysis. By streamlining the preprocessing workflow, these methods can significantly enhance the scalability and accessibility of EEG-based mental state detection.

Features or essential information embedded in the EEG signal are usually extracted after preprocessing as they are crucial for classification tasks (Guler & Ubeyli, 2005; Kordylewski et al., 2001; Übeyli, 2008). Both temporal and spatial features can be extracted from the EEG signals. For pilot mental state classification, temporal features in the time, frequency, and time-frequency domains are commonly extracted (Stancin et al., 2021). One such method that originates in the frequency domain is the power spectrum density. The presence or absence of shifts in the power spectra of individual EEG bands is an important indication of different mental states. In brain-computer interface (BCI) applications, spatial features are commonly extracted. They represent the active area of the brain. For pilot mental state classification, they are rarely used as input.

Features extracted from EEG signals are then fed into an ML model to predict various types of mental states. ML models are trained to distinguish between either binary or multiple classes. Fatigue, workload, stress, and drowsiness are examples of detected mental states in the literature. Most studies have attempted to establish a clear distinction between normal (NE) and each mental state (i.e., a binary classification) or to categorize a single mental state into three or more distinct levels. In addition, only a few studies have focused on assessing and detecting attention-related pilot performance deficiencies (APPD). To the best of our knowledge, no attempts have been made to simultaneously recognize different APPD states (i.e., multi-class classification), particularly CA, DA, SS, and NE, using solely EEG data.

In this work, we aimed to explore the viability of identifying APPD states using publicly released EEG data. Thus, we propose a novel multifaceted approach that decontaminates the EEG signals, extracts meaningful features, and detects the APPD states using heterogeneous cleaned EEG signals collected from 18 pilots. The main contributions of the paper are as follows:

- Development of automatic preprocessing pipeline to automatically repair or remove corrupted EEG data.
- Development of feature extraction and selection methodology based on Riemannian geometry analysis of the cleaned EEG data, that handles the issues of imbalanced dataset and curse of dimensionality and extract meaningful features from the EEG signals.
- Development of a novel APPD system-based hybrid ensemble learning for classifying CA, DA, SS, and NE states.

Recognition of APPD mental states was critically examined using several different ensemble learning algorithms, including Random Forests (RF), Extremely Randomized Trees (ERT), Gradient Tree Boosting (GTB), AdaBoost, and hybrid ensemble learning (Voting).

The remaining sections of this work are structured as follows: In Section 2, we briefly examine relevant works. The existing EEG recordings, the proposed multifaceted approach, and the proposed ML classification models are described in Section 3. In Section 4, we report and discuss experimental findings. Section 6 wraps up the investigation and suggests some directions to explore next in terms of research.

4.3 Related Work

The process of identifying mental states typically involves four steps: collecting data, cleaning it, selecting relevant features, and making predictions. The first step involves capturing signals from the brain and converting them into digital form. Then, to ensure accurate analysis, any extraneous noise or artifacts present in the data are removed through preprocessing. Next, specific

characteristics of the data are selected and extracted in preparation for classification. These extracted features are then used by a classifier to make predictions about which class the data belongs to. As this process specifically relates to EEG data, the following provides a summary of previous research on the three stages of mental state detection: preprocessing, feature extraction, and classification.

4.3.1 Signals Preprocessing

An assortment of neuronal activity, physiological artefacts, and non-physiological noise can be found in raw EEG data. As their presence may hinder the performance of ML models (Cesar Cavalcanti Roza & Adrian Postolache, 2019), identifying and removing artefacts is a crucial preprocessing step before their use (Jiang et al., 2019). Although most research preprocessed their EEG data, there were a few exceptions (Jiao et al., 2018; Terwilliger, 2020; Ziegler et al., 2016). To increase the signal-to-noise ratio (SNR), it is necessary to undertake a preprocessing procedure to eliminate extraneous noise and artefacts.

For the pilot's mental states classification, conventional preprocessing techniques including filtering (Binias et al., 2018; Cesar Cavalcanti Roza & Adrian Postolache, 2019; Han et al., 2020; Jaquess et al., 2017; Johnson et al., 2015; Nittala et al., 2018; Oh et al., 2015; Zhang et al., 2017) and ICA (Han et al., 2020; Wu et al., 2019; Zhang et al., 2017) were employed on the EEG recordings. For example, Roza et al. (Cesar Cavalcanti Roza & Adrian Postolache, 2019) used a band-pass filter with a center frequency of 12-30 Hz to isolate the beta rhythm. Han et al. (Han et al., 2020) used band-pass filtering at 0.1-50 Hz to remove the high frequency prior to removing eyes-related artefacts using the ICA algorithm. Similarly, Alreshidi et al. (Alreshidi et al., 2022) used previously released pilot EEG data to analyze the influence of three preprocessing procedures on the efficiency of two ML models. The results demonstrated no discernible changes in the performance accuracies of the models when the data was filtered or when ICA was applied for eyes-related artefact detection after data filtration. It has been established in the literature

that typical preprocessing procedures for EEG data analysis necessitate knowledge and experience on the part of the analyst. Furthermore, they are only applicable when applied manually, requiring inspection, identification, and removal of faulty channels and contaminated data segments.

The past few years have seen the development of a number of partially or completely automated EEG preprocessing procedures that provide ways to clean EEG data. The Autoreject algorithm is an example of an automated preprocessing procedure that can be employed in EEG analysis pipelines (Jas et al., 2017). It is a novel approach for automatically identifying and repairing erroneous segments in EEG data from single trials. It uses advanced statistical learning techniques such as Bayesian hyperparameter optimization and cross-validation to select amplitude thresholds to use for rejecting or repairing bad segments in EEG data. The Autoreject technique was used by Bonassi et al. (Bonassi et al., 2021) to automatically repair or reject contaminated epochs in EEG data. Pousson et al. (Pousson et al., 2021) preprocessed the EEG data that was recorded from pianists doing musical tasks using the Autoreject method. There was a total of 10% erroneous epochs that were uncovered by the method and subsequently omitted from the investigation. Previous research has established that Autoreject is a significant role in the automatic purification of EEG data.

4.3.2 Feature Extraction

EEG is a sort of stochastic signals that conceals extremely intricate data. Because of its high nonlinearity, its features are prone to sudden fluctuations. Human mental states, however, transition gradually from one state to the next (Wang et al., 2014). Feature extraction aims to extract relevant features from data to map EEG segments to mental states.

Various features such as statistical (Cesar Cavalcanti Roza & Adrian Postolache, 2019; Harrivel et al., 2017; Nittala et al., 2018) and power spectral density features (Cesar Cavalcanti Roza & Adrian Postolache, 2019; Han et al., 2020; Harrivel et al., 2016; Harrivel et al., 2017; Jaquess et al., 2017; Johnson

et al., 2015; Nittala et al., 2018; Wu et al., 2019; Zhang et al., 2017; Ziegler et al., 2016) have been extracted from pilots' EEG recordings in earlier research in order to classify pilots' mental states. For example, Wu et al. (Wu et al., 2019) used the power spectrum curve area representation of the decomposed delta, theta, alpha, and beta brain waves obtained using wavelet packet transform as features to perform the classification. Roza et al. (Cesar Cavalcanti Roza & Adrian Postolache, 2019) derived 15 distinct features from EEG and other physiological signals. The wavelet coefficients and several statistical features were extracted from the EEG signals. Furthermore, Binias et al. (Binias et al., 2018) extracted logarithmic band-power features using common spatial pattern (CSP) spatial filtering, which is widely used in BCI applications, from pilots' EEG recordings.

There has been a recent uptick in the use of Riemannian geometry-based feature extraction and classification algorithms for BCIs. Riemannian geometry offers a framework for analysing data lying on curved spaces, like covariance matrices of EEG signals. The first implementation of these techniques in BCI applications was published in (Barachant et al., 2012) , where the authors employed the Riemannian mean covariance matrix distance as a feature for classification purposes. This distance, mathematically expressed as $d(R_1, R_2) = \left\|\log(R_1^{-1/2} R_2 R_1^{-1/2})\right\|$, where $R_1$and R_2 are covariance matrices, measures the dissimilarity between two points (matrices) on the Riemannian manifold.

Additionally, Barachant et al. demonstrated how covariance matrices can be represented as vectors in the tangent space of the Riemannian manifold, a process crucial for applying conventional ML algorithms. Majidov and Whangbo (Majidov & Whangbo, 2019) computed the covariance matrices obtained by using Common Spatial Pattern (CSP) spatial filtering, a technique for enhancing SNR in multichannel EEG data. The CSP algorithm projects the EEG data onto a set of spatial filters, mathematically defined as the solution to the optimization problem: maximize $v^T C_1 v$ subject to $v^T C_2 v = 1$, where v is the spatial filter, and C_1 and C_2 are the covariance matrices of two classes. These matrices were then mapped onto the tangent space of the Riemannian manifold for classification.

Singh et al. (Singh et al., 2019) used the data from the EEG electrodes to create spatial filters that reduce the dimensionality prior to employing Riemannian distance as a pattern recognition metric for classification. In addition, classifiers based on Riemannian geometry were used by Appriou et al. (Appriou et al., 2021) in the proposed BioPyC toolbox. One such classifier is the tangent space classifier.

4.3.3 Mental State Classification

After EEG signals have had their features extracted, they must be classified using either a binary or multiclass ML approach. Because of the increased efficiency with which neural data may be analysed and the need to decode brain activity, ML and particularly Deep Learning (DL) algorithms have found widespread use in the field of computational neuroscience. Supervised ML algorithms, for instance, must first be trained using example data. The model and its learnt properties are then used to make predictions about the class label of new data that has not yet been seen.

For the detection of pilot various mental states, previous studies implemented various ML (Avots et al., 2022; Binias et al., 2018; Dehais et al., 2019; Han et al., 2020; Harrivel et al., 2016; Harrivel et al., 2017; Johnson et al., 2015; Nittala et al., 2018; Oh et al., 2015; Zhang et al., 2017; Ziegler et al., 2016) and DL (Binias et al., 2018; Cesar Cavalcanti Roza & Adrian Postolache, 2019; Han et al., 2020; Harrivel et al., 2016; Hernández-Sabaté et al., 2022; Wu et al., 2019; Zhang et al., 2019; Ziegler et al., 2016) algorithms. For instance, Han et al. (Han et al., 2020) proposed a detection system based on multimodal physiological signals including EEG, ECG, Electrodermal Activity, and Respiration. They transformed the EEG signal into topographical images for training in two dimensional convolutional neural network (2D-CNN), while the non-brain data were treated as numerical time series and trained in a long short-term memory (LSTM) network. This system was designed to detect mental states such as distraction, workload, fatigue, and normal state in pilots. Roza et al. (Cesar Cavalcanti Roza & Adrian Postolache, 2019) proposed an emotion recognition system based on multimodal physiological signals, namely

EEG, Galvanic Skin Response (GSR), and heart rate, and artificial neural network (ANN). The system was developed to detect 5 emotional states, namely happy, sad, angry, surprise and scared. To identify the various states of mental fatigue, Wu et al. (Wu et al., 2019) presented a deep contractive autoencoder network; up to 91.67 percent of cases of the fatigued mental status of pilots could be correctly identified. In a flight simulator experiment, Johnson et al. (Johnson et al., 2015) investigated probe-independent methods for categorization three layers of task-complexity. The investigation was carried out using six classification algorithms, namely naïve bayes, decision trees, quadratic discriminant analysis, linear discriminant analysis (LDA), k-nearest neighbours (KNN), and support vector machine (SVM). Dehais et al. (Dehais et al., 2019) devised a scenario in which twenty-two pilots using a six-dry-electrode EEG system performed a low-load and high-load traffic pattern, as well as a passive auditory oddball. Zhang and Wang (Zhang et al., 2017) proposed a concatenated structure of deep recurrent and 3D CNN to learn spatial-spectral-temporal EEG features for cross-task mental workload assessment. The findings reveal that the proposed approach achieved an average accuracy of 88.9%. Distinguishing between stages of brain activity related to idle but concentrated anticipation of visual cue and reaction to it using LDA, KNN, SVM, RF, and ANN algorithms was the focus of Binias et al. (Binias et al., 2018) research.

Detecting and assessing APPD was also addressed in previous studies. For example, Harrivel et al. (Harrivel et al., 2016) employed RF, extreme gradient boosting, and deep neural network classifiers to predict CA, DA, and low workload states. As preliminary study, through the use of different sensing modalities in high-fidelity flight simulators, the authors classified three types of mental states. Harrivel et al. (Harrivel et al., 2017) employed RF, gradient boosting, and two SVM classifiers to identify CA and SS states in further studies. The authors recommended that the data quality issues need to be addressed. The authors stressed the need of addressing the data quality issues. Terwilliger et al. (Terwilliger, 2020) aggregated 3 mental states classes, namely CA, DA, and SS, into one class called event. To distinguish the event

class from NE mental state class, the authors presented a convolutional autoencoder approach. In previous research, we examined the effects of two preprocessing procedures on SVM and ANN using EEG data from a pilot exposed to CA, DA, SS, and NE states (Alreshidi et al., 2022). Although the models demonstrated the viability of combining data from two scenarios, the curse of dimensionality prevented them from accurately predicting the DA and SS states.

In the field of aviation, several studies have been conducted to evaluate the efficacy of EEG data in predicting mental states of pilots. Some of these studies have employed a binary classification approach to detect different mental states, while others have utilized EEG data in combination with other physiological data to improve performance. However, a notable limitation of previous studies is the limited sample size, with many only incorporating EEG data from less than 10 participants. This raises questions regarding the generalizability of their results, as the findings may only be applicable to a small subset of the population. While incorporating additional signals can sometimes improve model performance, it can also introduce additional noise and complexity to the system, making it more challenging to interpret the results.

Additionally, some studies have not disclosed the necessary information to make their work easily reproducible, while others have failed to make their datasets publicly available. This makes it challenging for other researchers to verify or build upon their findings. Furthermore, some studies have not performed proper preprocessing techniques on their EEG data, such as advanced filtering and artefact removal, potentially compromising the validity of their results. The noise can interfere with the extraction of meaningful features and patterns in the EEG signal, leading to a decrease in the accuracy and reliability of the resulting model. To minimize the impact of noise on the performance of ML techniques, it is important to preprocess the EEG signal and remove as much noise as possible before training the model. Researchers have hardly ventured beyond statistical and PSD features in their pursuit of meaningful feature extraction. Regarding the classification of APPD states,

current research has, to the best of our knowledge, not attempted a multiclass classification of CA, DA, SS, and NE.

The innovative nature of this study lies in the development of a novel multifaceted approach to detect and classify APPD states using heterogeneous cleaned EEG data. EEG signals from 18 pilots were collected from a variety of conditions to form the heterogeneous EEG data. The approach involves the automatic preprocessing of the EEG signals, feature extraction and selection methodology based on Riemannian geometry analysis, and a novel APPD system that classifies the APPD states. The system addresses the issues of corrupted EEG data, imbalanced dataset, and curse of dimensionality, and provides meaningful features from the EEG signals, making it a unique contribution to the field.

4.4 Materials and Methods

4.4.1 Dataset Description

In November 2020, a dataset was obtained from NASA's open data portal website, which comprised experimental data collected from 18 pilots. The pilots participated in four experiments, three of which took place in a non-flight environment and one in a high-fidelity motion-based flight simulator. The non-flight environment experiments lasted approximately 6 minutes, while the flight simulator experiment lasted approximately 1 hour. The benchmark tasks included NE, CA, DA, and SS, with a typical snapshot and schematic of each experiment depicted in Figure 4-1. The dataset utilized in this study is a composite of the recordings gathered from the 18 pilots in both flight and non-flight environments. This approach was adopted to enrich the dataset's diversity and volume. As delineated in Chapter 3, the inclusion of non-flight data was proven feasible and beneficial for the comprehensiveness of the research. Information regarding the utilized EEG recording headset and the flight simulator is reported in Appendix A and Appendix B.

The dataset was segmented into one-second epochs for several key reasons. Firstly, this duration aligns well with the temporal dynamics of mental states,

some of which manifest within a one-second timeframe. It allows for capturing these brief yet significant fluctuations in cognitive states. Secondly, segmenting into one-second epochs increases the dataset size, creating a larger pool of samples that is beneficial for the training of ML models. This approach enhances the robustness and generalizability of the models by providing a more diverse range of data points for learning.

This dataset has great potential for advancing research in the fields of BCI and human factors in aviation and can be used to develop new models and algorithms to predict pilot performance under different conditions, as well as training programs to improve pilot performance in high-stress situations. Additionally, the dataset can be utilized to evaluate the design of flight deck interfaces and test the effectiveness of new technologies, such as augmented reality and virtual reality, in enhancing pilot performance.

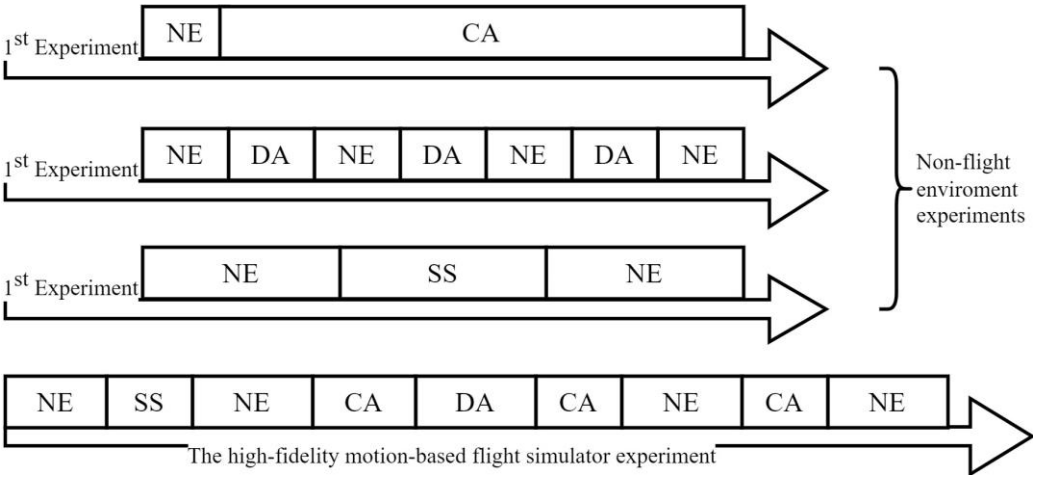

Figure 4-1 A typical snapshot and schematic of each experiment

4.4.2 The Automatic Preprocessing Pipeline

This study implemented advance preprocessing techniques using an open-source library called MNE-Python. The proposed EEG data preprocessing pipeline is shown in Figure 4-2. A brief description of the preprocessing steps is discussed below.

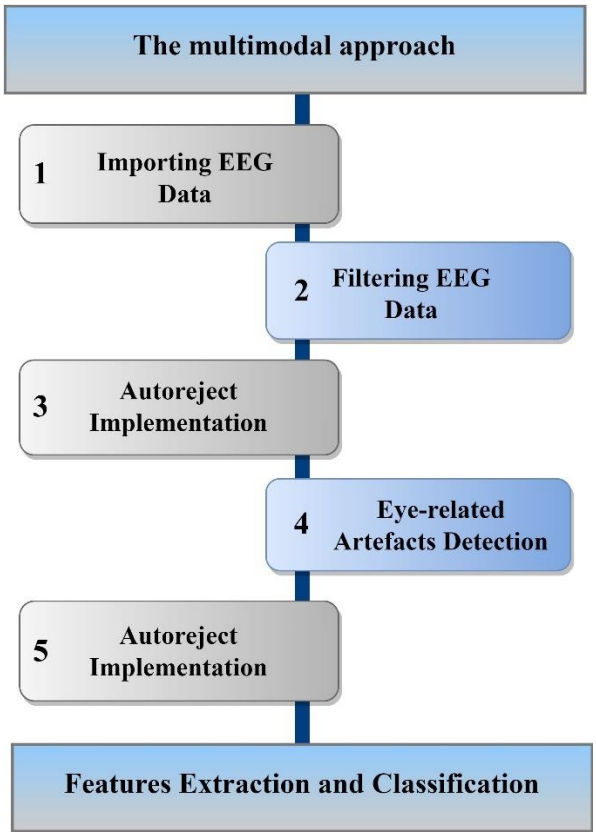

Figure 4-2 An outline of the multifaceted approach based on EEG

The EEG data was given in a CSV file. We used the MNE-Python library to apply advanced preprocessing methods. A "raw" object, a core data structure for continuous EEG data, was created and included information such as channel names and types, standard montage labeling, and the sample rate.

The first step is to filter the EEG signals. This is achieved by applying a digital filter to the data, which suppresses specific frequency components that fall outside of a designated range. There are two main types of digital filters used in digital signal processing (DSP): finite impulse response (FIR) and infinite impulse response (IIR). In the present study, we applied band-pass filtering to the EEG signals using an FIR filter, with a cutoff range of 1-50 Hz. We then segmented the EEG data into one-second non-overlapping epochs. The epochs that have a maximum peak-to-peak signal amplitude of more than 700 µV or a minimum peak-to-peak signal amplitude of less than 1 µV were dropped from the dataset as their existence negatively affect the applicability of the next preprocessing steps. Afterward, we employed the Autoreject method to repair or

discard corrupted epochs. Bayesian optimization and cross-validation are leveraged in Autoreject to automatically determine an artefact threshold for each channel/sensor; thereafter, faulty channels/sensors are interpolated, or the epoch is discarded. Figure 4-3 is a diagram depicting the operation of the Autoreject algorithm in a simplified form. For a detailed discussion of how and why this algorithm works, we suggest reading (Jas et al., 2017), written by the program's creators. To identify and eradicate blinks and other forms of artifactuality, we employed an MNE-Python function that used the EEG channel Fp1 as a surrogate electrooculogram. These components have a lot of variation and tend to be located in the frontotemporal region of the head. The EEG signals were reconstructed after the blinking component was eliminated from the source matrix. Finally, we used Autoreject again to encounter any distortions that could be found after repairing the blink artefacts.

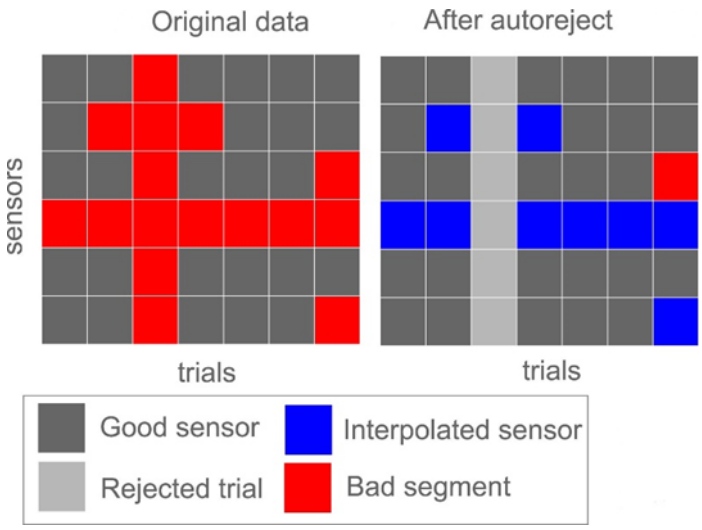

Figure 4-3 A simplified form of the Autoreject algorithm operation

With more than 80% of the data coming from the NE class, it’s possible that the trained model will be biased toward that class. This makes a model's predictions seem naive, even if they have a high degree of accuracy. To counteract the preponderance of the NE class, we undersampled the data with the intention of creating a more even distribution across all classes.

4.4.3 EEG Feature Extraction

After preprocessing the EEG data, two methods expanded upon previous work on EEG BCI were adopted. First, the EEG data is subjected to specialised spatial filtering in order to boost SNR. We used an algorithm modified from the xDawn algorithm to estimate the spatial filters. Second, we extracted the features from a particular form of the EEG epochs' covariance matrices and adjusted them using techniques from Riemannian geometry. Indeed, the covariance matrices, being Symmetric and Positive-Definite Matrices (SPD), are topologically localised on a Riemannian manifold. To reduce the covariance matrices dimensionality by discarding irrelevant information, we performed the Fisher Geodesic Discriminant Analysis (FGDA) algorithm proposed by (Barachant et al., 2010; Barachant et al., 2012). Be aware that the features are matrices, rather than the typical vectors. Because we need to maintain the special structure of these matrices, we cannot simply vectorize them. As an alternative, we employed techniques from Riemannian geometry introduced in (Barachant et al., 2013) to map the covariance matrices, belonging to a manifold, onto the Riemannian tangent space, where they may be vectorized and treated as Euclidean objects. Each matrix is represented as a vector of size *n(n+1)/2*, where *n* is the dimension of the SPD matrices.

In Figure 4-4, we provide a visual representation of the tangent space mapping process utilized in this study. The manifold M represents the set of all covariance matrices, with each point, such as P, corresponding to a specific covariance matrix that is SPD. The tangent space at point P , denoted T_P, is a flat space that approximates the manifold locally around P. In this space, operations of vector calculus can be applied. The logarithm map, denoted as $\log_P(.)$, projects a point P_i from the manifold to a tangent vector S_i in T_P, while the exponential map, denoted as $Exp_P\ (.)$, performs the inverse operation, mapping a tangent vector S_i back to a point on the manifold. The geodesic $\Gamma(t)$ is the shortest path on the manifold between two points and is central to defining these mappings. The derivative of $\Gamma(t)$ at $t = 0$ gives us the tangent vector S_i. The transformation equations are as follows:

$$Exp_P(S_i) = P^{\frac{1}{2}} exp(\, P^{-\frac{1}{2}} S_i P^{-\frac{1}{2}})\, P^{\frac{1}{2}} \quad \textbf{(4-1)}$$

$$Log_P(S_i) = P^{\frac{1}{2}} Log(\, P^{-\frac{1}{2}} P_i P^{-\frac{1}{2}})\, P^{\frac{1}{2}} \quad \textbf{(4-2)}$$

These equations allow us to transition between the non-Euclidean geometry of the manifold and the Euclidean geometry of the tangent space, which is crucial for applying ML algorithms that require vector space representations. Furthermore, in our study, we applied Principal Component Analysis (PCA) and ANOVA methods to the vectors in the tangent space to reduce dimensionality, thus alleviating the curse of dimensionality commonly encountered in high-dimensional data analysis.

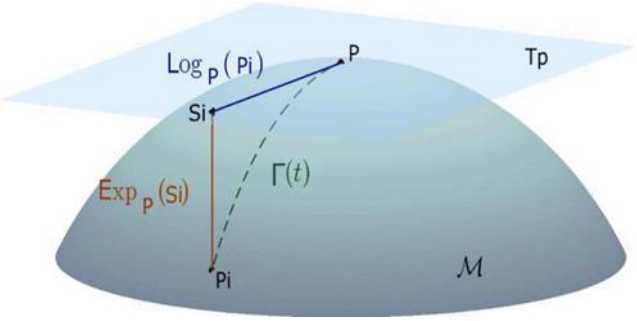

Figure 4-4 A geometric depiction of the tangent space mapping process

In the context of this study, the decision to employ a modified xDawn algorithm and the FGDA was informed by the specific requirements of EEG data analysis in pilot mental state detection. The modified xDawn algorithm was chosen for its proven efficiency in enhancing EEG data's SNR, a vital step for ensuring data quality. More critically, FGDA was selected over other dimension reduction or manifold learning techniques due to its unique suitability for handling EEG covariance matrices, which are SPD and thus lie on a Riemannian manifold. FGDA not only facilitates effective dimensionality reduction but also preserves the intrinsic geometric structure of the data, a crucial aspect for maintaining the fidelity of EEG features. This approach, therefore, offers a more precise and theoretically sound framework for EEG data processing, aligning with the specific challenges and nuances of the dataset at hand, compared to other

methods that might not adequately address these specialized aspects of EEG data.

4.4.4 EEG Classification

In this study, we rigorously tested multiple ensembles learning algorithms, including Random Forests (RF), Extremely Randomized Trees (ERT), Gradient Tree Boosting (GTB), AdaBoost, and Voting, for their ability to recognise APPD mental states. A modified version of the 5-fold cross-validation process based on stratification was used to assess the quality of the proposed approach.

The 5-fold cross-validation method is a commonly employed technique in ML to assess the performance of algorithms. The method involves dividing the original data set into five equal-sized subsets, referred to as folds. In turn, each fold serves as the validation data once while the remaining four folds are utilized as training data. This process is repeated five times with each fold being used exactly once as the validation data. The performance of the algorithm is then evaluated based on the average of the results obtained from the five trials. This approach to evaluating performance provides a more reliable estimate compared to a single train/test split. This is due to the reduction of variance in performance estimates and the assurance that all data is utilized for both training and testing.

RF. In 2001, L. Breiman presented the random forest algorithm as a general-purpose classification and regression technique, and it has since seen tremendous success (Breiman, 2001). The method has shown effective in situations when there are more variables than observations, as it mixes multiple randomised decision trees and averages their predictions. It can be scaled up to address complex issues, customized to meet the needs of a wide range of ad hoc learning projects, and designed to yield metrics of varying significance. The entropy function was used as a metric of split quality in our work, with the number of estimators fixed at 200.

ERT. It is a classifier that works in a way that's similar to RF, but with a slight twist: it introduces randomization to the training process (Geurts et al., 2006).

Each tree in ExtraTrees's multiple trees is trained independently using the entire dataset used for the classification. The optimum branching at a node is determined by considering a subset of all features, much like the Random Decision Forest. Each feature has a single threshold picked at random rather than multiple, less optimal ones. In our research, we used a total of 200 estimators and the entropy function to evaluate split quality.

GTB. It provides a prediction model in the shape of a collection of weak prediction models, most often decision trees (Friedman, 2001). GTB is the name of the resulting procedure when a decision tree is the weak learner. The method extends the boosting algorithm to any loss function that can be differentiated. In our study, split quality was assessed using the 'friedman_mse' function and a total of 100 estimators.

AdaBoost. The statistical classification meta-algorithm known as Adaptive Boosting, was developed by Yoav Freund and Robert Schapire in 1995 (Schapire, 2003). Its performance can be enhanced by combining it with a variety of different learning methods. This method creates a model in which each piece of information is given the same amount of consideration. Incorrectly labelled points are thus given more weight. After this new model is created, the points with greater weights will be given more consideration. A model will be trained repeatedly until a reduced error is received. Because of its rapid convergence to a smaller test error after fewer boosting iterations, the 'SAMME.R' method was chosen in our research.

The hybrid model (Voting). The goal is to predict class labels using a majority vote or the average projected probability (soft vote), based on the results of a collection of ML classifiers that are conceptually distinct from one another. A classifier like this can help even out the performance of a group of otherwise comparable models. Based on the outcomes of RF, ERT, and GTB, we used the average projected probability to make predictions about class labels.

4.5 Results and Discussion

In this study, a multifaceted approach was proposed to identify attention-related pilot performance-limiting states based on heterogeneous EEG data. We employed an automated preprocessing pipeline to clean the EEG data by either removing or repairing corrupted epochs. We employed an extraction and selection methodology based on Riemannian geometry analysis to obtain meaningful features from the cleaned data. Using these extracted features, we trained a hybrid ensemble learning model in addition to four other ensemble learning models to detect APPD states.

4.5.1 EEG Signal Analysis

This section presents and discusses the results of employing the automated preprocessing pipeline. Figure 4-5 reveals the size of the dataset before and after preprocessing the dataset.

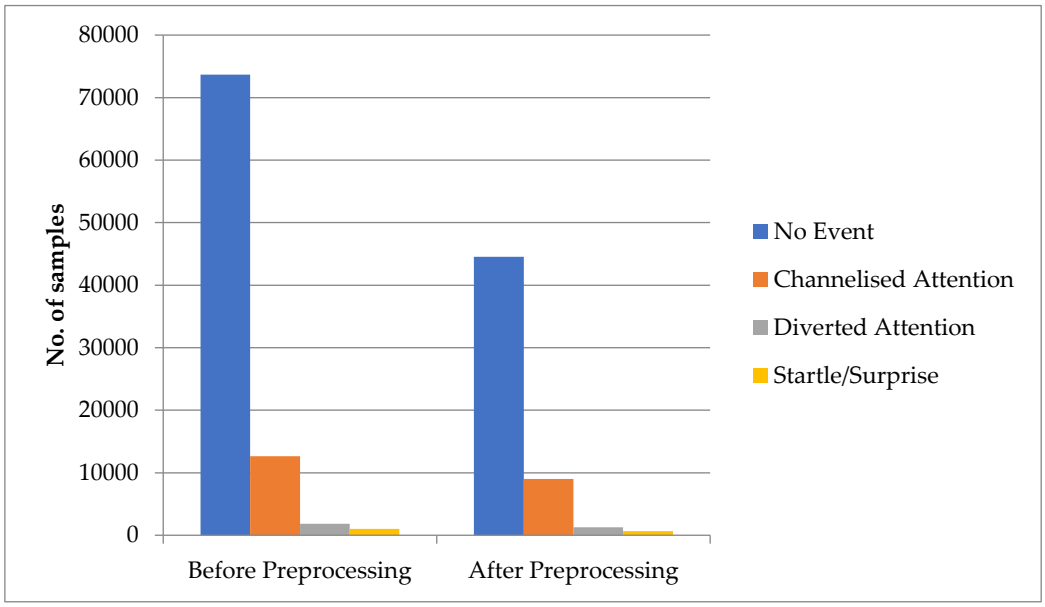

Figure 4-5 The size of the dataset before and after preprocessing the dataset

We observed that the proposed pipeline identified and discarded a total of 33786 contaminated epochs in the dataset; to be precise, 29175 epochs from NE class, 3632 epochs from CA class, 598 from DA class, and 381 epochs from SS class were dropped from the dataset as they were considered artefacts.

The proposed EEG preprocessing pipeline aims to improve the quality of EEG data by removing artifacts and other sources of noise, ultimately leading to more accurate and reliable results in downstream analyses. The employed pipeline removed 33,786 out of 89,198 epochs were recorded, resulting in a final dataset of 55,412 epochs. While some may argue that removing such a large number of epochs may lead to a loss of valuable data, it is important to consider the rationale behind the preprocessing steps and the impact they have on the quality of the remaining epochs.

While visually inspecting the discarded epochs, we observed that the epochs were contaminated by physiological artefacts such as muscles tension and clenching the jaw and non-physiological / technical artifacts such as body movements and powerline interference. As an illustration, Figure 4-6 (A) depicts an 8-epoch window of the original EEG data, whereas Figure 4-6 (B) depicts an 8-epoch window of the EEG data that has been preprocessed using the preprocessing pipeline. Figure 4-6 (A) reveals that ocular activity artefacts such as blinks and lateral eye movements were spotted and colour-coded as red in epochs 15, 18, and 20. These three epochs were deleted in addition to epochs 19, 21, and 25 as indicated in Figure 4-6 (B). We also noticed that some epochs, epoch 16 for instance, were repaired.

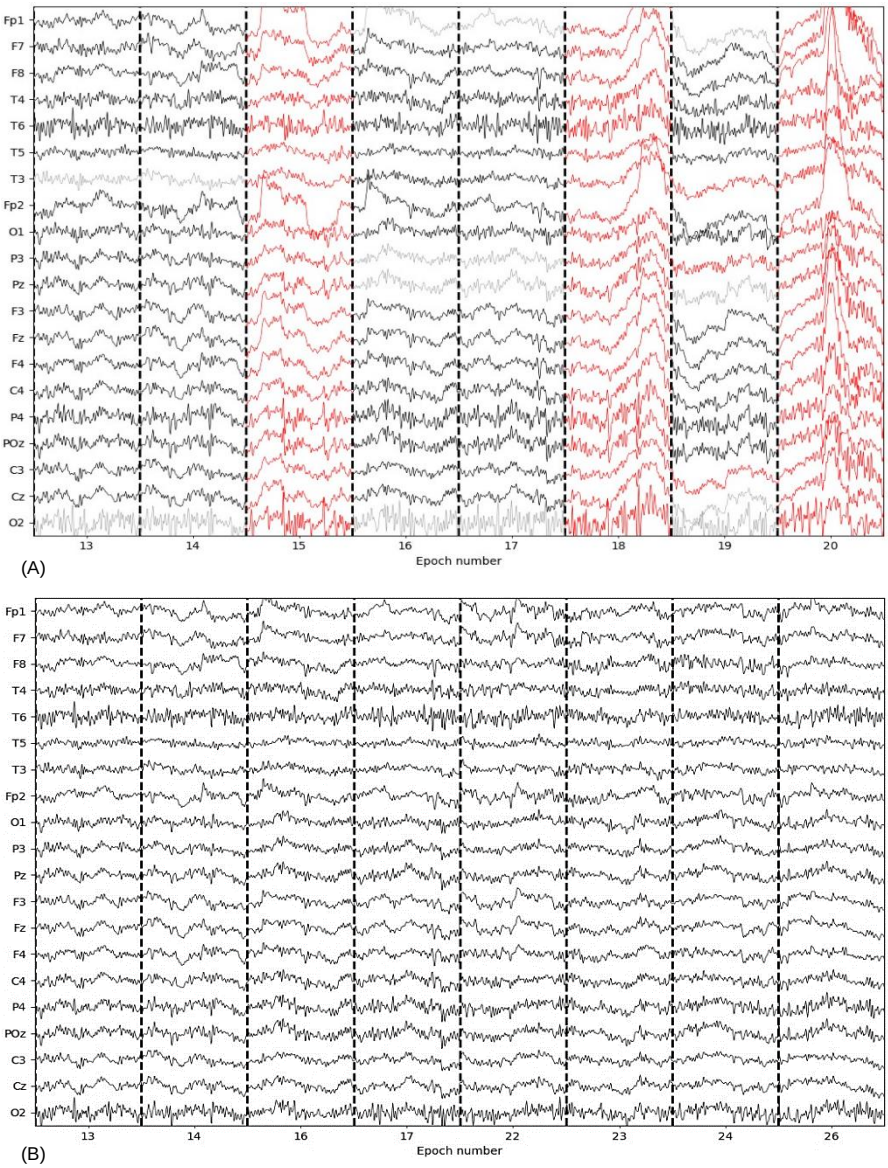

Figure 4-6 An 8-epoch example of the EEG signals before and after preprocessing.

Based on the results presented, the EEG preprocessing pipeline appears to be effective in improving the quality of the EEG data. The visual comparison of the EEG signal before and after preprocessing indicates a reduction in noise and artifacts, resulting in a cleaner and more consistent signal.

The use of Autoreject for artifact rejection and correction, followed by eye-related artefacts removal, and a second stage of Autoreject for further correction, provides a comprehensive approach to minimizing the impact of

artefacts on the EEG signal. The use of these methods in combination is likely to capture a wide range of artefacts and improve the overall quality of the data.

The effectiveness of the pipeline is also supported by the quantitative analysis of the EEG data. For example, the reduction in the number of epochs removed after preprocessing may indicate that the pipeline was successful in identifying and removing a significant proportion of the artifacts. Furthermore, the comparison of the EEG data before and after preprocessing may provide evidence of the improvements made in the EEG data quality.

However, it is important to note that the effectiveness of the pipeline may depend on various factors, such as the quality of the initial EEG data and the parameters used for each stage of the pipeline. Therefore, a careful evaluation of the resulting EEG data and the quality of the analysis should be conducted to determine the overall effectiveness of the pipeline.

In addition, while the use of automated methods for artefact detection and correction can provide several advantages, such as consistency and efficiency, they may not capture all sources of noise and artifacts. Therefore, it may be beneficial to supplement the automated methods with visual inspection, especially in cases where subtle sources of noise may be present.

We also report the spectral power analysis of one pilot while performing the high-fidelity motion-based flight simulator experiment to examine the overall activity level of the brain at different frequencies. Figure 4-7 illustrates the spectral power topography during APPD mental states, namely A) NE, B) SS, C) CA, and D) DA. The power spectral density was computed for each frequency band (delta (0-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), beta (12-30 Hz), and gamma (30-45 Hz)).

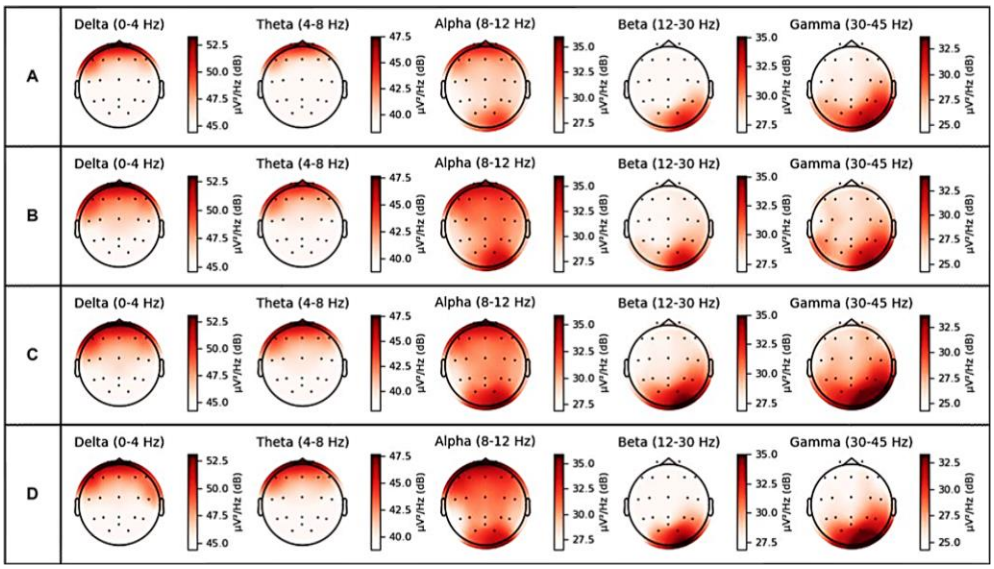

Figure 4-7 Spectral power topography during APPD mental states, namely A) NE, B) SS, C) CA, and D) DA.

In all frequency bands, we commonly found an increase mean power of CA, DA, and SS states compared to NE state. We also observed a lower frequency power increase in all frequency bands ranges during SS state. For the delta activity, the highest mean spectral power was located in the frontal lobe during CA and DA states. For the theta and alpha activity, the highest spectral power was observed in the frontal lobe for theta activity (max: 47.5 dB) and in the frontal and occipital lobes for alpha activity (max: 36.7 dB) during the DA state. Theta oscillations have been linked to mental states of relaxation and drowsiness, while alpha oscillations have been associated with decreased cognitive engagement and mind-wandering. For the beta (max: 33.3 dB) and gamma activity (max: 33 dB), the highest spectral power was observed in the occipital lobe during the CA state. Both beta and gamma oscillations have been connected to engaged cognitive processing, including perception and memory, while beta oscillations have been associated with focused attention and concentration.

Spectral power analysis is a well-established method for analyzing EEG data that has been used in many studies to investigate the spectral properties of the EEG signal. In our study, we used spectral power analysis to visualize the

topography of EEG activity during four different mental states – CA, DA, SS, NE. By calculating the power spectral density of the EEG signal in different frequency bands, we were able to obtain topographical maps that showed the distribution of power across the scalp. These maps provided a global view of the EEG patterns that were associated with each mental state, and allowed us to identify the scalp regions that exhibited the strongest or weakest power in different frequency bands. This information was useful in identifying patterns of EEG activity that were associated with each mental state, and in validating the results of our subsequent classification analysis. Thus, the use of spectral power analysis was essential to achieving the primary objective of our study, which was to gain a better understanding of the EEG patterns underlying the four mental states.

4.5.2 Evaluation of Machine Learning Models

Five ensemble learning models, namely RF, ERT, GTB, AdaBoost, and Voting, were trained with tangent space features generated from cleaned EEG data using 5-fold cross-validation technique. First, we estimated the spatial covariance matrices from the cleaned EEG data and obtained a set of SPD matrices of shapes (48,48). Each matrix was vectorized obtaining 1176 tangent space features which then projected to a lower dimensional space using PCA. In Table 4-1, we show the performances of the employed ensemble learning models. We considered the Macro average of evaluation metrics Accuracy, Recall, Precision and F1-score. We also show the standard error (SE) in parentheses for each class because we trained the models using 5-fold cross validation technique. The SE was calculated based on Recall, Precision and F1-score.

Table 4-1 Ensemble learning models' performances. SE is provided in parentheses.

Methods	Mental Class	Accuracy	Precision (SE)	Recall (SE)	F1-score (SE)
	NE		91.43% (0.010)	91.82% (0.011)	91.11% (0.010)
RF	SS		81.65% (0.014)	81.22% (0.006)	81.87% (0.009)
	CA		86.69% (0.022)	86.40% (0.015)	86.99% (0.013)

	DA	81.60% (0.020)	82.88% (0.015)	83.21% (0.011)
	Macro average 86.48%	86.40%	85.92%	86.30%
ERT	NE	89.62% (0.012)	91.03% (0.013)	90.00% (0.011)
	SS	79.96% (0.022)	80.08% (0.012)	80.05% (0.016)
	CA	85.81% (0.016)	85.32% (0.010)	86.09% (0.010)
	DA	80.66% (0.020)	82.00% (0.014)	81.94% (0.012)
	Macro average 83.72%	83.89%	83.72%	83.64%
GTB	NE	91.22% (0.015)	90.03% (0.020)	91.11% (0.016)
	SS	81.69% (0.012)	82.31% (0.010)	82.16% (0.009)
	CA	87.22% (0.020)	87.42% (0.017)	87.36% (0.012)
	DA	83.15% (0.020)	83.80% (0.016)	83.47% (0.011)
	Macro average 85.80%	85.63%	85.82%	85.76%
AdaBoost	NE	90.91% (0.013)	88.28% (0.013)	89.44% (0.009)
	SS	79.61% (0.015)	80.41% (0.017)	80.01% (0.007)
	CA	83.33% (0.017)	82.11% (0.021)	83.16% (0.010)
	DA	79.19% (0.030)	79.95% (0.020)	79.62% (0.023)
	Macro average 82.53%	82.51%	82.53%	82.52%
Voting	NE	91.07% (0.016)	92.18% (0.014)	91.59% (0.013)
	SS	82.11% (0.012)	82.42% (0.007)	82.27% (0.009)
	CA	86.88% (0.023)	86.34% (0.014)	86.55% (0.012)
	DA	82.92% (0.020)	83.78% (0.015)	83.28% (0.013)
	Macro average 86.26%	86.15%	86.25%	86.19%

To provide thoroughly analysis, the degree of confusion generated by each model was computed. The confusion matrix for 5-fold cross-validation results using RF classifier is shown in Figure 4-8 (A); The ERT was employed in (B), GTB in (C), AdaBoost in (D), and Voting in (E). The values of the diagonal elements represent the percentage of correctly predicted classes.

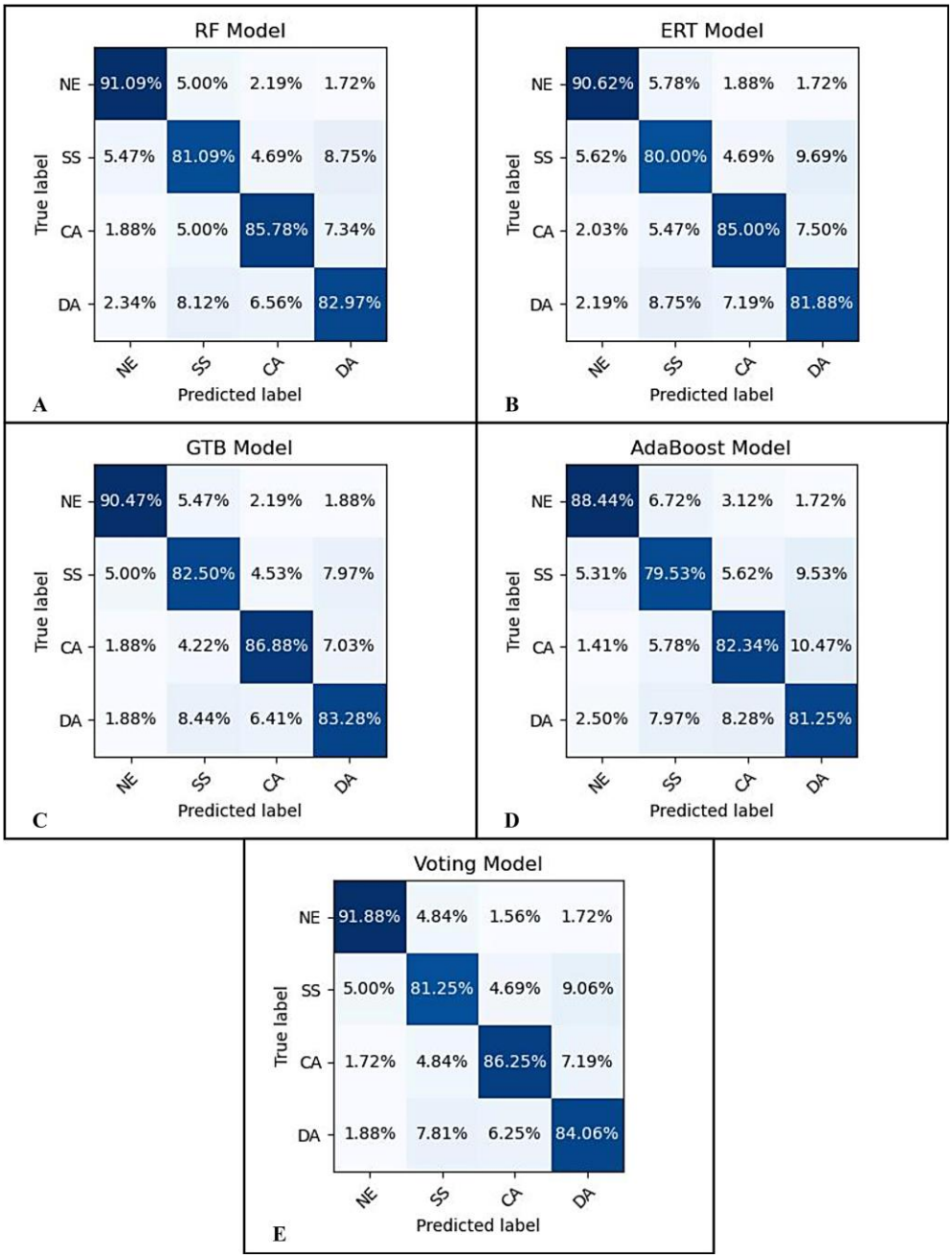

Figure 4-8 The confusion matrix for 5-fold cross-validation results. The RF model's confusion matrix is shown in (A); the ERT in (B), GTB in (C), AdaBoost in (D), and Voting in (E).

Based on the data from Table 4-1, we observed that all 5 models provided good detection performances. The best accuracy performance achieved was 86% which achieved by RF, GTB, and Voting models, followed by AdaBoost (84%) and ERT (83%). The same trend can be seen across different metrics including precision, recall, and F1-score. We believe the reason why ERT did not perform

as well as the RF model, although both algorithms are based on the bagging or bootstrap aggregation technique, is because of the randomness in the way splits are computed; while the most discriminative thresholds are picked as the splitting rule in RF, thresholds in ERT are drawn at random which slightly increased biasness in the model. Similarly, we also observed a slight difference in performances of GTB and AdaBoost even though both algorithms are based on the boosting technique. We suspect the reason of the increase in GTB model performance is due to the use of the log-loss loss function which is more robust to mislabeled examples in the dataset; unlike GTB, AdaBoost algorithm uses the exponential loss function.

Figure 4-8 further shows that all models made accurate classification predictions. The NE mental state was predicted by all five models to be the easiest to distinguish, with an accuracy performance range of 88.44%–91.88%, followed by the CA with a range of 82.34%–86.88%. It was also discovered that across all five models, DA was the third best at recognizing class with an accuracy performance of 81.25%-84.06%, while SS was the worst at recognizing class with an accuracy performance of 79.53%-82.50%. Nevertheless, these performances levels can be enhanced if the dataset is more cohesive. With regards to predicting NE and DA states, the Voting classifier performed best, whereas the GTB classifier performed best with regards to predicting CA and SS states.

The use of ensemble models has become increasingly popular in ML due to their ability to leverage the strengths of different models to improve performance. In this study, we compared the performance of several popular ensemble models, including RF, ERT, GTB, and AdaBoost, with a hybrid ensemble model. The results showed that the hybrid ensemble model outperformed ERT and AdaBoost and achieved comparable performance to RF and GTB. One of the key advantages of the hybrid ensemble model is its flexibility. By combining different models, the hybrid ensemble approach can handle various types of data and tasks, making it a versatile option for different applications. In contrast, the other models tested in this study were each based

on a single algorithm, limiting their flexibility to some extent. Another advantage is its improved generalization ability. The use of a combination of models in the hybrid ensemble approach can help to mitigate the risk of overfitting. This can lead to more accurate predictions on new, unseen data, making the hybrid ensemble model a promising approach for practical applications.

Several studies have investigated the classification of mental states using EEG data. However, some of these studies did not make their dataset publicly accessible, did not achieve clear or consistent results, employed different sensors and conventional preprocessing techniques, or did not classify the same number of mental states. In order to compare the results of our multifaceted approach with other studies, we evaluated our approach in the context of studies that have used the same dataset.

Harrivel et al. (Harrivel et al., 2016) implemented a broad suite of sensors to classify pilot mental states. Although this study provided initial insights into the use of physiological signals to measure attention in aviation, their datasets were limited in size. In addition, their results were not conclusive because it was based on only one pilot. Harrivel et al. (Harrivel et al., 2017), on the other hand, considered a larger sample size and employed multiple sensors including EEG, ECG, GSR, and respiration. However, the study relied on spectral power features and did not classify 4 mental states. Moreover, the results were not as good as in our study, likely due to the limited classification capabilities of spectral power features. Similarly, (Terwilliger, 2020) considered a larger sample size of 18 users, but did not clean their data from artifacts and merged three mental states into one called event state. The lack of artifact removal may have contributed to unclear results and the use of different metrics limited comparison with our study.

we also evaluated our approach in the context of studies that have used the different dataset. For example, Han et al. (Han et al., 2020), proposed a multimodal deep learning network to classify four mental states (Distraction, baseline, workload, and fatigue) using a dataset of eight pilots. The authors employed conventional preprocessing techniques, including filtering and ICA for

removing eye-related artifacts. They also extracted PSD features from the EEG signals, and provided three topographic maps as an input to a CNN model. In addition, the authors employed ECG, GSR, and respiration signals as input to an LSTM network. However, the dataset used by Han et al. was not a publicly accessible dataset unlike our study and studies (Alreshidi et al., 2022; Harrivel et al., 2016; Harrivel et al., 2017; Terwilliger, 2020), which were all publicly available. While their results were promising, the small sample size and lack of a public dataset may limit the generalizability of the findings. In addition, our approach achieved an accuracy of 86% in detecting mental states, which is a substantial improvement over Han et al. study's performance of 77.7%. Hernández-Sabaté et al. (Hernández-Sabaté et al., 2022) developed a CNN model to classify different mental workloads of pilots using EEG signals. Although they made their dataset publicly available, they divided a signal state to multiple states.

In comparison to our previous study (Alreshidi et al., 2022), where we evaluated the impact of different preprocessing techniques on the performance of ML algorithms for classifying pilots' mental states, the current study represents a significant improvement in mental state detection.

In this study, we developed a novel multifaceted approach that includes advanced automated preprocessing techniques, Riemannian geometry-based feature extraction, and a hybrid ensemble learning technique that combines the results of several machine learning classifiers. The use of Riemannian geometry analysis for feature extraction and the hybrid ensemble learning technique outperforms traditional approaches and shows the importance of advanced techniques in improving the accuracy of mental state detection. This study can have significant implications for improving pilots' performance and safety in the aviation industry.

Our approach has the potential to benefit several sectors within the aviation industry. One important application is in pilot training and performance evaluation. By accurately characterizing pilot mental states using EEG data, the proposed approach can be used to identify areas where pilots may need

additional training or support, and to evaluate the effectiveness of training programs in improving cognitive performance. Another potential application is in aviation safety, particularly in identifying potential safety hazards related to pilot mental states. By providing a detailed and accurate characterization of pilot mental states during flight, the proposed approach can help identify situations where pilots may be at higher risk of making errors or experiencing cognitive overload, allowing for proactive interventions to be taken to prevent accidents and improve safety. Additionally, our approach has the potential to improve human-machine interaction in the aviation industry. By using EEG data to monitor pilot mental states, future BCI systems can be developed that are better able to adapt to the cognitive state of the pilot, improving the efficiency and safety of the aviation system as a whole.

Overall, the potential applications of our approach are diverse and have the potential to make a significant contribution to the aviation industry by improving safety, training, and human-machine interaction.

4.5.3 Enhancing EEG Data Classification Through the Proposed Approach: A Comparative Analysis

In this subsection, we delve into a detailed examination of the efficacy of the advanced automated preprocessing approach introduced in this chapter. This approach, which incorporates filtering, automatic detection of eye-related artefacts, and the utilization of the Autoreject algorithm, is analyzed for its impact on the performance of SVM and ANN models. To comprehensively evaluate the advancements achieved through this method, we draw a comparative analysis with the results obtained in Chapter 3, where conventional preprocessing techniques were employed. This comparison allows us to contextualize the improvements and understand the extent to which the proposed preprocessing strategy enhances the accuracy and reliability of EEG data classification in mental state detection.

The implementation of the advanced automated preprocessing approach introduces a significant shift in the performance of both the SVM and ANN

models as illustrated in Table 4-2. This is particularly notable when comparing the results of this approach with those obtained from previous preprocessing methods (filtering and filtering combined with ICA) on the merged flight and non-flight data.

Table 4-2 Comparative Performance Metrics of SVM and ANN Models: Conventional Preprocessing Techniques (Chapter 3) vs. Proposed Approach (Chapter 4)

Preprocessing method	Model type	Evaluation Metric			
		Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Unprocessed	SVM	86.86	45.09	35.87	37.59
	ANN	83.14	44.40	41.68	42.90
Filtered	SVM	87.23	45.42	35.19	37.80
	ANN	82.95	45.28	43.40	44.32
Filtered + ICA	SVM	90.08	46.82	40.14	42.20
	ANN	85.77	51.82	48.35	50.20
Proposed Approach	SVM	87.11	87.47	86.96	87.08
	ANN	87.09	87.32	86.81	87.01

Improvement in Uniformity Across Metrics: A striking aspect of the results from the proposed approach is the uniformity across all metrics (accuracy, precision, recall, and F1-score) for both SVM and ANN models, each achieving 87% in all four categories. This uniformity suggests a highly balanced performance, indicating that the models are not only accurate overall but also equally proficient in their precision, recall, and F1-score. This is a significant improvement over the previous methods, where we observed discrepancies between different performance metrics.

Comparison with Previous Approaches: When compared to the filtering and filtering + ICA approaches, the proposed approach demonstrates a more

consistent and balanced enhancement in model performance. For instance, in the merged dataset with filtering + ICA, the SVM and ANN models showed improvements in accuracy and other metrics, but not to the extent observed with the proposed approach. The advanced automated preprocessing seems to address the limitations observed in the previous methods, particularly in terms of balancing the trade-offs between different performance measures.

Implications for Model Reliability: The results from the proposed approach suggest a significant improvement in the reliability and robustness of the models. The high precision and recall values indicate that the models are not only correctly identifying a high proportion of true positive instances but are also avoiding false positives effectively as shown in Figure 4-9. This enhanced reliability is crucial in applications like mental state classification using EEG data, where accuracy and consistency are paramount.

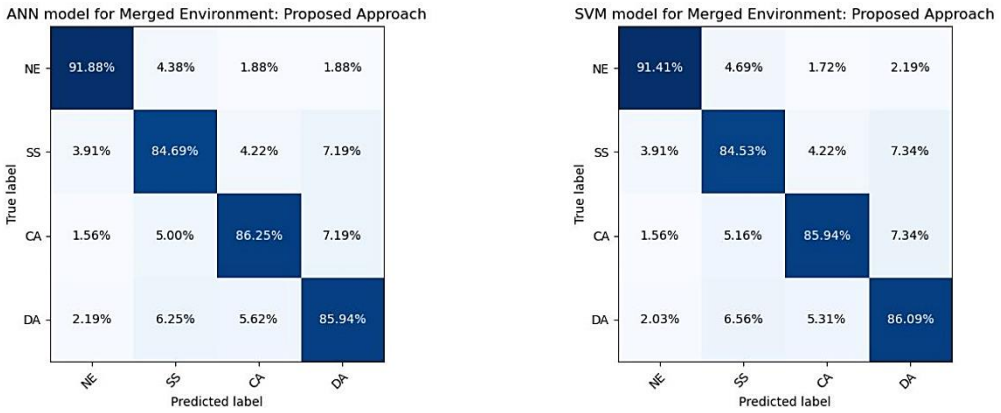

Figure 4-9 Confusion Matrices for SVM and ANN Models Using Data Preprocessed with the Proposed Approach

Potential Impact on EEG Data Analysis: The introduction of the Autoreject algorithm, in combination with filtering and automatic artefact detection, highlights the potential of advanced preprocessing techniques in refining EEG data for ML applications. This approach seems to effectively mitigate the challenges associated with noise and artefacts in the data, which are common issues in EEG analysis.

In conclusion, the proposed advanced automated preprocessing approach significantly elevates the performance of SVM and ANN models in classifying mental states from EEG data. The uniformity and high values across all key performance metrics reflect the effectiveness of this comprehensive preprocessing strategy, marking a substantial improvement over previous methods. This suggests that such advanced preprocessing techniques could be pivotal in enhancing the accuracy and reliability of EEG-based ML models in various applications.

4.6 Conclusions

To this end, we conducted an exploratory investigation using uncontaminated EEG data and ensemble learning algorithms to characterize the pilot's mental states (i.e., CA, DA, SS, and NE). We also demonstrated how the pilot's varied mental states impacted physiological indicators. With the goal of identifying the neural signal related to cognitive processes reflective of brain activity while disregarding the other artefacts and extracting significant information, we proposed a feasible approach for automatically preprocessing EEG data. In order to proceed to the classification phase, the processed data underwent feature extraction, during which spatial covariance matrices were calculated and subsequently mapped onto the Riemannian tangent space. Four ensemble learning models, namely RF, ERT, GTB, and AdaBoost, and a hybrid ensemble model were trained using tangent space vectors.

Based on the findings, it was clear that the proposed method successfully identified artefacts in the EEG epochs and either fixed or discarded them automatically. In addition, the results indicated the viability of implementing EEG-based BCI systems such as tangent space mapping to characterize the pilot's mental states. According to the findings of the pilot's mental states detection investigation, we observe that the RF, GTB, and the hybrid ensemble models are the best at predicting NE, CA, SS, and DA states, with an accuracy rate of 86%.

The innovative nature of your study lies in its combination of advanced automated preprocessing techniques, Riemannian geometry-based feature extraction, and ensemble learning models, which together provide a detailed and accurate characterization of pilot mental states, ultimately leading to a safer and more efficient aviation system.

The models' performance will be further refined, and the training dataset will be enlarged, in subsequent work. We also aim to apply the aforementioned approach to a broad range of ML and DL models. In further studies, we can also investigate the possibility of extracting other meaningful features.

4.7 Appendices

4.7.1 Appendix A: Advanced Brain Monitoring X24 EEG Headset

The X24 EEG headset was employed to gather the EEG dataset. This headset offers a wireless option for acquiring and recording EEG signals without the need for scalp abrasion. It is equipped with 20 electrodes arranged in the standard 10-20 format and one additional electrode, POz, as shown in Figure 1. These electrodes are located at specific locations on the head, such as Fz, Cz, Pz, F3, F4, C3, C4, P3, P4, O1, O2, T5, T3, F7, Fp1, Fp2, F8, T4, T6, and Linked Mastoids. The wireless technology allows for freedom of movement for the user during data collection and display in real-time. The headset collects EEG signals from the sensors on the participant and processes the signals through analog-to-digital conversion, encoding, formatting, and transmission. It operates at a sample rate of 256 Hz and uses the system's bi-directional capabilities to check scalp-electrode impedance and monitor battery capacity in the X24 Headset. **Error! Reference source not found.** illustrates the names and locations of the electrodes on the EEG sensor.

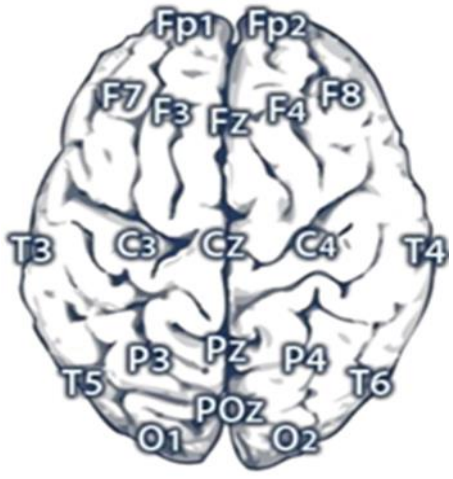

Figure 4-10 EEG electrodes' names and locations

4.7.2 Appendix B: Flight Simulator

The dataset was obtained from 18 commercial aviation pilots who participated in a research flight deck simulation at NASA Langley Research Centre. The flight deck, which is known as the cockpit motion facility, is an all-glass reconfigurable cockpit that is equipped with a programmable sidestick and pedal control inceptors. The simulator, which can operate in both motion-based and fixed-base modes, is designed to provide a high-fidelity, full-systems flight experience for pilots. It is used to evaluate and improve research concepts related to flight crew operations, covering everything from engine start-up to engine shutdown.

4.7.3 Appendix C: Data and Reproducibility Code

In the interest of promoting transparency and reproducibility, the data utilised in this chapter, along with the associated code for analyses, have been made publicly accessible. The dataset and the code for replicating the analyses can be found under the Digital Object Identifier (DOI): https://doi.org/10.17862/cranfield.rd.22232062

REFERENCES

Alreshidi, I. M., Moulitsas, I., & Jenkins, K. W. (2022). Miscellaneous EEG Preprocessing and Machine Learning for Pilots' Mental States Classification: Implications. 2022 The 6th International Conference on Advances in Artificial Intelligence,

Appriou, A., Pillette, L., Trocellier, D., Dutartre, D., Cichocki, A., & Lotte, F. (2021). BioPyC, an Open-Source Python Toolbox for Offline Electroencephalographic and Physiological Signals Classification. *Sensors (Basel)*, *21*(17). https://doi.org/10.3390/s21175740

Association, I. A. T. (2019). *Loss of Control In-Flight Accident Analysis Report*.

Avots, E., Jermakovs, K., Bachmann, M., Paeske, L., Ozcinar, C., & Anbarjafari, G. (2022). Ensemble Approach for Detection of Depression Using EEG Features. *Entropy (Basel)*, *24*(2). https://doi.org/10.3390/e24020211

Barachant, A., Bonnet, S., Congedo, M., & Jutten, C. (2010). *Riemannian geometry applied to BCI classification* LVA/ICA 2010 - 9th International Conference on Latent Variable Analysis and Signal Separation,

Barachant, A., Bonnet, S., Congedo, M., & Jutten, C. (2012). Multiclass brain-computer interface classification by Riemannian geometry. *IEEE Trans Biomed Eng*, *59*(4), 920-928. https://doi.org/10.1109/TBME.2011.2172210

Barachant, A., Bonnet, S., Congedo, M., & Jutten, C. (2013). Classification of covariance matrices using a Riemannian-based kernel for BCI applications. *Neurocomputing*, *112*, 172-178. https://doi.org/10.1016/j.neucom.2012.12.039

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K. M., & Robbins, K. A. (2015). The PREP pipeline: standardized preprocessing for large-scale EEG analysis. *Front Neuroinform*, *9*, 16. https://doi.org/10.3389/fninf.2015.00016

Binias, B., Myszor, D., & Cyran, K. A. (2018). A Machine Learning Approach to the Detection of Pilot's Reaction to Unexpected Events Based on EEG Signals. *Comput Intell Neurosci*, *2018*, 2703513. https://doi.org/10.1155/2018/2703513

Boksem, M. A., & Tops, M. (2008). Mental fatigue: costs and benefits. *Brain Res Rev*, *59*(1), 125-139. https://doi.org/10.1016/j.brainresrev.2008.07.001

Bonassi, A., Ghilardi, T., Gabrieli, G., Truzzi, A., Doi, H., Borelli, J. L., Lepri, B., Shinohara, K., & Esposito, G. (2021). The Recognition of Cross-Cultural Emotional Faces Is Affected by Intensity and Ethnicity in a Japanese Sample. *Behav Sci (Basel)*, *11*(5). https://doi.org/10.3390/bs11050059

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Cesar Cavalcanti Roza, V., & Adrian Postolache, O. (2019). Multimodal Approach for Emotion Recognition Based on Simulated Flight Experiments. *Sensors (Basel)*, *19*(24). https://doi.org/10.3390/s19245516

Dehais, F., Dupres, A., Blum, S., Drougard, N., Scannella, S., Roy, R. N., & Lotte, F. (2019). Monitoring Pilot's Mental Workload Using ERPs and Spectral Power with a Six-Dry-Electrode EEG System in Real Flight Conditions. *Sensors (Basel)*, *19*(6). https://doi.org/10.3390/s19061324

Flo, A., Gennari, G., Benjamin, L., & Dehaene-Lambertz, G. (2022). Automated Pipeline for Infants Continuous EEG (APICE): A flexible pipeline for developmental cognitive studies. *Dev Cogn Neurosci*, *54*, 101077. https://doi.org/10.1016/j.dcn.2022.101077

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, *29*(5), 1189-1232. http://www.jstor.org/stable/2699986

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3-42. https://doi.org/10.1007/s10994-006-6226-1

Guler, I., & Ubeyli, E. D. (2005). Adaptive neuro-fuzzy inference system for classification of EEG signals using wavelet coefficients. *J Neurosci Methods*, *148*(2), 113-121. https://doi.org/10.1016/j.jneumeth.2005.04.013

Hamann, A., & Carstengerdes, N. (2022). Investigating mental workload-induced changes in cortical oxygenation and frontal theta activity during simulated flights. *Sci Rep*, *12*(1), 6449. https://doi.org/10.1038/s41598-022-10044-y

Han, S. Y., Kwak, N. S., Oh, T., & Lee, S. W. (2020). Classification of pilots' mental states using a multimodal deep learning network. *Biocybernetics and Biomedical Engineering*, *40*(1), 324-336. https://doi.org/10.1016/j.bbe.2019.12.002

Hankins, T. C., & Wilson, G. F. (1998). A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. *Aviat Space Environ Med*, *69*(4), 360-367. https://www.ncbi.nlm.nih.gov/pubmed/9561283

Harrivel, A. R., Liles, C., Stephens, C. L., Ellis, K. K., Prinzel, L. J., & Pope, A. T. (2016). Psychophysiological Sensing and State Classification for Attention Management in Commercial Aviation. AIAA Infotech @ Aerospace,

Harrivel, A. R., Stephens, C. L., Milletich, R. J., Heinich, C. M., Last, M. C., Napoli, N. J., Abraham, N., Prinzel, L. J., Motter, M. A., & Pope, A. T. (2017). Prediction of Cognitive States during Flight Simulation using Multimodal

Psychophysiological Sensing. AIAA Information Systems-AIAA Infotech @ Aerospace,

Hernández-Sabaté, A., Yauri, J., Folch, P., Piera, M. À., & Gil, D. (2022). Recognition of the Mental Workloads of Pilots in the Cockpit Using EEG Signals. *Applied Sciences*, *12*(5). https://doi.org/10.3390/app12052298

Hogervorst, M. A., Brouwer, A. M., & van Erp, J. B. (2014). Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. *Front Neurosci*, *8*, 322. https://doi.org/10.3389/fnins.2014.00322

International_Air_Transport_Association. (2022).

Jaquess, K. J., Gentili, R. J., Lo, L. C., Oh, H., Zhang, J., Rietschel, J. C., Miller, M. W., Tan, Y. Y., & Hatfield, B. D. (2017). Empirical evidence for the relationship between cognitive workload and attentional reserve. *Int J Psychophysiol*, *121*, 46-55. https://doi.org/10.1016/j.ijpsycho.2017.09.007

Jas, M., Engemann, D. A., Bekhti, Y., Raimondo, F., & Gramfort, A. (2017). Autoreject: Automated artifact rejection for MEG and EEG data. *Neuroimage*, *159*, 417-429. https://doi.org/10.1016/j.neuroimage.2017.06.030

Jiang, G., Chen, H., Wang, C., & Xue, P. (2022). Mental Workload Artificial Intelligence Assessment of Pilots' EEG Based on Multi-Dimensional Data Fusion and LSTM with Attention Mechanism Model. *International Journal of Pattern Recognition and Artificial Intelligence*, *36*(11), 2259035. https://doi.org/10.1142/S0218001422590352

Jiang, X., Bian, G. B., & Tian, Z. (2019). Removal of Artifacts from EEG Signals: A Review. *Sensors (Basel)*, *19*(5). https://doi.org/10.3390/s19050987

Jiao, Z. C., Gao, X. B., Wang, Y., Li, J., & Xu, H. J. (2018). Deep Convolutional Neural Networks for mental load classification based on EEG data. *Pattern Recognition*, *76*, 582-595. https://doi.org/10.1016/j.patcog.2017.12.002

Johnson, M. K., Blanco, J. A., Gentili, R. J., Jaquess, K. J., Oh, H., & Hatfield, B. D. (2015). Probe-Independent EEG Assessment of Mental Workload in Pilots. 7th Annual International IEEE EMBS Conference on Neural Engineering,

Kelly, D., & Efthymiou, M. (2019). An analysis of human factors in fifty controlled flight into terrain aviation accidents from 2007 to 2017. *J Safety Res*, *69*, 155-165. https://doi.org/10.1016/j.jsr.2019.03.009

Kordylewski, H., Graupe, D., & Liu, K. (2001). A novel large-memory neural network as an aid in medical diagnosis applications. *Ieee Transactions on Information Technology in Biomedicine*, *5*(3), 202-209. https://doi.org/Doi 10.1109/4233.945291

Liu, Y., Ayaz, H., & Shewokis, P. A. (2017). Multisubject "Learning" for Mental Workload Classification Using Concurrent EEG, fNIRS, and Physiological Measures. *Front Hum Neurosci*, *11*, 389. https://doi.org/10.3389/fnhum.2017.00389

Majidov, I., & Whangbo, T. (2019). Efficient Classification of Motor Imagery Electroencephalography Signals Using Deep Learning Methods. *Sensors (Basel)*, *19*(7). https://doi.org/10.3390/s19071736

Nittala, S. K. R., Elkin, C. P., Kiker, J. M., Meyer, R., Curro, J., Reiter, A. K., Xu, K. S., & Devabhaktuni, V. K. (2018). *Pilot Skill Level and Workload Prediction for Sliding-Scale Autonomy* 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA),

Oh, H., Hatfield, B. D., Jaquess, K. J., Lo, L.-C., Tan, Y. Y., Prevost, M. C., Mohler, J. M., Postlethwaite, H., Rietschel, J. C., Miller, M. W., Blanco, J. A., Chen, S., & Gentili, R. J. (2015). A Composite Cognitive Workload Assessment System in Pilots Under Various Task Demands Using Ensemble Learning. In *Foundations of Augmented Cognition* (pp. 91-100). https://doi.org/10.1007/978-3-319-20816-9_10

Pousson, J. E., Voicikas, A., Bernhofs, V., Pipinis, E., Burmistrova, L., Lin, Y. P., & Griskova-Bulanova, I. (2021). Spectral Characteristics of EEG during Active Emotional Musical Performance. *Sensors (Basel)*, *21*(22). https://doi.org/10.3390/s21227466

Santos, L., & Ferreira, L. (2023). Atlantic—Automated data preprocessing framework for supervised machine learning. *Software Impacts*, *17*. https://doi.org/10.1016/j.simpa.2023.100532

Schapire, R. E. (2003). The Boosting Approach to Machine Learning: An Overview. In D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, & B. Yu (Eds.), *Nonlinear Estimation and Classification* (pp. 149-171). Springer New York. https://doi.org/10.1007/978-0-387-21579-2_9

Singh, A., Lal, S., & Guesgen, H. W. (2019). Reduce Calibration Time in Motor Imagery Using Spatially Regularized Symmetric Positives-Definite Matrices Based Classification. *Sensors (Basel)*, *19*(2). https://doi.org/10.3390/s19020379

SKY_Brary. (2019). *SE211_ Airplane State Awareness - Training for Attention Management (R-D)*. https://skybrary.aero/articles/se211-airplane-state-awareness-training-attention-management-r-d

Stancin, I., Cifrek, M., & Jovic, A. (2021). A Review of EEG Signal Features and their Application in Driver Drowsiness Detection Systems. *Sensors (Basel)*, *21*(11). https://doi.org/10.3390/s21113786

Terwilliger, P. S., Jack; Walker, Shannon; Harrivel, Angela. (2020). *A ResNet Autoencoder Approach for Time Series Classification of Cognitive State* MODSIM,

Übeyli, E. D. (2008). Wavelet/mixture of experts network structure for EEG signals classification. *Expert Systems with Applications*, *34*(3), 1954-1962. https://doi.org/10.1016/j.eswa.2007.02.006

Wang, X. W., Nie, D., & Lu, B. L. (2014). Emotional state classification from EEG data using machine learning approach. *Neurocomputing*, *129*, 94-106. https://doi.org/10.1016/j.neucom.2013.06.046

Wu, E. Q., Peng, X. Y., Zhang, C. Z. Z., Lin, J. X., & Sheng, R. S. F. (2019). Pilots' Fatigue Status Recognition Using Deep Contractive Autoencoder Network. *Ieee Transactions on Instrumentation and Measurement*, *68*(10), 3907-3919. https://doi.org/10.1109/Tim.2018.2885608

Yen, J. R., Hsu, C. C., Yang, H., & Ho, H. (2009). An investigation of fatigue issues on different flight operations. *Journal of Air Transport Management*, *15*(5), 236-240. https://doi.org/10.1016/j.jairtraman.2009.01.001

Zhang, P., Wang, X., Chen, J., & You, W. (2017). Feature Weight Driven Interactive Mutual Information Modeling for Heterogeneous Bio-Signal Fusion to Estimate Mental Workload. *Sensors (Basel)*, *17*(10). https://doi.org/10.3390/s17102315

Zhang, P., Wang, X., Zhang, W., & Chen, J. (2019). Learning Spatial-Spectral-Temporal EEG Features With Recurrent 3D Convolutional Neural Networks for Cross-Task Mental Workload Assessment. *IEEE Trans Neural Syst Rehabil Eng*, *27*(1), 31-42. https://doi.org/10.1109/TNSRE.2018.2884641

Ziegler, M. D., Kraft, A., Krein, M., Lo, L.-C., Hatfield, B., Casebeer, W., & Russell, B. (2016). Sensing and Assessing Cognitive Workload Across Multiple Tasks.*Lecture Notes in Computer Science* Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience,

5 A Comprehensive Analysis of Machine Learning and Deep Learning Models for Identifying Pilots' Mental States from Imbalanced Physiological Data

5.1 Abstract

This study focuses on identifying pilots' mental states linked to attention-related human performance-limiting states (AHPLS) using a publicly released, imbalanced physiological dataset. The research integrates electroencephalography (EEG) with non-brain signals, such as electrocardiogram (ECG), galvanic skin response (GSR), and respiration, to create a deep learning architecture that combines one-dimensional Convolutional Neural Network (1D-CNN) and Long Short-Term Memory (LSTM) models. Addressing the data imbalance challenge, the study employs resampling techniques, specifically downsampling with cosine similarity and oversampling using Synthetic Minority Over-sampling Technique (SMOTE), to produce balanced datasets for enhanced model performance. An extensive evaluation of various machine learning and deep learning models, including XGBoost, AdaBoost, Random Forest (RF), Feed-Forward Neural Network (FFNN), standalone 1D-CNN, and standalone LSTM, is conducted to determine their efficacy in detecting pilots' mental states. The results contribute to the development of efficient mental state detection systems, highlighting the XGBoost algorithm and the proposed 1D-CNN+LSTM model as the most promising solutions for improving safety and performance in aviation and other industries where monitoring mental states is essential.

5.2 Introduction

The evolution of the aviation industry is intricately linked to the enhancement of safety standards, technological advancements in aircraft design, and operational methodologies that collectively contribute to a notable decrease in aviation accidents globally (Oehling & Barry, 2019; Pan et al., 2021). This progress is underscored by a comprehensive understanding of the complex

interplay between human cognitive processes and the high-tech environment of cockpit operations. Pilots, central to the operational hierarchy of flight safety, encounter various scenarios, ranging from navigating through extreme weather conditions to responding to cockpit alerts, each presenting unique challenges to maintaining attentional focus and situational awareness. The significance of cognitive tendencies, particularly in high-stakes situations like take-off and landing, cannot be overstated. Such moments demand acute mental agility and the ability to swiftly adapt to changing circumstances, factors that are crucial in mitigating risks associated with flight operations.

Research has illuminated the critical role of cognitive functions in aviation safety, noting that human factors contribute to a significant proportion of aviation accidents (Boksem & Tops, 2008; Hankins & Wilson, 1998; Yen et al., 2009). The International Air Transport Association (IATA) has documented instances where lapses in pilots' mental states have led to catastrophic outcomes, emphasizing the urgent need for a deeper exploration into cognitive tendencies that might predispose pilots to operational errors (Jiang et al., 2021; Walmsley & Gilbey, 2016). The dynamics of attentional focus under stress and its correlation with operational proficiency have been a focal point of recent studies, highlighting the need for strategies to enhance mental sharpness and cognitive resilience in pilots.

In addition to identifying cognitive biases and situational awareness deficiencies, there is a growing body of research dedicated to developing methodologies for detecting and managing pilots' attention-related mental states. Studies employing physiological signals and machine learning (ML) techniques have made strides in this arena, suggesting that a multimodal approach to monitoring cognitive states could pave the way for significant improvements in aviation safety (Giraudet et al., 2015; Jiang et al., 2020). For instance, electroencephalography (EEG) has been spotlighted for its potential to detect rapid changes in brain activity that are indicative of cognitive load or stress, albeit with challenges such as susceptibility to environmental interference (Khanna et al., 2015). To counteract these limitations, researchers

have advocated for the concurrent collection of additional physiological data such as electrocardiogram (ECG), galvanic skin response (GSR), and respiration (Resp.) to create a more robust framework for understanding pilots' physiological responses under varied operational conditions (Han et al., 2020).

However, the development of effective mental state detection systems is not without its challenges. One of the most pressing issues is the imbalance in data available for training ML models (Chawla et al., 2002), a problem that arises from the variability in the frequency of different mental states encountered in real-world flight operations. This imbalance can lead to detection systems that are biased towards more commonly occurring states, reducing their efficacy in identifying less frequent but potentially critical mental states (Han et al., 2019). Addressing these challenges requires not only technical solutions but also a holistic approach that includes training pilots to recognize and mitigate the effects of cognitive biases and stress on their decision-making processes.

The aim of this study is to investigate the potential for identifying mental states associated with attention-related human performance limiting states (AHPLS), namely channelized attention, diverted attention, startle/surprise, and normal states, using an imbalanced, publicly released physiological dataset. The research offers two primary contributions:

First, it develops a comprehensive deep learning (DL) architecture, consisting of one-dimensional Convolutional Neural Network (1D-CNN) and Long Short-Term Memory (LSTM) models, designed to effectively combine EEG and non-brain signals. Although this architecture was used to classify emotions using physiological signals in other applications, it has not been used to AHPLS in the aviation-context. For instance, Tripathi et al. (Tripathi et al., 2017) utilized a 1D-CNN+LSTM model for accurate emotion classification on the Dataset for Emotion Analysis using Physiological and Audiovisual Signals (DEAP) (Koelstra et al., 2012), which contains EEG and peripheral physiological signals. Similarly, Zheng and Lu (Wei-Long & Bao-Liang, 2015) investigated critical frequency bands and channels for EEG-based emotion recognition using a 1D-CNN+LSTM model.

Second, it addresses the data imbalance issue by employing data resampling techniques, such as downsampling and oversampling, to create more balanced datasets for improved model performance. In addition to the 1D-CNN and LSTM fusion model, the study also incorporates and critically analyzes the performance of other ML and DL models, including eXtreme Gradient Boosting (XGBoost), Adaptive Boosting (AdaBoost), Random Forest (RF), Feed-Forward Neural Network (FFNN), standalone 1D-CNN, and standalone LSTM. This comprehensive analysis aims to provide a more robust understanding of the strengths and weaknesses of each model when dealing with imbalanced physiological data and detecting pilots' mental states.

The performance results of ensemble learning models and DL models, along with the impact of data resampling techniques like Synthetic Minority Over-sampling Technique (SMOTE) and the combination of Cosine Similarity (CS) and SMOTE, are reported and compared. This comparison sheds light on the effectiveness of different models and resampling techniques in handling data imbalance and improving mental state detection in the context of AHPLS.

The remainder of the research paper is structured as follows: Section II offers an overview of the relevant literature. Section III delineates the utilized dataset, the preprocessing methods, feature extraction methods, data imbalance approaches, and the classification methods utilized in this study. Section IV presents and discusses the experimental findings. Finally, Section V. concludes the investigation and suggests future research directions.

5.3 Related Work

The literature on previous studies that have attempted to detect AHPLS and address the data imbalanced issue are reviewed in this section.

5.3.1 Mental States Detection in the Context of AHPLS

Prior research has delved into the detection and evaluation of AHPLS. For instance, Harrivel et al. (Harrivel et al., 2016) employed RF, XGBoost, and Deep Neural Network (DNN) classifiers in a sophisticated flight simulator

environment to predict CA, DA, and low workload states using various sensing modalities. In subsequent research, Harrivel et al. (Harrivel et al., 2017) utilized RF, gradient boosting, and two SVM classifiers to discern CA and SS states. Terwilliger et al. (Terwilliger, 2020) aggregated CA, DA, and SS mental states into an "event" category and introduced a convolutional autoencoder method to differentiate the event class from the normal state (NE). In earlier investigations, the impact of two preprocessing techniques on SVM and Artificial Neural Network (ANN) models using EEG data from a pilot exposed to CA, DA, SS, and NE states was examined (Alreshidi et al., 2022). However, there were certain limitations to these studies: 1) the performance was not optimal, and 2) no study conducted a multiclass classification categorizing CA, DA, SS, and NE. Notably, the curse of dimensionality restricted the accuracy of predicting DA and SS states, despite the potential of merging data from two distinct scenarios.

5.3.2 Addressing Data Imbalance Issue

A critical challenge in mental state detection using biosignals is the data imbalance issue, where certain mental states may be underrepresented in the dataset (Haibo & Garcia, 2009; Haixiang et al., 2017). This issue can lead to biased model predictions and poor generalization to real-world scenarios (Weiss & Provost, 2001). This subsection explores various innovative approaches developed to address this challenge, highlighting key findings and methodologies from recent studies.

The DL techniques have shown promise in handling class imbalanced data, as explored by Johnson and Khoshgoftaar (Johnson & Khoshgoftaar, 2019), who reviewed existing DL strategies for class imbalance. They found that traditional methods like data sampling and cost-sensitive learning are effective when applied to DL, with advanced methods leveraging neural network features showing promising results. This indicates a potential direction for future research in DL applied to imbalanced data, especially in domains like fraud and cancer detection where imbalance is prevalent. In another studies, Ahlawat and Singh (Ahlawat & Singh, 2020) demonstrated the effectiveness of fuzzy logic

combined with MapReduce architecture in classifying imbalanced big data for human activity recognition. Their findings suggest that fuzzy algorithms can significantly improve prediction accuracy as the imbalance ratio increases, showcasing the potential of fuzzy logic in handling complex, imbalanced datasets. Sun et al. (Sun et al., 2019) proposed an imbalanced learning model for epileptic seizure detection from EEG signals. By employing discrete wavelet transform and uniform 1D-LBP feature extraction along with an ensemble of SVM classifiers, they achieved improved seizure detection performance. This approach highlights the importance of specialized feature extraction and ensemble learning in managing imbalanced datasets. Wu et al. (Wu et al., 2016) addressed the imbalance dataset problem in human activity recognition with a mixed-kernel based weighted extreme learning machine (MK-WELM). Their method effectively reduced the influence of imbalance datasets, demonstrating the applicability of cost-sensitive methods and mixed-kernel approaches in extreme learning machines. Puspaningrum et al. (Puspaningrum et al., 2020) explored the use of DeepCNN for Alzheimer's disease stage classification from imbalanced MRI data. They highlighted the role of oversampling in improving classification accuracy, emphasizing the need for data augmentation in dealing with imbalanced distributions in health datasets. Sharma and Verbeke (Sharma & Verbeke, 2020) utilized an XGBOOST ML model on a large biomarkers Dutch dataset to improve the diagnosis of depression, addressing the class imbalance problem through various resampling strategies. Their study achieved high accuracy, precision, recall, and F1 score, showcasing the effectiveness of resampling strategies in dealing with imbalanced datasets. Kaur et al. (Kaur et al., 2019) provided a comparative analysis of approaches for dealing with imbalanced data challenges, emphasizing the hybrid approach that combines algorithm level and data level methods for effective imbalance data analysis. This systematic review highlights the diverse strategies available for managing data imbalance across different domains. Alani et al. (Alani et al., 2020) explore DL approaches to handling imbalanced multi-modal sensor data for human activity recognition in smart homes, demonstrating the effectiveness of data fusion and resampling methods

in improving classification accuracy. Finally, Pamplona-Beron et al. (Pamplona-Beron et al., 2021) evaluated penalized support vector machines (SVM) for classifying human activities, demonstrating significant advancements in detecting micro-movements compared to non-penalized paradigms. Their research underscores the potential of penalized models in enhancing the performance of classifiers dealing with imbalanced data.

Collectively, these studies illustrate a broad spectrum of methodologies and strategies to tackle the data imbalance issue in ML applications involving physiological signals for human behaviour detection. However, none of them have not adequately addressed this problem in the context of pilot's mental state detection using DL models, traditional ML techniques, and multimodal biosignals (Kim et al., 2004; Minguillon et al., 2017).

The novelty of the proposed research lies in its application of a 1D-CNN+LSTM architecture to predict pilots' mental states (CA, DA, SS, and NE) using EEG signals and non-brain signals, such as ECG, GSR, and R. Furthermore, this study addresses the data imbalance issue by employing resampling strategies, including downsampling using the CS method and oversampling using the SMOTE method. To the best of our knowledge, no previous study has combined these specific mental states, DL architecture, multimodal biosignals, and data balancing techniques with traditional ML methods to predict pilots' mental states, making this research a unique contribution to the field.

5.4 Materials and Methods

5.4.1 AHPLS Dataset

The AHPLS dataset, collected by Harrivel et al. (Harrivel et al., 2016), is publicly released on the NASA open portal website. It comprises psychophysiological data gathered from 18 pilots during various scenario events designed to induce CA, DA, SS, and NE states. These data were recorded using the Advanced Brain Monitoring X24 EEG and Mind Media B.V. Nexus Mark II systems. For each pilot, four sets of data, including EEG, ECG, GSR, and Respiration, were provided. Three of the four sets were collected in a non-flying environment,

while the fourth set was obtained in a high-fidelity flight simulator, featuring approximately one hour of labeled benchmark data. This set consists of 25 columns, which include a time stamp, 20 EEG channels, an ECG channel, an Resp. channel, a GSR channel, and an event label.

In this research, the fourth set was utilized as it was collected in a flight simulator and contains labeled benchmark data that induced the states of interest (NE, SS, CA, and DA). The NE, SS, CA, and DA states are annotated as Class 0, Class 1, Class 2, and Class 3, respectively. The dataset exhibits significant class imbalance; for each pilot, Class 0 constitutes approximately 83% of the data, followed by Class 2 at about 14%, Class 3 at around 2%, and Class 1 comprising only 1% of the data.

5.4.2 Signal Preprocessing

The EEG, ECG, GSR, and Resp. signals were preprocessed using open-source libraries, specifically MNE-Python (Gramfort et al., 2013) and BioSSPy (Carreiras et al., 2015). MNE-Python was employed to implement advanced preprocessing techniques for cleaning artifacts from the EEG data, ensuring the highest quality signal for subsequent analysis. Initially, the dataset was transformed into a compatible format to facilitate the use of MNE-Python functions (Alreshidi et al., 2022). For the EEG signal, an automated preprocessing pipeline was employed to identify and eliminate artifacts (Alreshidi et al., 2023), ensuring the data's integrity and reliability. In parallel, the ECG, GSR, and Resp. signals were filtered using BioSSPy, a specialized library for biosignal processing. With the aid of BioSSPy, one distinctive feature was extracted from each of these channels, providing a comprehensive representation of the physiological data. The combination of MNE-Python and BioSSPy allowed for effective preprocessing and feature extraction, setting the foundation for accurate and reliable analysis of the pilots' psychophysiological states.

5.4.3 Features Extraction

For the EEG signals, the Power Spectral Density (PSD) features were extracted using Welch's method (WELCH, 1967), a widely recognized technique for spectral estimation. Welch's method employs the Fast Fourier Transform (FFT) algorithm to estimate power spectra, providing an accurate representation of the signals' frequency domain characteristics. The parameters utilized for extracting the PSD values using the MNE-Python library are summarized in Table 5-1.

Table 5-1 Parameters utilized for PSD values extraction.

Parameters	Description	Value
sfreq	The sampling frequency	256
fmin	The lower frequency of interest.	1
fmax	The upper frequency of interest	50
n_fft	The length of FFT used	1280
n_overlap	The number of points of overlap between segments	255
n_per_seg	Length of each Welch segment	1280
window	Windowing function to use	boxcar

The EEG signals were recorded at a sampling frequency rate of 256 Hz. Consequently, the 'sfreq' parameter was set to 256, matching the sampling frequency rate. The 'fmin' and 'fmax' parameters were set to 1 and 50, respectively, generating 50 periodograms (i.e., features) for each channel within each epoch. These parameters define the range of periodograms and yield an equal number of PSD values for each epoch. The default length of FFT and the Welch segment is 256, equivalent to 1 second. Both the length of FFT and the Welch segment can adopt values that are multiples of the sampling frequency. In this study, the length of FFT and the Welch segment was set to 1280,

corresponding to 5 seconds with an overlap of one second. The key equations associated with Welch's method (Smith, 2011) are outlined below:

Let $x_m(n) \triangleq w(n)x(n+mR)$ represent the m^{th} windowed segment of the signal, where $n = 0,1,\dots,M-1$ and $m = 0,1,\dots,K-1$. R denotes the window hop size, and K indicates the total number of segments. The periodogram of the m^{th} block is calculated as:

$$P_{x_m,M}(\omega_k) = \frac{1}{M}\left|FFT_{N,k}(x_m)\right|^2 \triangleq \frac{1}{M}\left|\sum_{n=0}^{N-1} x_m\,(n)\; e^{-\frac{j2\pi nk}{N}}\right|^2 \quad \textbf{(5-1)}$$

The Welch estimate of the power spectral density is given by the average of periodograms across all segments:

$$S_x^W \triangleq \frac{1}{K}\sum_{m=0}^{K-1} P_{x_m,M}(\omega_k). \quad \textbf{(5-2)}$$

This method computes an average of periodograms derived from non-overlapping successive blocks of data when $w(n)$ is a rectangular window.

Welch's method produces 50 features for each channel, totaling 1000 for each epoch. Figure 5-1 illustrates the Welch's periodogram for a single epoch and channel.

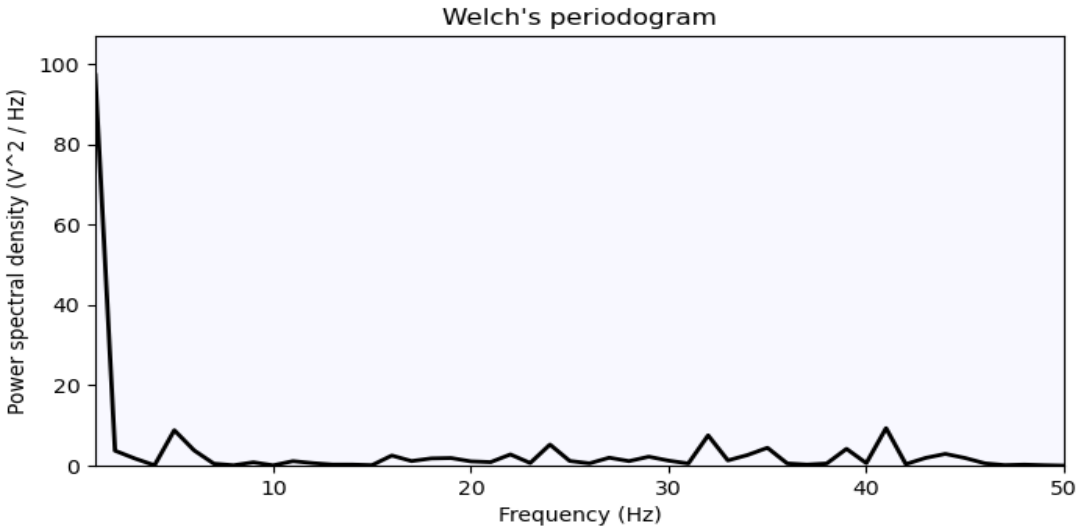

Figure 5-1 Welch's periodogram for a single epoch and channel

To reduce the dimensionality of the feature space from 1000 to 100 features, the absolute PSD values for five distinct frequency bands were computed. These frequency bands include delta (0-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (13-20 Hz), and gamma (20-50 Hz). For instance, to extract the PSD values intersecting the delta band, the 'logical AND' operation from the NumPy library was employed. As illustrated in Figure 5-2, the absolute PSD values for the delta band were determined using the Area Under Curve (AUC) method. Due to the curve's indefinite shape, the Composite Simpson's Rule (CSR) was utilized to compute the AUC. The absolute PSD values for other bands were calculated similarly. The CSR operates on the principle of dividing the larger area into smaller parabolic segments and subsequently calculating the sum of the area under each parabola.

The total number of EEG features generated per epoch was 5 × 20. In addition to these 100 features, ECG, R, and GSR signals were incorporated into the dataset after filtering and feature extraction using the BioSSPy library, contributing one feature per channel. Consequently, the total number of features generated per epoch amounted to 103.

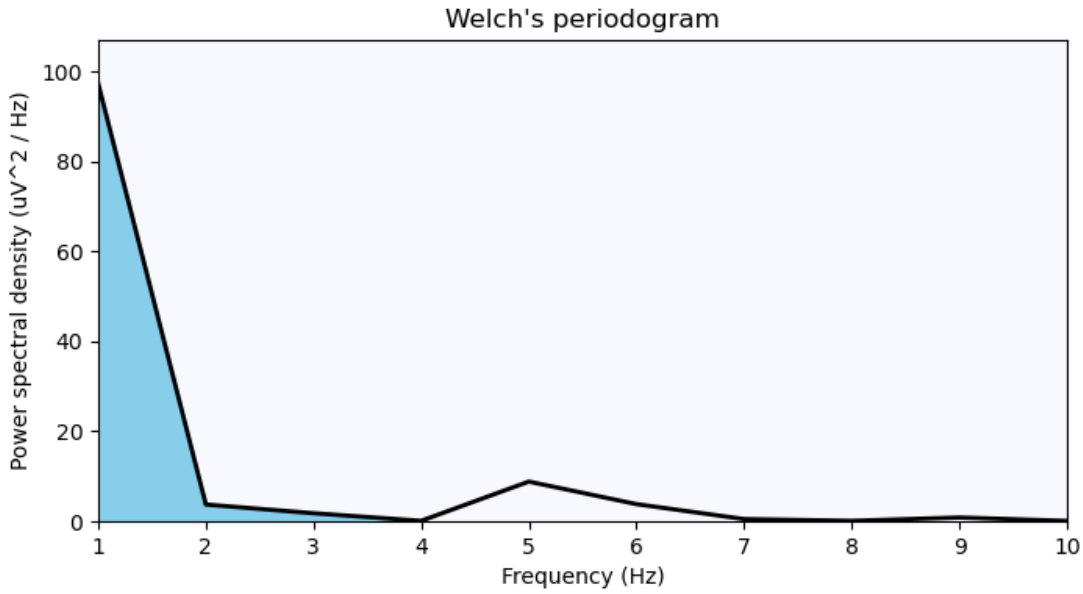

Figure 5-2 Delta Band’s Absolute PSD

In this chapter, the feature extraction for EEG signals primarily focused on PSD using Welch's method. This method was chosen for its effectiveness in providing a clear depiction of the frequency domain characteristics of EEG signals, which are crucial for the mental state detection tasks addressed in this chapter. While in the previous chapter, tangent space features derived from Riemannian geometry were utilized, the focus in Chapter 5 shifted to PSD features due to their direct applicability in distinguishing specific mental states relevant to this phase of the research.

The selection of PSD features over tangent space features is a reflection of the distinct objectives and methodological approaches of each chapter. While tangent space features offer a comprehensive view of EEG signal characteristics and are advantageous in certain contexts, PSD features were deemed more suitable for the specific analysis conducted in Chapter 5. This choice aligns with the chapter's aim to utilize spectral information directly related to the cognitive processes under investigation.

Additionally, connectivity features, which analyse the interrelations between different EEG channels, were not included in this chapter. The decision to exclude these features was based on the desire to concentrate on spectral content for a more straightforward analysis in the context of this chapter. While connectivity features can provide valuable insights into brain network dynamics, they were not aligned with the primary focus of this chapter. However, the importance of such features is recognized, and they represent a promising avenue for future exploration to build on the findings of this research.

5.4.4 Data Balancing

In this subsection, two resampling techniques, namely Cosine Similarity and Synthetic Minority Over-sampling Technique (SMOTE), are introduced and explained in detail. These techniques are employed in the study to address the data imbalance issue, which is a critical challenge in mental state detection using multimodal biosignals.

5.4.4.1 Cosine Similarity (CS)

The CS is a widely used similarity metric to measure the angular distance between two vectors in a multi-dimensional space, providing a value between -1 and 1. It is particularly effective in high-dimensional datasets, as it is less sensitive to the size of the vectors compared to Euclidean distance. The CS between two non-zero vectors A and B is calculated using the following formula:

$$Cosine\ Similarity(A,B) = \frac{(A.B)}{(\|A\|\|B\|)} \quad \textbf{(5-3)}$$

where A and B are two non-zero vectors, $A.B$ denotes the dot product of A and B, and $||A||$ and $||B||$ represent the magnitudes of the vectors A and B, respectively.

In this study, the CS method is utilized as a downsampling technique to identify and remove similar instances within the majority class (i.e., Class 0). By computing the similarity between instances in the majority class, the most representative samples can be retained, thus reducing the data imbalance and mitigating the impact of duplicate or highly similar instances on the model's performance.

5.4.4.2 Synthetic Minority Over-sampling Technique (SMOTE)

The SMOTE method is an oversampling approach that generates synthetic samples for the minority classes such as DA and SS to balance the class distribution. Unlike simple oversampling techniques that replicate minority class epochs, SMOTE generates synthetic epochs that lie along the line segments joining the minority class instances and their k-nearest neighbours in the feature space.

The process of generating synthetic epochs using SMOTE involves identifying the k-nearest neighbours in the feature space for each epoch in the minority classes. Then, select a random epoch from the minority class and one of its k-nearest neighbours. Finally, generate a synthetic epoch by interpolating between the chosen epoch and its neighbour using the following formula:

$$Synthetic\ Epoch = Epoch + \lambda * (Neighbor - Epoch) \quad \textbf{(5-4)}$$

where $Epoch$ is the randomly selected epoch from the minority class, $Neighbor$ is one of its k-nearest neighbors, and λ is a random number between 0 and 1.

In this study, the SMOTE method is applied to generate synthetic samples for the underrepresented mental states (i.e., CA, DA, and SS) in the dataset. By employing both the CS and SMOTE methods, the methodology effectively addresses the data imbalance issue in the dataset, which is essential for improving the performance and generalization of all the adopted models for predicting pilots' mental states.

5.4.5 Classification Methods

In the present chapter, following the completion of data cleaning, feature extraction, and data balancing, four DL models and three ensemble learning models were employed to perform a multiclass classification task. The DL models included LSTM, 1D-CNN, a combined 1D-CNN and LSTM architecture, and FFNN.

The decision to employ the 1D-CNN for analysing EEG data, specifically with PSD features, and the LSTM network for ECG, GSR, and Resp. signals was grounded in the inherent characteristics and dimensional structure of the respective datasets. The EEG data, characterised by its temporal and spectral dimensions (channels and frequencies), were effectively handled by the 1D-CNN after flattening the features, which simplified the data structure to a single dimension of features per sample. Conversely, the inclusion of 2D-CNN was not deemed necessary, as the primary dataset did not exhibit the spatial relationships across two dimensions that 2D-CNNs excel in capturing. Moreover, the computational efficiency and simplicity of 1D-CNNs offered a pragmatic advantage, given the already complex nature of integrating multiple biometric signals.

The ensemble learning models, encompassing AdaBoost, XGBoost, and RF, were trained on a dataset composed of combined pilot data. In comparison to

other algorithms, such as Logistic Regression and SVM, the ensemble learning algorithms demonstrated superior performance due to their ability to derive hyper-rectangles in the feature space.

For ensemble learning models, the hyperparameters tuning was performed using the *GridSearchCV* function from the scikit-learn library. This function takes a dictionary of hyperparameters and their values as input and constructs a grid of all possible combinations of hyperparameters using the k-fold cross-validation method. In this study, the learning rate, sub-sample, algorithm, bootstrap, n_estimators, and max_depth hyperparameters were fine-tuned using a 3-fold cross-validation method. To fine-tune the DL models developed in this study, a trial-and-error approach was adopted. The hyperparameters of the FFNN, 1D-CNN, LSTM, and CNN+LSTM models that were fine-tuned include learning rate, batch size, and epochs. Fine-tuning these hyperparameters can lead to overfitting or underfitting, which in turn affects the performance of the DL models. Each of the aforementioned models is briefly described in the following subsections, providing an overview of their structure and function in the context of this multiclass classification task.

5.4.5.1 Adaptive Boosting (AdaBoost)

The AdaBoost algorithm is a powerful ensemble method that combines multiple weak classifiers, each trained on different subsets of the training data, to form a robust and accurate strong classifier. The primary objective of this approach is to create a more efficient classifier by capitalizing on the strengths of the individual weak classifiers while minimizing their weaknesses. Initially, the algorithm assigns equal weights to each sample in the training set. Subsequently, a weak classifier is trained on this training set, and its error rate is computed. Based on the error rate, the algorithm calculates the weight of the weak classifier and updates the weights of the samples in the training set. This iterative process continues for a predetermined number of iterations or until a specified threshold is achieved. Upon completion of the iterative process, the weak classifiers are combined by weighting their individual outputs based on their calculated weights, thus forming a strong classifier. The final prediction is

made using this combined classifier, which is expected to exhibit improved performance compared to its constituent weak classifiers. By continuously updating the weights of the samples in the training set and retraining the weak classifiers, AdaBoost effectively focuses on the samples that are challenging to classify, thereby enhancing the overall performance of the final classifier. In the present study, several hyperparameters are optimized to achieve the best performance for the AdaBoost classifier. The learning rate, max depth, number of estimators, and loss function parameters are set to 0.6, 5, 200, and 'SAMME', respectively.

5.4.5.2 Extreme Gradient Boosting (XGBoost)

The XGBoost algorithm is a state-of-the-art ensemble learning method that iteratively trains a sequence of weak decision trees and combines their predictions to form a powerful and accurate model. It employs a gradient boosting framework, which involves fitting a model on the residual errors of the preceding iteration. In each iteration, the algorithm calculates the gradient of the loss function with respect to the predicted values, subsequently updating the weights of the decision trees to minimize the loss. XGBoost incorporates regularization techniques, such as L1 and L2 regularization, to prevent overfitting, ensuring a more robust model capable of generalizing well to unseen data. Additionally, it includes a feature selection method that evaluates the importance of each feature, contributing to a more efficient and interpretable model. By integrating these techniques, XGBoost produces highly accurate models that can effectively handle complex datasets with numerous features, making it a popular choice for various ML tasks and applications. In the current study, several hyperparameters are fine-tuned to achieve optimal performance for the XGBoost classifier. The learning rate, max depth, number of estimators, and subsample parameters are set to 0.6, 2, 200, and 0.9, respectively.

5.4.5.3 Random Forest (RF)

RF is an ensemble learning technique that constructs multiple decision trees on distinct subsets of the training data, subsequently integrating their predictions to form a robust model. Each decision tree within the forest is trained on a unique

subset of the training data, and at each node, a random subset of features is chosen for splitting. This strategy serves to mitigate overfitting and enhances the model's generalization capabilities. During the prediction phase, each decision tree in the forest independently forecasts the outcome. The final prediction is then determined by aggregating the individual predictions, typically through a majority vote mechanism. This methodology yields highly accurate models capable of managing intricate datasets characterized by a multitude of features. Additionally, it allows for the assessment of the relative importance of each feature within the dataset. In the present study, the hyperparameters for the RF model are configured as follows: the maximum depth is set to 5, limiting the extent of tree growth and complexity; the number of trees is established at 600, providing a large enough ensemble to capture diverse patterns in the data; and the bootstrap parameter is set to True, enabling the usage of bootstrapped samples for training each individual tree.

5.4.5.4 Feed-Forward Neural Network (FFNN)

FFNN is a type of multi-layer ANN wherein the information flow proceeds unidirectionally, transitioning from the input layer through one or more hidden layers, ultimately reaching the output layer. Each neuron in the network receives a weighted sum of inputs from the preceding layer, applies an activation function to this sum, and conveys the outcome to the subsequent layer. Throughout the training process, the weights and biases of the neurons are adjusted using an optimization algorithm to minimize the discrepancy between predicted and actual outputs. Activation functions can be either linear or nonlinear, and the number of layers and neurons in the network can be fine-tuned to enhance performance. FFNNs are particularly well-suited for complex problems involving large datasets, as they can learn to extract meaningful features from input data. In the present study, hyperparameters such as learning rate, batch size, and epochs are set to 0.0001, 32, and 150, respectively. The FFNN architecture is configured with 103 perceptron units in the input layer, 50 in the hidden layer, and 4 in the output layer as shown in Figure 5-3. The Rectified Linear Unit (ReLU) activation function is employed for

both the input and hidden layers, while the Softmax activation function is utilized in the output layer to provide class probabilities.

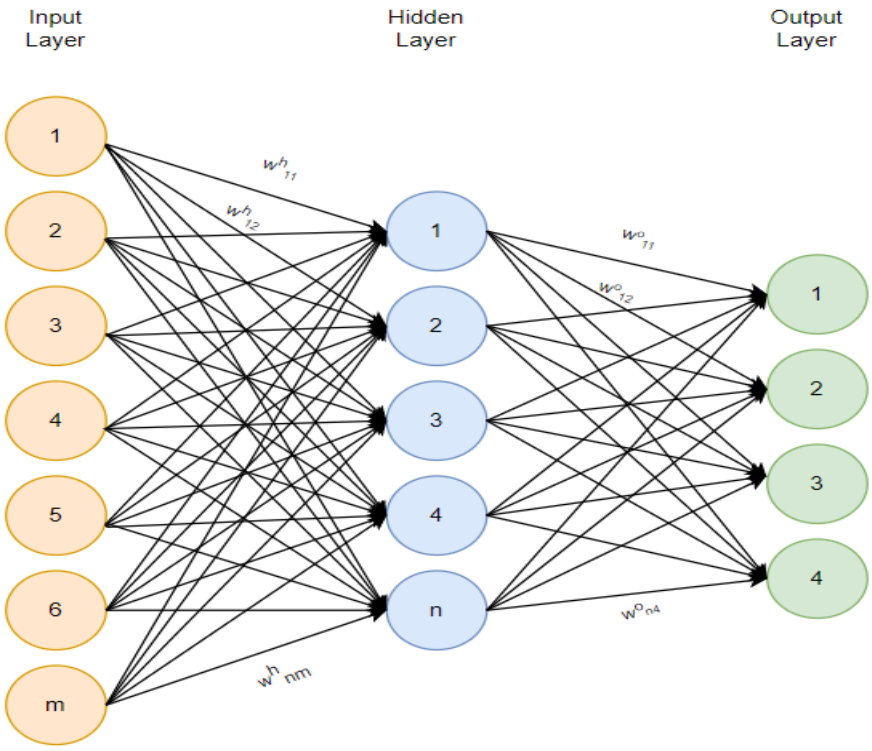

Figure 5-3 The FFNN architecture

5.4.5.5 One-Dimensional Convolution Neural Network (1D-CNN)

The 1D-CNN is a specialized type of neural network designed for processing time-series data. The architecture typically consists of one or more convolutional layers, followed by one or more fully connected layers. Convolutional layers apply a set of filters to the input data to extract relevant features, such as changes in frequency or amplitude over time. During the training process, the filter weights are adjusted to minimize the difference between the predicted and actual output. The fully connected layers then combine the features extracted by the convolutional layers to make a final prediction. 1D-CNNs are especially useful for detecting patterns in sequential data and can accommodate data with variable lengths. Figure 5-4 depicts each convolutional stage as a collection of learnable convolutional filters.

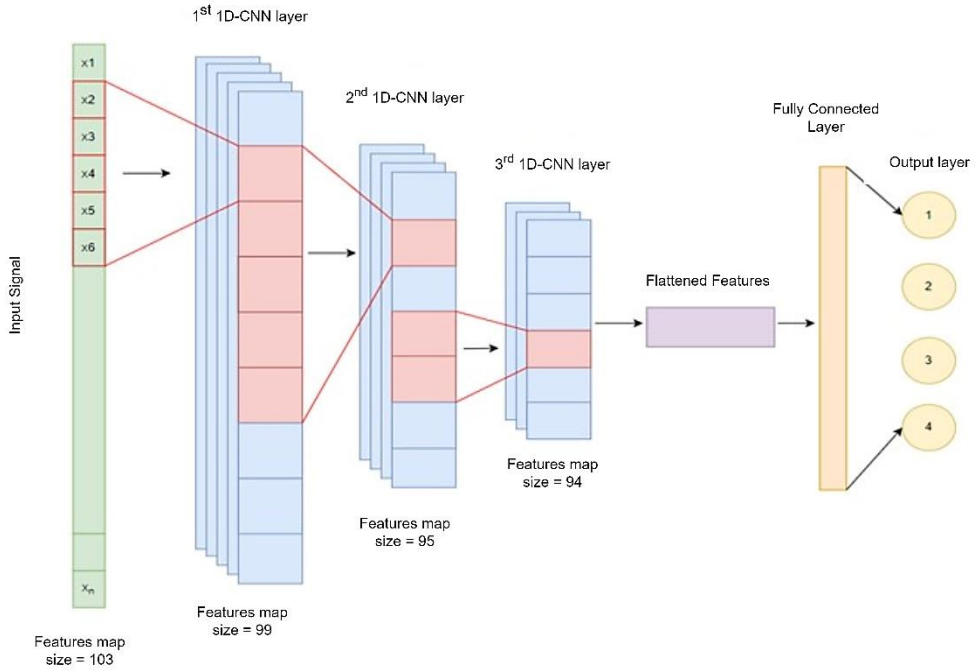

Figure 5-4 One-Dimensional Convolution Neural Network

A set of input signals $x2, x3, x4, x5\ and\ x6$, corresponding to the filter size, is chosen for the application of convolution. This process involves the utilization of convolutional filters, which are assigned specific weights. These filters are designed to extract high-level features from a given input signal by applying ReLU activation function. Given that there are xn features in the input signal, the output features of the first layer can be calculated using the following formula, taking into account the filter size (k) and stride (s):

$$Output\ size = \frac{input\ features - k}{s} + 1 \quad \textbf{(5-5)}$$

The outputs generated by the first layer are subsequently fed into a second convolution layer. This layer extracts features for the subsequent layer using the same formula as before. This iterative process reduces the spatial scale of the features extracted by the convolutional filters, while simultaneously emphasizing the salient features learned by each filter. The output of the second layer is then passed through a third convolution layer to generate the final convolution output features. As the input signals progress through the convolutional layers, the network becomes increasingly adept at learning problem-specific

characteristics. Upon reaching the final stage, the extracted features are flattened and passed through a densely connected hidden layer. This layer is connected to an output layer consisting of four nodes, which ultimately yields the final output.

In this study, the 1D-CNN architecture is configured with three convolutional layers, each followed by a dropout of 0.5 to prevent overfitting. The first convolutional layer consists of 128 filters, each with a kernel size of 5 and a ReLU activation function. The second convolutional layer comprises 64 filters, each with a kernel size of 5 and a ReLU activation function, followed by a dropout of 0.5. The third and final convolutional layer has 32 filters, each with a kernel size of 2 and a ReLU activation function, followed by a dropout of 0.5. After the convolutional layers, the features are flattened and passed through a fully connected layer with 128 nodes. The output layer consists of 4 nodes, corresponding to the four classes, and employs the Softmax activation function to yield class probabilities. The learning rate, batch size, and epochs hyperparameters are set to 0.0001, 32, and 150, respectively.

5.4.5.6 Long Short Term Memory Network (LSTM)

LSTM networks are designed to address the vanishing gradient problem and process sequential data with long-term dependencies. They use a memory cell that can store information over extended periods, a set of input, output, and forget gates to control the flow of information, and a set of cell state transformations to manipulate the stored information. The input gate controls the addition of new information to the memory cell, the forget gate determines the discarding of old information, and the output gate controls the information exposure to the subsequent layer. During training, backpropagation through time adjusts the weights of the gates and transformations to minimize the difference between the predicted and actual output.

To train the data using LSTM, the sequence of input data $x1, x2, x3, \ldots\ldots, xm$ is fed to the input gate of the LSTM layer in the network, as illustrated in Figure 5-5. The features generated by a single LSTM cell $(a1)$ are stored in the cell

memory and then passed to the next cell $(a2, a3, \ldots\ldots, am)$. The output of each cell is computed using the features passed on by the previous cell, and each cell provides the output through the output gate. The outputs provided by the output gate of each cell $(h1, h2, h3, \ldots, hm)$ are then multiplied by the weights of each cell. This stored data in memory is then used to derive new features or to observe the pattern in time series. After the LSTM layer, the data proceeds to the hidden layer by multiplying the weights of the hidden layer $(w_{11}^{h}, w_{12}^{h}, \ldots\ldots, w_{nm}^{h})$ with the output of each cell. The weighted output values of each cell are then combined to obtain a sum on each node of the hidden layer.

The LSTM network uses the same architecture and number of nodes as the previously discussed FFNN model. Instead of using the dense layer for input, the LSTM layer is employed for the LSTM model. The LSTM model offers an advantage over the FFNN because the biosignal data is in time-series format, allowing it to generalize results more effectively. In this study, the learning rate, batch size, and epochs hyperparameters are set to 0.0001, 32, and 150, respectively. Additionally, the LSTM architecture is configured by setting the LSTM layer, hidden layer, and output layer to 103, 50, and 4, respectively.

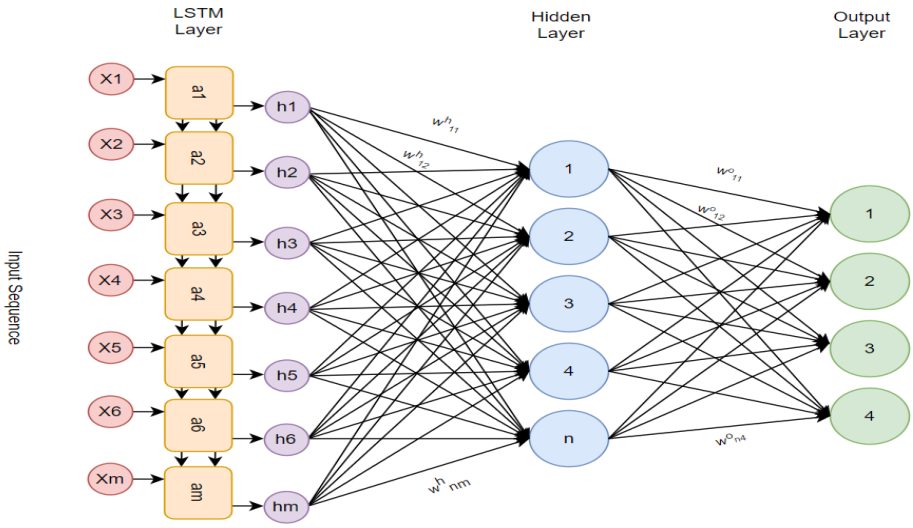

Figure 5-5 The LSTM Neural Network

5.4.5.7 The 1D-CNN+LSTM Architecture

This chapter presents a custom network that combines 1D-CNN and LSTM architectures. The 1D-CNN performs well with graphical and sparse data, while the LSTM demonstrates superior performance with time-series data. In contrast to Han et al. (2020), who trained a 2D-CNN with the EEG data focusing on the top three discriminative PSD frequency bands, along with ECG, EDA, and Resp. for LSTM, our architecture trained the 1D-CNN with all the frequency bands' PSD features, whereas the remaining three signals (i.e., ECG, Resp., and GSR) are trained with LSTM. While their model transformed EEG into a 2D format (topographical images) for CNN processing, our model directly processes the one-dimensional EEG signal, considering the entire frequency spectrum. This not only simplifies the preprocessing pipeline but also ensures that no spatial information inherent in the raw EEG signals is lost in translation to a 2D format.

Employing a 1D-CNN for EEG data alongside LSTM for non-brain signals capitalizes on the inherent strengths of each architecture, leading to several distinct advantages. This approach optimizes feature extraction by utilizing 1D-CNN to effectively capture temporal and spectral features from EEG data, crucial for interpreting cognitive states, while LSTMs excel in identifying complex temporal patterns and long-term dependencies within non-brain signals, offering a comprehensive analysis of physiological responses. Such a strategy not only ensures computational efficiency, facilitating quicker training and inference without sacrificing model performance but also enhances predictive accuracy by leveraging the complementary capabilities of both networks for a holistic understanding of both brain activity and peripheral physiological signals. Additionally, this dual-model framework allows for the separate optimization of models based on the unique characteristics of each signal type, enabling precise adjustments to improve overall model effectiveness.

The input sequences ($x1, x2, x3, \ldots, xn$) are fed to the 1D-CNN part of the model, as shown in Figure 5-6, and processed in a manner similar to that

described in Subsection 5.4.5.5. The remaining features $(x_{n+1}, x_{n+2}, \text{and } x_{n+3})$ are input to the LSTM portion of the model and processed as detailed in Subsection 5.4.5.6. The output features generated by both models are then concatenated and passed through a fully connected layer before reaching the output layer, consisting of 4 nodes to classify the four classes with the 'softmax' activation function.

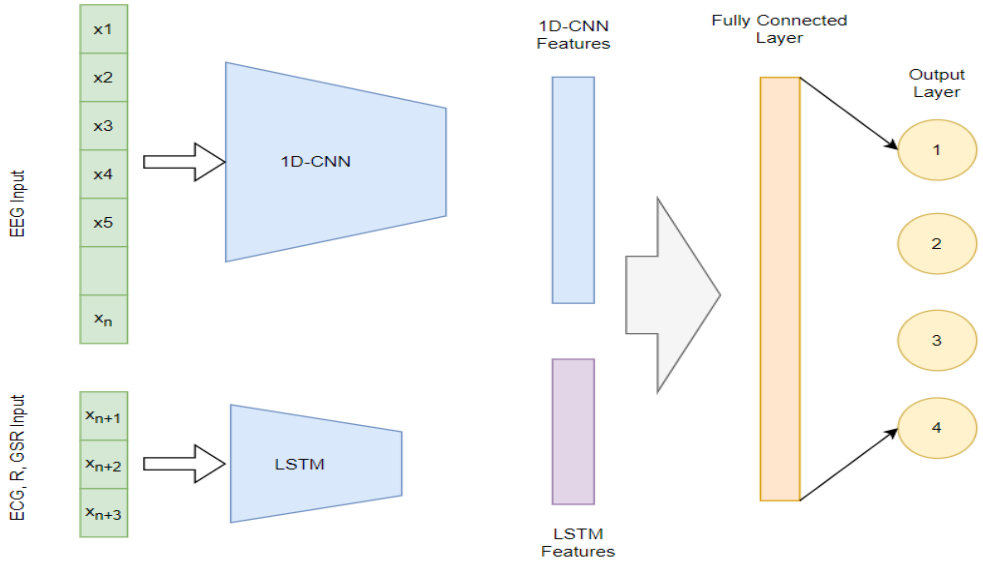

Figure 5-6 Overview of the proposed 1D-CNN+LSTM architecture

This architecture represents a modular approach, where each signal type is processed independently by the most suitable model for its characteristics. This specificity allows for targeted feature extraction and analysis, optimizing the model's ability to interpret each signal type accurately. In contrast, an integrated architecture where EEG, ECG, GSR, and Resp. signals are first processed through a series of 1D-CNN layers and subsequently through a series of LSTM layers adopts a sequential processing method. This sequential model attempts to universally apply convolutional layers to extract features across all signal types before using recurrent layers to analyse temporal sequences, potentially diluting the specificity of feature extraction for each signal type. While the sequential model can capture a wide array of features and temporal patterns in a single flow, it may not offer the same level of tailored analysis and efficiency as the modular approach. The modular approach ensures that each signal's

unique properties are addressed optimally, leading to potentially more accurate and computationally efficient predictions of mental states.

In addition, the choice to not consider Transformer models, despite their state-of-the-art performance in various domains, was informed by several factors. Transformers, which excel in capturing long-range dependencies within data, typically require large datasets and substantial computational resources for training. Given the relatively compact nature of the dataset (with flattened features for EEG, ECG, GSR, and Resp. signals) and the computational overhead associated with Transformers, the benefit of employing such models might not outweigh the increased complexity and resource demands, especially when the primary focus was on extracting meaningful temporal and spectral features from the EEG data and temporal patterns from ECG, GSR, and Resp. signals. Although this approach was not adopted in the current research, it represents a promising direction for future research.

5.5 Results and Discussion

In this section, the results of the proposed multimodal DL architecture are presented, and its performance is evaluated in comparison to various ensemble learning and DL models. Furthermore, the effectiveness of integrating CS with SMOTE to address data imbalance issues in the dataset is assessed.

The results and discussion are organized into three subsections. In subsection 5.5.1, the performance outcomes of the proposed architecture alongside other ensemble learning and DL models, both before and after incorporating CS, are presented. This comparison will provide a comprehensive understanding of each model's strengths and weaknesses. Subsection 5.5.2 focuses on evaluating the training and validation performance of the DL models, considering the impact of the sampling techniques on their performance. The convergence and generalization capabilities of these models before and after the utilization of CS are discussed. In subsection 5.5.3, the overall impact of CS on the performance of all the trained models is assessed. The influence of the combined approach of SMOTE and CS on the models’ performances are

analysed using the confusion matrix. This will provide a deeper understanding of the benefits and potential limitations of using this combined sampling technique.

5.5.1 Performance Comparison of ML and DL Models

In this subsection, the classification performance of ensemble and DL models is evaluated, using features extracted from EEG, ECG, Resp., and GSR signals. Welch's method and FFT were employed to generate 100 EEG features per channel, which were subsequently reduced to five features per channel using absolute PSD. The resulting combined dataset consisted of 32,867 epochs, each containing 103 features.

For model evaluation, the dataset was divided into 80% for training and 20% for testing for ensemble learning models (i.e., XGBoost, AdaBoost, and RF). Meanwhile, the DL models (i.e., FFNN, 1D-CNN, and CNN+LSTM) used a 60% training, 20% validation, and 20% testing split. The SMOTE method was applied to address class imbalance in the training data for all models. The performance results of these models, evaluated using the unseen testing dataset, are presented in Table 5-2. The testing dataset comprised 6,574 epochs, including 5,913 epochs of the NE class, 43 epochs of the SS class, 538 epochs of the CA class, and 80 epochs of the DA class.

Table 5-2 Classification performance of the pilots' mental states using only the SMOTE method (without CS).

Model	Mental Class	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	Support
XGBoost	NE		95.87	98.95	97.38	5913
	SS		100	6.97	13.04	43
	CA		83.84	71.37	77.10	538
	DA		40	5	8.88	80

	Macro Avg		79.92	45.57	49.10	6574
	Weighted Avg	94.94	94.23	94.94	94.09	6574
AdaBoost	NE		95.16	96.53	95.84	5913
	SS		50	9.30	15.68	43
	CA		64.33	63.38	63.85	538
	DA		13.15	6.25	8.47	80
	Macro Avg		55.66	43.86	45.96	6574
	Weighted Avg	92.15	91.34	92.15	91.63	6574
RF	NE		92.28	67.39	77.90	5913
	SS		2.11	51.16	4.05	43
	CA		32.47	23.42	27.21	538
	DA		24.18	25	4.41	80
	Macro Avg		32.32	41.74	28.39	6574
	Weighted Avg	63.17	85.70	63.17	72.37	6574
FFNN	NE		92.46	85.74	88.97	5913
	SS		11.11	4.65	6.55	43
	CA		20.33	36.43	26.09	538
	DA		0.91	1.25	1.05	80
	Macro Avg		31.20	32.01	30.67	6574
	Weighted Avg	80.14	84.91	80.14	82.22	6574
1D-CNN	NE		92.61	79.90	85.79	5913
	SS		2.59	4.65	3.33	43

	CA		17.42	42.19	24.66	538
	DA		4.34	5	4.65	80
	Macro Avg		29.24	32.93	29.60	6574
	Weighted Avg	75.41	84.79	75.41	79.26	6574
LSTM	NE		91.58	88.11	89.81	5913
	SS		1.38	2.32	1.73	43
	CA		19.64	24.90	21.96	538
	DA		3.81	6.25	4.73	80
	Macro Avg		29.10	30.39	29.56	6574
	Weighted Avg	81.38	84.03	81.38	82.64	6574
1D-CNN+LSTM	NE		91.79	92.84	92.31	5913
	SS		6.25	2.32	3.38	43
	CA		27.49	27.13	27.31	538
	DA		2.17	1.25	1.58	80
	Macro Avg		31.92	30.88	31.15	6574
	Weighted Avg	85.76	84.87	85.76	85.31	6574

As illustrated in Table 5-2, the XGBoost algorithm displayed the best performance among the evaluated models, followed by AdaBoost, 1D-CNN+LSTM, LSTM, FFNN, 1D-CNN, and RF. Both XGBoost and AdaBoost achieved high mean accuracies of 94.94% and 92.15%, respectively, while the RF model lagged behind with a mean accuracy of 63.17%. These results indicate that while ensemble methods employ multiple weak learners to create a more powerful model, they rely on distinct mechanisms and configurations, which lead to differences in performance. Regarding DL models, all of them demonstrated strong performance. The proposed 1D-CNN+LSTM model

achieved the highest mean accuracy of 85.76%. Although it was outperformed by XGBoost and AdaBoost, the incorporation of 1D-CNN in this domain is a contribution. The 1D-CNN has proven effective in other areas such as speech recognition and provides the advantage of computational efficiency.

While the SMOTE method was employed to balance the dataset, the majority of the trained models struggled to accurately detect the SS, CA, and DA classes. A closer examination of the precision, recall, and F1-score metrics for the NE class reveals that the models demonstrated exceptional detection performance for this class. However, performance for the other classes was considerably lower, as evidenced by the macro average values. Among the remaining classes, the CA class exhibited the second-best detection performance. This observation suggests that if the dataset were not as imbalanced, the models may have achieved better overall performance across all classes.

To investigate this hypothesis, cosine similarity was applied to the NE epochs. The CS method aims to reduce the number of epochs with high similarity between the rows of the dataset containing the CA and NE classes. This process resulted in a reduction of NE epochs from 29,561 to 6,327, leaving the dataset with a total of 9,633 epochs. The modified dataset was then divided into 80% training and 20% testing for the ensemble learning models (i.e., XGBoost, AdaBoost, and Random Forest), and 60% training, 20% validation, and 20% testing for the DL models (i.e., FFNN, 1D-CNN, and CNN+LSTM). After that, the SMOTE method was employed on the training dataset. Table 5-3 presents the performance results of these models, evaluated using the unseen testing dataset.

It is important to note that the updated testing dataset consists of 1,927 epochs, including 1,266 NE class epochs, 43 SS class epochs, 538 CA class epochs, and 80 DA class epochs. This modified dataset allows for a more balanced evaluation of the models' performance across all classes.

Table 5-3 Classification performance of the pilots' mental states using SMOTE and CS methods.

Model	Mental Class	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	Support
XGBoost	NE		94.48	97.47	95.95	1266
	SS		77.41	55.81	64.86	43
	CA		87.34	91.07	89.17	538
	DA		65.51	23.75	34.86	80
	Macro Avg		81.19	67.02	71.21	1927
	Weighted Avg	91.69	90.90	91.69	90.08	1927
AdaBoost	NE		92.42	94.47	93.43	1266
	SS		90.90	46.51	61.53	43
	CA		80.31	84.94	82.56	538
	DA		35.71	18.75	24.59	80
	Macro Avg		74.84	61.16	65.53	1927
	Weighted Avg	87.59	86.65	87.59	86.83	1927
RF	NE		85.57	91.86	88.60	1266
	SS		14.13	60.46	22.90	43
	CA		89.95	39.96	55.34	538
	DA		17.24	31.25	22.22	80
	Macro Avg		51.17	55.88	47.27	1927
	Weighted Avg	74.15	82.36	74.15	75.09	1927
FFNN	NE		86.92	92.41	89.58	1266

	SS		26.92	16.27	20.28	43
	CA		70.75	66.54	68.58	538
	DA		18.36	11.25	13.95	80
	Macro Avg		50.74	46.62	48.10	1927
	Weighted Avg	80.12	78.22	80.12	79.03	1927
1D-CNN	NE		84.75	91.31	87.90	1266
	SS		14.54	18.60	16.32	43
	CA		70.72	61.52	65.80	538
	DA		20	10	13.33	80
	Macro Avg		47.50	45.36	45.84	1927
	Weighted Avg	77.99	76.58	77.99	77.04	1927
LSTM	NE		85.99	92.65	89.20	1266
	SS		30	6.97	11.32	43
	CA		70.05	69.14	69.59	538
	DA		13.63	3.75	5.88	80
	Macro Avg		49.92	43.13	44	1927
	Weighted Avg	80.48	77.29	80.48	78.53	1927
1D-CNN+LSTM	NE		83.06	92.57	87.56	1266
	SS		26.08	13.95	18.18	43
	CA		70.10	59.29	64.24	538
	DA		13.15	6.25	8.47	80
	Macro Avg		48.10	43.01	44.61	1927

Weighted Avg	77.94	75.27	77.94	76.22	1927
Weighted Avg	77.94	75.27	77.94	76.22	1927

Using the same hyperparameters and configuration settings, the performance of the ensemble and DL models trained on the new dataset is displayed in Table 5-3. Once again, the XGBoost algorithm achieved the highest performance, followed by AdaBoost, LSTM, FFNN, 1D-CNN, CNN+LSTM, and RF. These results indicate that XGBoost is particularly suitable for this specific task, outperforming the other models. Interestingly, the proposed 1D-CNN+LSTM model did not perform as well on the new dataset as it did on the original dataset. This can be attributed to the fact that DL models typically perform better when trained with larger datasets.

The application of CS method considerably improved the detection performance for each mental state, as evidenced by the macro average values shown in Table 5-3. This improvement is further corroborated by examining the precision, recall, and F1-score of each model. These findings confirm the hypothesis that the skewed distribution of the dataset was one of the factors impacting the models' performance. Notably, employing CS to remove the NE epochs with similar row data as the CA class substantially enhanced the detection performance of the CA mental state, especially for the DL models.

It could be argued that the performance of the models trained on the original dataset, as displayed in Table 5-2, is superior to that of the models trained on the modified dataset shown in Table 5-3. However, this comparison is complicated by the different testing dataset sizes, as indicated in the support column. To accurately evaluate the performance of the models trained with the original dataset, these models were tested on a dataset identical to the testing dataset used for assessing the models trained on the modified dataset. Table 5-4 displays the classification performance of the models that were trained using the original dataset and evaluated with the updated testing dataset. This approach allows for a fair comparison between the models, accounting for differences in testing dataset sizes.

Table 5-4 Classification performance of the pilots' mental states using the updated testing dataset.

Model	Accuracy	Precision	Recall	F1 score
XGBoost	85.21	84.78	85.21	82.6
AdaBoost	81.94	80.84	81.94	79.37
RF	52.93	65.95	52.93	56.08
FNN	65.39	51.99	65.39	52.44
1D-CNN	64.45	61.72	64.45	62.89
LSTM	65.23	60.65	65.23	61.0
1D-CNN+LSTM	68.34	63.69	68.34	63.09

The results presented in Table 5-4, illustrating the comparative performance of DL and ML methods, merit a detailed discussion. It was observed that ML methods, particularly XGBoost and AdaBoost, outperformed the DL models, including the 1D-CNN+LSTM architecture. Several factors contribute to this phenomenon, which are important to consider in the context of mental state detection using physiological signals.

1. **Data Size and Complexity**: DL models are typically more data-intensive compared to ML models. They require large datasets to effectively learn and generalize complex patterns, especially in domains involving intricate signals like EEG. The size of our dataset, while adequate, may not have been sufficient to fully exploit the DL models' capacity to capture and learn the nuances present in EEG and other physiological signals.
2. **Model Suitability to Data Characteristics**: ML models like XGBoost and AdaBoost are particularly effective in extracting patterns from tabular data and can provide robust results even with relatively smaller datasets. These

models are capable of identifying simpler, yet highly predictive, features, which seemed to have aligned well with the characteristics of our dataset.

3. **Impact of Data Resampling on Model Performance**: The application of data resampling techniques, such as SMOTE and Cosine Similarity, may influence DL and ML models differently. These resampling methods, while addressing data imbalance, might also modify the data distribution in ways that are more conducive to the learning patterns of ML models. This effect could potentially lead to a disadvantage for DL models, especially if the resampled data loses some of the intricate patterns that DL models are adept at capturing.

Despite the observed performance differences in this specific study, the incorporation of 1D-CNN in this domain is a contribution. The 1D-CNN has proven effective in other areas such as speech recognition and provides the advantage of computational efficiency. Also, DL models have the advantage of automatic feature extraction and the ability to model non-linear and high-dimensional relationships. This is particularly relevant for tasks involving temporal data like EEG signals.

The findings from this study underscore the importance of a balanced approach, considering both ML and DL methods based on the dataset's nature and the research goals. It also opens avenues for further research, particularly in exploring strategies to enhance the performance of DL models with the available dataset or in scenarios where larger and more diverse datasets are accessible.

In short, while ML models demonstrated superior performance in our study, the potential of DL in mental state detection, especially with advancements in data collection and model architecture, should not be underestimated. Future research might focus on overcoming the current limitations and harnessing the full capabilities of DL models in this evolving field.

5.5.2 Training and Validation Analysis of DL Models

This study developed four distinct DL models to detect the AHPLS states. In addition to presenting the performance metrics of the DL models in Table 5-2 and Table 5-3, the learning curves (i.e., accuracy and loss curves) for each model are also provided. Figure 5-7 (A), (B), (C), and (D) display the accuracy and loss curves of the FFNN, 1D-CNN, LSTM, and 1D-CNN+LSTM models, respectively, prior to the application of CS. In general, all the DL models demonstrated strong performance, as evidenced by the increasing training and validation accuracies and the decreasing training and validation losses as the models learned.

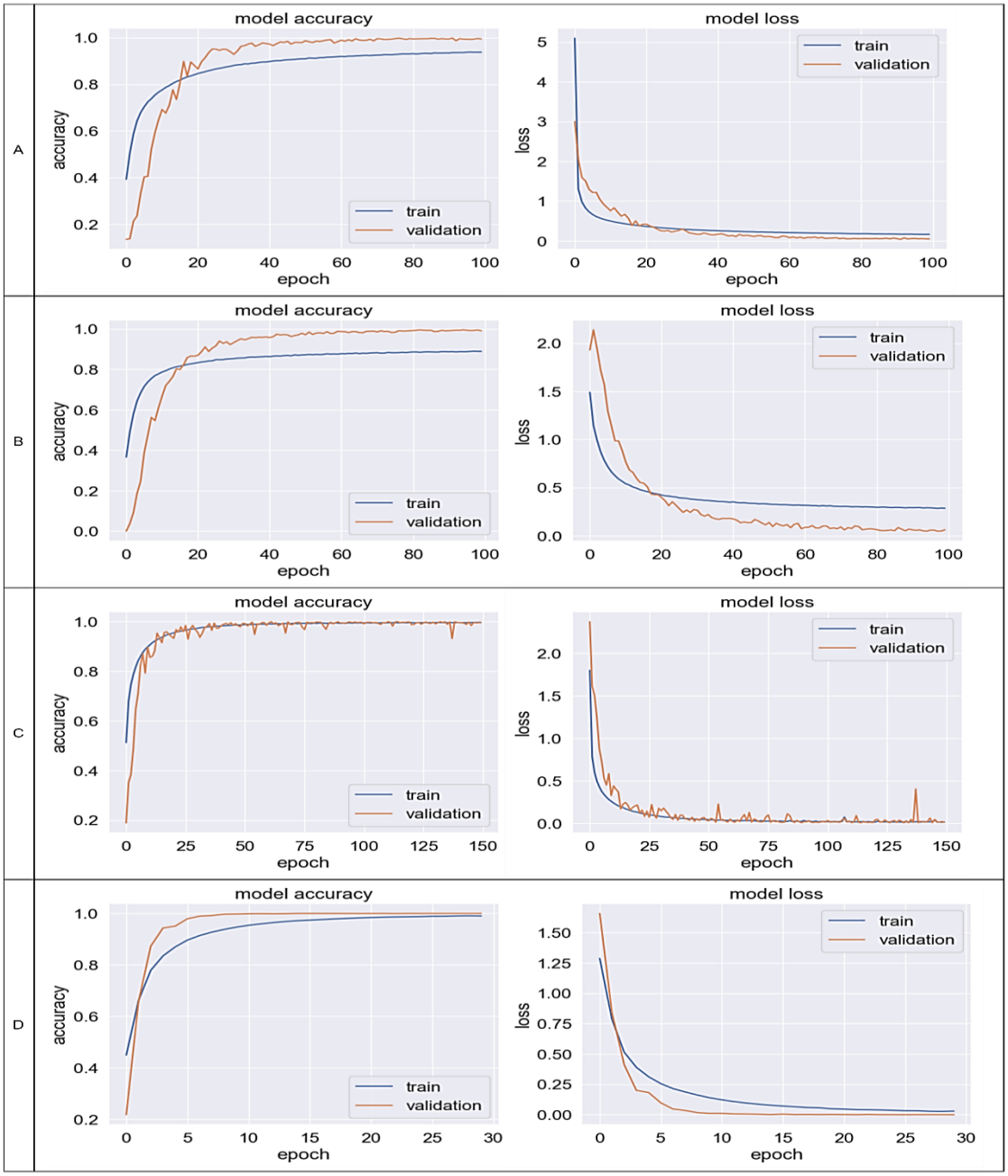

Figure 5-7 The DL models' learning curves before incorporating the CS method

Figure 5-7 (A) shows that the validation accuracy and loss are slightly better than the training accuracy and loss, which typically indicates that the training data is somewhat more challenging to model than the validation data. This is likely due to the non-linearity of the dataset. However, this is not the case for the 1D-CNN and 1D-CNN+LSTM models shown in Figure 5-7 (B) and (D), as dropout was used during their training. During the training process, a

percentage of the features are set to zero, while all features are used during validation. This results in higher validation accuracy, suggesting that the model is more robust. Although the LSTM model's accuracy and loss curves in Figure 5-7 (C) display negligible differences between training and validation, as the model is fully converged, the fluctuations in the validation data imply that the model is not generalizing well to the validation data. Consequently, among all the DL models, the proposed 1D-CNN+LSTM model is considered the best for the dataset prior to the incorporation of CS.

The DL models were also trained on the modified dataset after applying the CS method. Figure 5-8 (A), (B), (C), and (D) depict the accuracy and loss curves of the FFNN, 1D-CNN, LSTM, and 1D-CNN+LSTM models, respectively, after training them on the modified dataset. All the DL models demonstrated strong performance, as indicated by the increasing training and validation accuracies and the decreasing training and validation losses as the models learned.

Examining the FFNN model's loss curve in Figure 5-8 (A), it is evident that it is a superb curve, as the training and validation losses initially correlated, then diverged slightly, and finally converged again. Similarly, the LSTM model shown in Figure 5-8 (C) and the 1D-CNN+LSTM model shown in Figure 5-8 (D) displayed good loss curves, as the training and validation curves exhibit minor differences. The fluctuations in the validation data suggest that the models were not generalized enough to work on different data, such as the validation data. Regarding the 1D-CNN model depicted in Figure 5-8 (B), the validation data appears unrepresentative compared to the training data; however, they begin to converge at the end. This implies that training the model on more epochs might yield better convergence. The reason behind this trend is likely the decrease in the number of training and validation samples compared to the old dataset, which justifies the observed behaviour. It is crucial to highlight that the observed variations are statistical in nature rather than systematic. As a result, it could be argued that the proposed 1D-CNN+LSTM model demonstrates promising and strong performance for the dataset both before and after the application of CS.

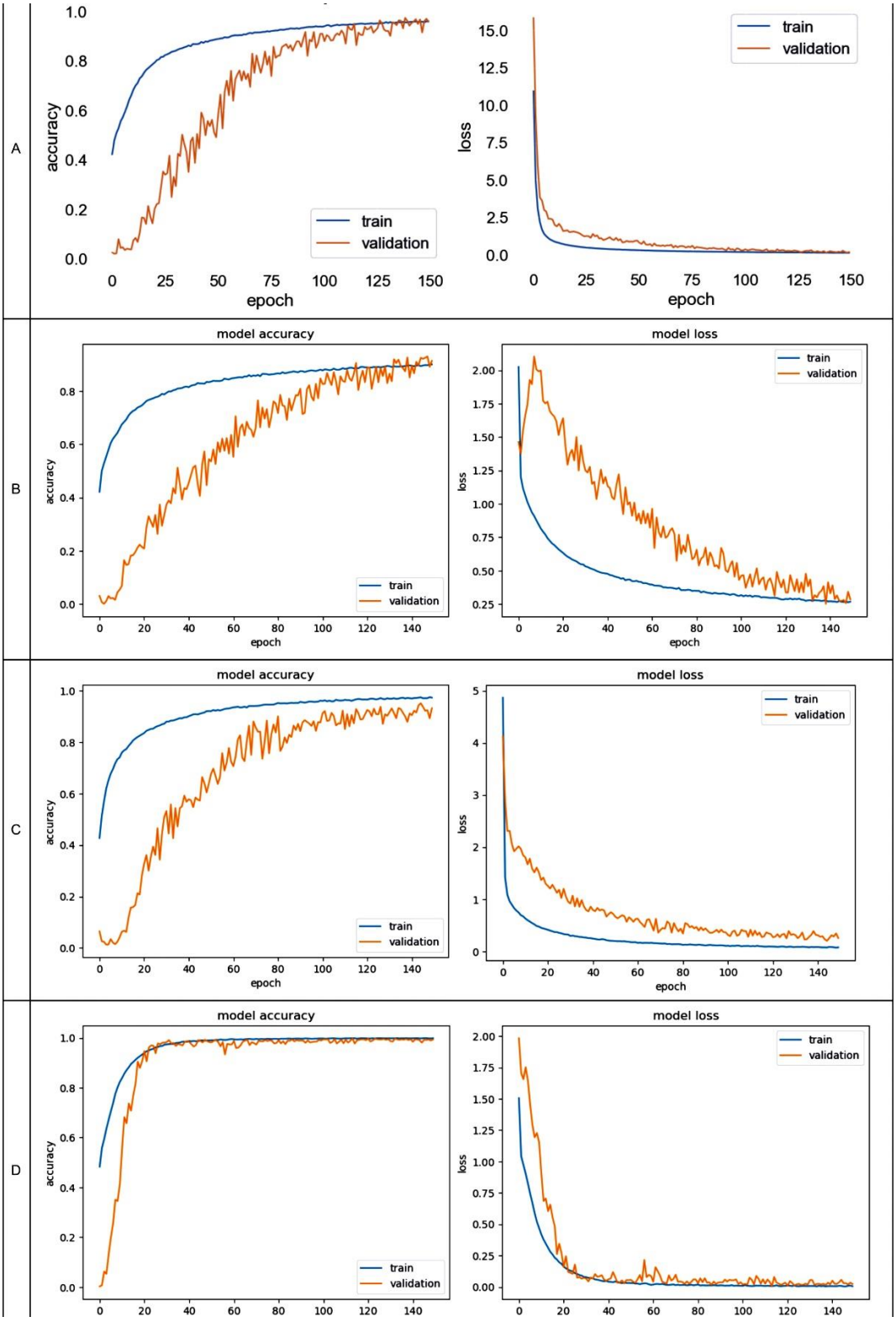

Figure 5-8 The DL models' learning curves after incorporating the CS method

5.5.3 Impact of Cosine Similarity on Model Performance: Confusion Matrix Analysis

To gauge the efficacy of the cosine similarity method, an identical testing dataset was utilized to measure the performance of the models trained both prior to and following the application of CS. The level of confusion produced by the models, before and after implementing the CS, was computed.

Figure 5-9 displays the confusion matrices for the models before applying CS. Specifically, Figure 5-9 (A) to (H) depict the confusion matrices for the XGBoost, AdaBoost, RF, FFNN, 1D-CNN, LSTM, and 1D-CNN+LSTM models, respectively. In contrast, Figure 5-10 presents the confusion matrices for the models after incorporating CS, where Figure 5-10 (A) to (H) illustrate the confusion matrices for the same models. The values of the diagonal elements in the matrices indicate the percentage of accurately predicted classes. This comparison allows for a more comprehensive understanding of the impact of CS on model performance.

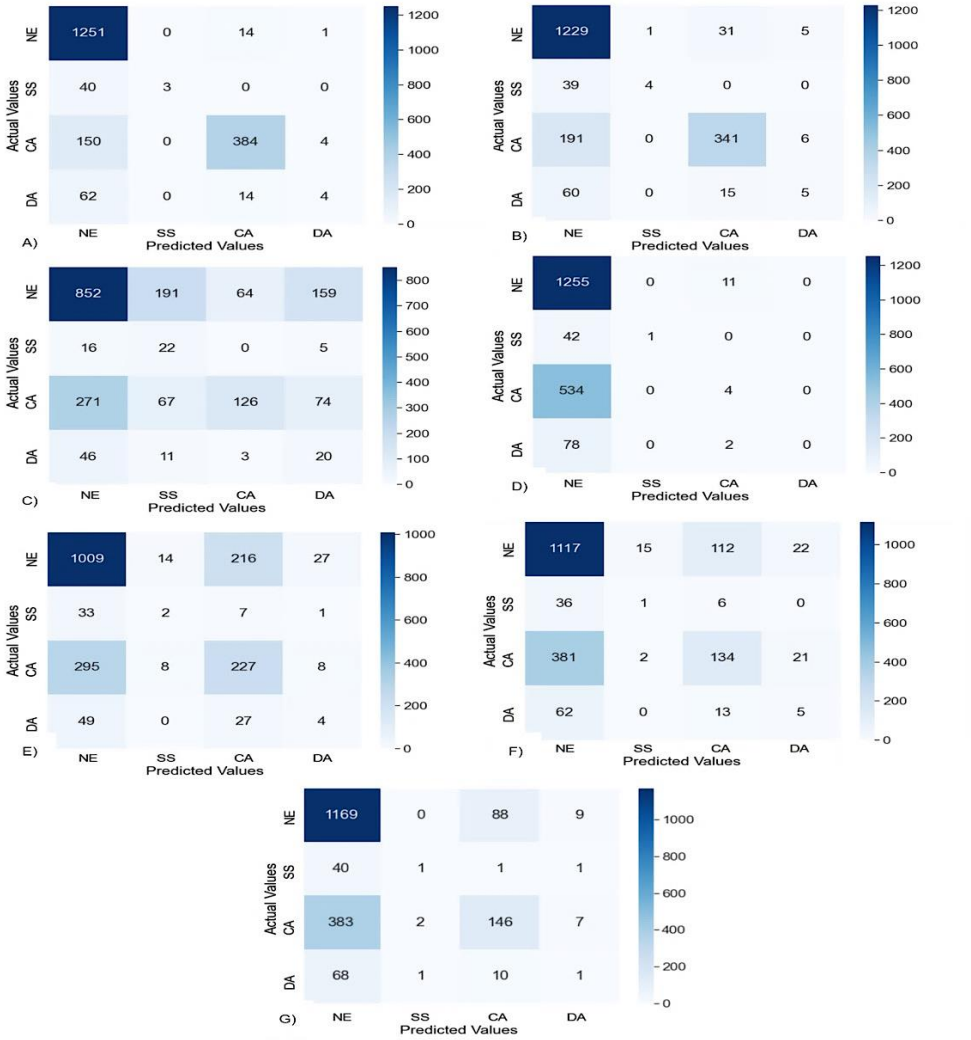

Figure 5-9 Confusion matrices for the models before incorporating CS method

Upon comparing the confusion matrices of the models before and after incorporating the CS method, as illustrated in Figure 5-9 and Figure 5-10, a significant improvement in model performance was observed. The NE and CA states were detected with relatively higher performance in almost all the models before employing the CS method, compared to the SS and DA states. After incorporating CS, more SS and DA samples were identified.

A few exceptions were noted, however. For the NE class, the XGBoost, AdaBoost, and FFNN models performed slightly better before using cosine similarity. This can be attributed to the reduction in the number of NE class samples, which affected the performance of these three models. Similarly, the LSTM model performed marginally better before applying cosine similarity,

detecting 5 out of 80 samples correctly, while only 3 samples were correctly predicted after using cosine similarity. This decrease in the identification of the DA state is hypothesized to be due to the configuration of the LSTM model.

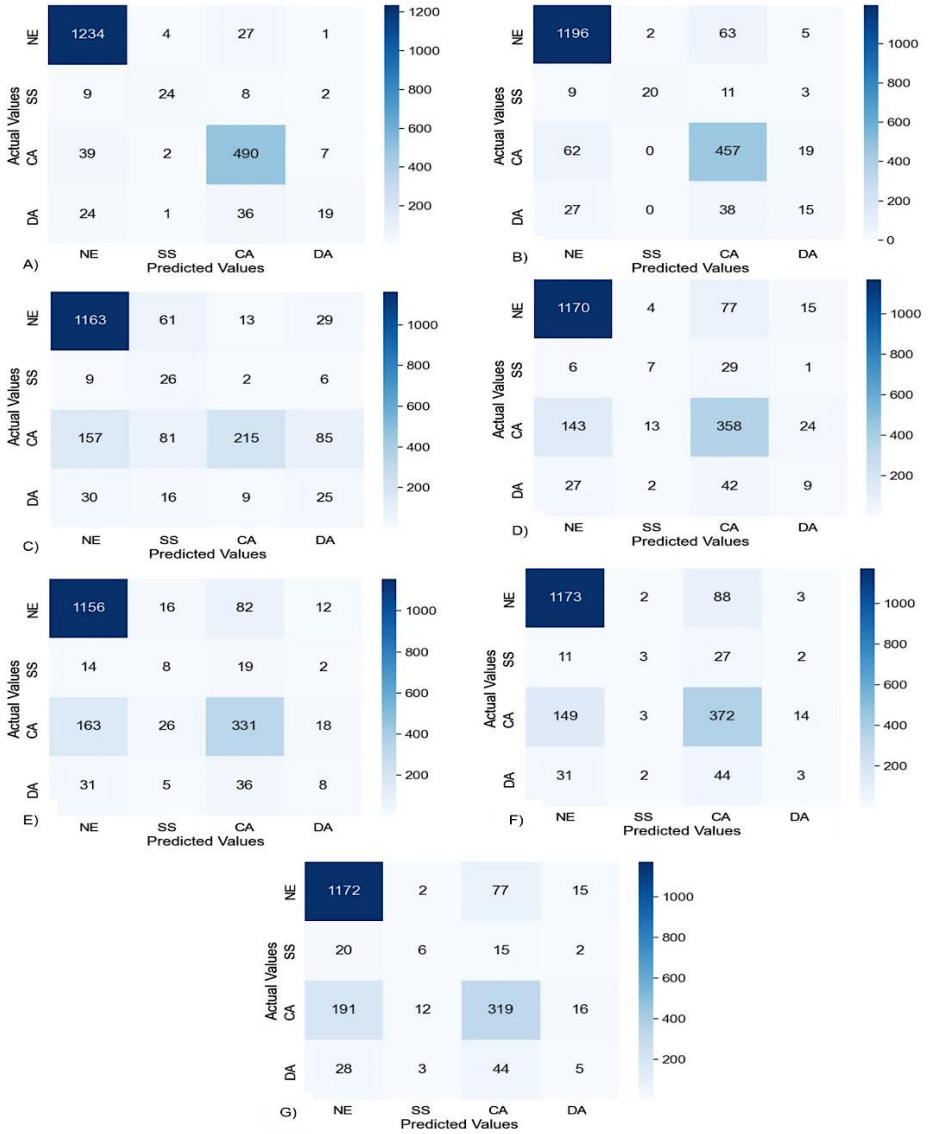

Figure 5-10 Confusion matrices for the models after incorporating CS method

In summary, this study presented a comprehensive evaluation of ensemble and DL models for detecting mental states using multimodal physiological signals. The performance of these models was assessed before and after the application of CS in conjunction with SMOTE to address the data imbalance issue. The results revealed that the XGBoost algorithm consistently

outperformed other models, while the proposed 1D-CNN+LSTM model demonstrated considerable potential as a DL solution.

Upon analysing the performance metrics, it was evident that the data imbalance issue had a significant impact on the models' ability to detect specific mental states. The implementation of CS led to a considerable improvement in the detection performance of each mental state, particularly for the CA class. This finding confirmed the hypothesis that data skewness was a major factor affecting the models' performance.

The learning curves of the DL models, both before and after the application of CS, displayed robust performance, with the proposed 1D-CNN+LSTM model deemed suitable for the given dataset. When comparing the confusion matrices before and after the use of CS, an overall improvement in model performance was observed. However, some models, such as XGBoost, AdaBoost, FFNN, and LSTM, exhibited slightly better performance in detecting certain mental states before applying CS, which could be attributed to the reduction in the number of samples for specific classes.

The incorporation of CS and SMOTE proved to be effective in addressing data imbalance and improving the performance of the models in detecting mental states. Among the considered models, the XGBoost algorithm and the proposed 1D-CNN+LSTM model emerged as the most promising solutions for the given dataset.

5.6 Conclusion

This research has made significant strides in addressing the critical need for detecting pilots' mental states, particularly AHPLS, to enhance aviation safety and reduce the likelihood of accidents. By employing a multimodal approach that combines EEG with non-brain signals such as ECG, GSR, and Resp., the study has developed a robust DL architecture that effectively fuses 1D-CNN and LSTM models.

The research has also tackled the challenge of data imbalance, which is prevalent in real-world datasets and often results in biased models with poor detection performance for underrepresented mental states. By incorporating data resampling techniques, including downsampling using the CS method and oversampling using the SMOTE method, the study has successfully created more balanced datasets, which led to improved model performance.

As part of future work, further refinement of the models' performance will be carried out, and the training dataset will be enlarged to enhance the generalization capability of the models. Additionally, the possibility of extracting other meaningful features from the multimodal sensor data will be explored to further enhance the accuracy and robustness of the classification model.

Overall, this study highlights the potential of using multimodal sensor data and the proposed 1D-CNN+LSTM model for classifying pilots' mental states. The findings contribute to the growing body of literature on human factors in aviation and have implications for the development of real-time mental state monitoring systems for aviation safety applications.

5.7 Appendices

5.7.1 Appendix A: Data and Reproducibility Code

In the interest of promoting transparency and reproducibility, the data utilised in this chapter, along with the associated code for analyses, have been made publicly accessible. The dataset and the code for replicating the analyses can be found under the Digital Object Identifier (DOI): https://doi.org/10.17862/cranfield.rd.24156345

REFERENCES

Ahlawat, K., & Singh, A. P. (2020, 2020//). Human Activity Recognition in Imbalanced Big Data Using Fuzzy Rule-Based Classification System. Soft Computing and Signal Processing, Singapore.

Alani, A. A., Cosma, G., & Taherkhani, A. (2020, 19-24 July 2020). Classifying Imbalanced Multi-modal Sensor Data for Human Activity Recognition in a Smart Home using Deep Learning. 2020 International Joint Conference on Neural Networks (IJCNN),

Alreshidi, I., Moulitsas, I., & Jenkins, K. W. (2023). Multimodal Approach for Pilot Mental State Detection Based on EEG. *Sensors (Basel)*, *23*(17). https://doi.org/10.3390/s23177350

Alreshidi, I. M., Moulitsas, I., & Jenkins, K. W. (2022). Miscellaneous EEG Preprocessing and Machine Learning for Pilots' Mental States Classification: Implications. 2022 The 6th International Conference on Advances in Artificial Intelligence,

Boksem, M. A., & Tops, M. (2008). Mental fatigue: costs and benefits. *Brain Res Rev*, *59*(1), 125-139. https://doi.org/10.1016/j.brainresrev.2008.07.001

Carreiras, C., Alves, A. P., Lourenço, A., Canento, F., Silva, H., & Fred, A. (2015). BioSPPy - Biosignal Processing in Python. In.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321-357. https://doi.org/DOI 10.1613/jair.953

Giraudet, L., St-Louis, M.-E., Scannella, S., & Causse, M. (2015). P300 Event-Related Potential as an Indicator of Inattentional Deafness? *PLoS One*, *10*(2), e0118556. https://doi.org/10.1371/journal.pone.0118556

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hamalainen, M. (2013). MEG and EEG data analysis with MNE-Python. *Front Neurosci*, *7*, 267. https://doi.org/10.3389/fnins.2013.00267

Haibo, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284. https://doi.org/10.1109/tkde.2008.239

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220-239. https://doi.org/10.1016/j.eswa.2016.12.035

Han, S. Y., Kim, J. W., & Lee, S. W. (2019, 18-20 Feb. 2019). Recognition of Pilot's Cognitive States based on Combination of Physiological Signals.

2019 7th International Winter Conference on Brain-Computer Interface (BCI),

Han, S. Y., Kwak, N. S., Oh, T., & Lee, S. W. (2020). Classification of pilots' mental states using a multimodal deep learning network. *Biocybernetics and Biomedical Engineering*, *40*(1), 324-336. https://doi.org/10.1016/j.bbe.2019.12.002

Hankins, T. C., & Wilson, G. F. (1998). A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. *Aviat Space Environ Med*, *69*(4), 360-367. https://www.ncbi.nlm.nih.gov/pubmed/9561283

Harrivel, A. R., Liles, C., Stephens, C. L., Ellis, K. K., Prinzel, L. J., & Pope, A. T. (2016). Psychophysiological Sensing and State Classification for Attention Management in Commercial Aviation. AIAA Infotech @ Aerospace,

Harrivel, A. R., Stephens, C. L., Milletich, R. J., Heinich, C. M., Last, M. C., Napoli, N. J., Abraham, N., Prinzel, L. J., Motter, M. A., & Pope, A. T. (2017). Prediction of Cognitive States during Flight Simulation using Multimodal Psychophysiological Sensing. AIAA Information Systems-AIAA Infotech @ Aerospace,

Jiang, H., Xu, K., Chen, X., Wang, Q., Yang, Y., Fu, C., Guo, X., Chen, X., & Yang, J. (2020). The Neural Underpinnings of Emotional Conflict Control in Pilots. *Aerosp Med Hum Perform*, *91*(10), 798-805. https://doi.org/10.3357/AMHP.5618.2020

Jiang, S., Chen, W., Kang, Y., Liu, J., & Kuang, W. (2021). Identifying Cognitive Mechanism Underlying Situation Awareness of Pilots' Unsafe Behaviors Using Quantitative Modeling. *International Journal of Environmental Research and Public Health*, *18*(6).

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0192-5

Kaur, H., Pannu, H. S., & Malhi, A. K. (2019). A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. *ACM Comput. Surv.*, *52*(4), Article 79. https://doi.org/10.1145/3343440

Khanna, A., Pascual-Leone, A., Michel, C. M., & Farzan, F. (2015). Microstates in resting-state EEG: current status and future directions. *Neurosci Biobehav Rev*, *49*, 105-113. https://doi.org/10.1016/j.neubiorev.2014.12.010

Kim, K. H., Bang, S. W., & Kim, S. R. (2004). Emotion recognition system using short-term monitoring of physiological signals. *Med Biol Eng Comput*, *42*(3), 419-427. https://doi.org/10.1007/BF02344719

Koelstra, S., Muhl, C., Soleymani, M., Jong-Seok, L., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., & Patras, I. (2012). DEAP: A Database for Emotion Analysis ;Using Physiological Signals. *IEEE Transactions on Affective Computing*, *3*(1), 18-31. https://doi.org/10.1109/t-affc.2011.15

Minguillon, J., Lopez-Gordo, M. A., & Pelayo, F. (2017). Trends in EEG-BCI for daily-life: Requirements for artifact removal. *Biomedical Signal Processing and Control*, *31*, 407-418. https://doi.org/10.1016/j.bspc.2016.09.005

Oehling, J., & Barry, D. J. (2019). Using machine learning methods in airline flight data monitoring to generate new operational safety knowledge from existing data. *Safety Science*, *114*, 89-104. https://doi.org/10.1016/j.ssci.2018.12.018

Pamplona-Beron, L. E., Henao Baena, C. A., & Calvo-Salcedo, A. F. (2021). Human activity recognition using penalized support vector machines and Hidden Markov Models in multimodal systems. *Revista Facultad de Ingeniería Universidad de Antioquia*. https://doi.org/10.17533/udea.redin.20210532

Pan, T., Wang, H., Si, H., Li, Y., & Shang, L. (2021). Identification of Pilots' Fatigue Status Based on Electrocardiogram Signals. *Sensors (Basel)*, *21*(9). https://doi.org/10.3390/s21093003

Puspaningrum, E. Y., Wahid, R. R., Amaliyah, R. P., & Nisa', C. (2020, 14-16 Oct. 2020). Alzheimer's Disease Stage Classification using Deep Convolutional Neural Networks on Oversampled Imbalance Data. 2020 6th Information Technology International Seminar (ITIS),

Sharma, A., & Verbeke, W. J. M. I. (2020). Improving Diagnosis of Depression With XGBOOST Machine Learning Model and a Large Biomarkers Dutch Dataset (n = 11,081) [Brief Research Report]. *Frontiers in Big Data*, *3*. https://www.frontiersin.org/articles/10.3389/fdata.2020.00015

Smith, J. O. (2011). *Spectral Audio Signal Processing*. W3K Publishing. http://ccrma.stanford.edu/~jos/sasp/

Sun, C., Cui, H., Zhou, W., Nie, W., Wang, X., & Yuan, Q. (2019). Epileptic Seizure Detection with EEG Textural Features and Imbalanced Classification Based on EasyEnsemble Learning. *International Journal of Neural Systems*, *29*(10), 1950021. https://doi.org/10.1142/S0129065719500217

Terwilliger, P. S., Jack; Walker, Shannon; Harrivel, Angela. (2020). *A ResNet Autoencoder Approach for Time Series Classification of Cognitive State* MODSIM,

Tripathi, S., Acharya, S., Sharma, R., Mittal, S., & Bhattacharya, S. (2017). Using Deep and Convolutional Neural Networks for Accurate Emotion Classification on DEAP Data. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(2), 4746-4752. https://doi.org/10.1609/aaai.v31i2.19105

Walmsley, S., & Gilbey, A. (2016). Cognitive Biases in Visual Pilots' Weather-Related Decision Making. *Applied Cognitive Psychology*, *30*(4), 532-543. https://doi.org/https://doi.org/10.1002/acp.3225

Wei-Long, Z., & Bao-Liang, L. (2015). Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks. *IEEE Transactions on Autonomous Mental Development*, *7*(3), 162-175. https://doi.org/10.1109/tamd.2015.2431497

Weiss, G. M., & Provost, F. (2001). *The effect of class distribution on classifier learning: an empirical study*.

WELCH, P. D. (1967). The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms. *IEEE TRANSACTIONS ON AUDIO AND ELECTROACOUSTICS*.

Wu, D., Wang, Z., Chen, Y., & Zhao, H. (2016). Mixed-kernel based weighted extreme learning machine for inertial sensor based human activity recognition with imbalanced dataset. *Neurocomputing*, *190*, 35-49. https://doi.org/https://doi.org/10.1016/j.neucom.2015.11.095

Yen, J. R., Hsu, C. C., Yang, H., & Ho, H. (2009). An investigation of fatigue issues on different flight operations. *Journal of Air Transport Management*, *15*(5), 236-240. https://doi.org/10.1016/j.jairtraman.2009.01.001

6 Illuminating the Neural Landscape of Pilot Mental States: A Convolutional Neural Network Approach with SHAP Interpretability

6.1 Abstract

Predicting pilots' mental states is a critical challenge in aviation safety and performance, with electroencephalogram data offering a promising avenue for detection. However, the interpretability of machine learning and deep learning models, which are often used for such tasks, remains a significant issue. This study aims to address these challenges by developing an interpretable model to detect four mental states—channelized attention, diverted attention, startle/surprise, and normal state in pilots using EEG data. The methodology involves training a convolutional neural network on power spectral density features of EEG data from 17 pilots. The model's interpretability is enhanced through the use of SHapley Additive exPlanations values, which identify the top 10 most influential features for each mental state. The results demonstrate high performance in all metrics, with an average accuracy of 96%, a precision of 96%, recall of 94%, and an F1-score of 95%. An examination of the effects of mental states on EEG frequency bands further elucidates the neural mechanisms underlying these states. The innovative nature of this study lies in its combination of high-performance model development, improved interpretability, and in-depth analysis of the neural correlates of mental states. This approach not only addresses the critical need for effective and interpretable mental state detection in aviation, but also contributes to our understanding of the neural underpinnings of these states. This study thus represents a significant advancement in the field of EEG-based mental state detection.

6.2 Introduction

The human brain, an intricate network of billions of neurones, is a dynamic system that constantly generates electrical activity. This electrical activity, which

reflects the complex interplay of neural processes, can be measured and analysed using electroencephalography (EEG). Since the advent of EEG in the early 20th century, these signals have been extensively studied for their potential to provide insights into various cognitive states, mental conditions, and neurological disorders (Buzsaki et al., 2012; Cohen, 2017; Schomer & Lopes da Silva, 2017). In recent years, the use of EEG in cognitive neuroscience has surged, driven by advancements in signal processing techniques and the development of portable and wearable EEG devices (Mihajlovic et al., 2015). The non-invasive nature of EEG, its relatively low cost, high temporal resolution, and the possibility of real-time monitoring make it a particularly attractive tool for studying brain dynamics in various contexts (Makeig et al., 2012; Teplan, 2002). One such context is the field of aviation, where understanding and monitoring the mental states of pilots is of paramount importance. The mental state of a pilot can significantly influence his decision-making ability, reaction times, and overall performance, particularly in high-stakes or stressful situations (Borghini et al., 2014). Therefore, the ability to accurately detect and classify different mental states based on EEG data could provide a valuable tool for enhancing safety and performance in aviation.

Different mental states are associated with different patterns of brain activity, which can be captured in the frequency domain of EEG signals as characteristics of power spectral density (PSD) (Klimesch, 1999). These features represent the distribution of signal power over various frequency bands, such as delta (δ), theta (θ), alpha (α), beta (β), and gamma (γ), each of which is associated with different cognitive processes and mental states (Basar et al., 2001). For instance, δ waves are typically associated with deep sleep or relaxation, θ waves with creativity and insight, α waves with relaxed alertness, β waves with active thinking and focus, and γ waves with higher mental activity and perception (Basar et al., 2001; Harmony, 2013). The analysis of these frequency bands can provide a window into the cognitive processes underlying different mental states, making them a valuable tool for mental state classification (Klimesch, 1999).

In recent years, the field of machine learning (ML), particularly Convolutional Neural Networks (CNN), has made significant strides in the analysis of EEG data for mental state classification (Schirrmeister et al., 2017). CNN models have demonstrated their efficacy in handling high-dimensional data, such as EEG signals, by automatically learning hierarchical representations from raw data. This ability to learn and extract salient features from raw data without the need for manual feature extraction is a significant advantage in EEG analysis, where the selection of appropriate features is often challenging (Bashivan et al., 2016; Oehling & Barry, 2019; Schirrmeister et al., 2017).

However, while CNN models have shown promise in terms of performance, understanding the decision-making process of these models remains a challenge. To address this, we employ SHapley Additive exPlanations (SHAP) values, a powerful tool for interpreting ML models. SHAP values provide a measure of the contribution of each feature to the model's prediction, thereby offering insights into the model's decision-making process (Chen et al., 2018; Lundberg & Lee, 2017).

The novelty of this study lies in the comprehensive approach to mental state detection in pilots, which includes preprocessing and extracting PSD features from EEG data, developing a one-dimensional CNN (1D-CNN) model with five convolutional layers, predicting four mental states, and interpreting the results using SHAP to identify important features for each class. This approach not only leverages the power of CNN models for EEG analysis but also addresses the critical need for model interpretability in this domain.

The decision to utilize a 1D-CNN approach, in preference to other time-series models such as LSTM or transformers, is rooted in several key considerations specific to EEG signal processing. EEG signals, while temporal in nature, exhibit distinct spatial patterns across electrodes that are critical for identifying different mental states. These spatial patterns are akin to the spatial features in images, making 1D-CNNs, which excel in spatial feature extraction, a suitable choice for EEG analysis. The convolutional layers in 1D-CNNs can effectively capture these spatial relationships within EEG signals, even though they are

manifested over time. Moreover, 1D-CNNs provide an efficient way to handle the high-dimensional nature of EEG data. They can process multiple channels of EEG simultaneously, effectively learning spatial features across these channels, which is essential for accurate mental state classification. This is particularly advantageous when dealing with multi-channel EEG recordings, where each channel provides valuable spatial information about brain activity.

Another crucial aspect is the computational efficiency of 1D-CNNs. Compared to LSTM and transformer models, which are inherently more complex and computationally demanding due to their architecture designed to capture long-term dependencies, 1D-CNNs are generally more lightweight and faster to train. This makes them well-suited for applications where rapid processing of EEG data is required, such as real-time monitoring in aviation contexts. Additionally, the interpretability of the model was a critical factor in our choice. While LSTM and transformer models offer advanced capabilities in capturing temporal dynamics, their complex internal structures can make the interpretation of their decision-making processes more challenging. On the other hand, 1D-CNNs, particularly when combined with SHAP values, provide a clearer avenue for understanding which features of the EEG data are most influential in determining mental states. This level of interpretability is crucial in high-stakes applications like aviation, where understanding the rationale behind a model's prediction is as important as the accuracy of the prediction itself.

The EEG data was sourced from the Attention-related Human Performance Limiting States (AHPLS) dataset, a rich and diverse dataset that has been used in several recent studies to understand and model human cognitive states (I. Alreshidi et al., 2023; Ibrahim Alreshidi et al., 2023; Alreshidi et al., 2022; Harrivel et al., 2016; Harrivel et al., 2017; Terwilliger, 2020). The AHPLS dataset is unique in its inclusion of data from pilots under various mental states, namely channelised attention (CA), diverted attention (DA), startle/surprise (SS), and normal/ no event (NE) states. This provides a robust and realistic dataset for training and testing our model. The use of such a dataset is a significant contribution to the field, as it allows for the exploration of EEG-based

mental state detection in a high stakes real-world environment, such as aviation (Prinzel et al., 2000).

This study aims to answer the following research questions:

1. How effectively can a 1D-CNN model, trained on PSD features of EEG signals, detect four mental states in pilots?
2. What are the key features of PSD that contribute to the successful detection of these mental states?
3. How does the performance of the model vary across different pilots and when trained on data from all pilots combined?

The answers to these questions will provide valuable insights into the potential of EEG-based mental state detection in aviation and other high-stakes environments and will guide future research in this area.

The paper is organised as follows: In the related work section, an overview of mental states and related research studies is presented. The proposed comprehensive approach section describes the methods used in this study. The results section presents the findings, and the discussion section presents a discussion of the results. Finally, the conclusion of the study is stated in the conclusion section.

6.3 Related Work

The application of ML techniques, particularly deep learning (DL), to the analysis of EEG data for the detection of mental states has been a topic of significant interest in recent years. This section provides an overview of the key research in this area, highlighting the methodologies used, the results obtained, and the gaps that this study aims to address.

6.3.1 Previous Studies on EEG-Based Mental State Detection

Several studies have explored the use of EEG data for the detection of mental state. For instance, Başar et al. (Basar et al., 2001) found that γ, α, δ, β, and θ oscillations govern cognitive processes, suggesting that these frequency bands

could be used to detect different mental states. Similarly, Klimesch (Klimesch, 1999) found that EEG α and θ oscillations reflect cognitive and memory performance, indicating their potential for mental state detection. Regarding the application of ML and DL methods, Giudice et al. (Giudice et al., 2020) developed a 1D-CNN model to detect and discriminate between voluntary and involuntary blinking of the eye using EEG data, demonstrating the potential of DL techniques for such tasks. Mattioli et al. (Mattioli et al., 2022) proposed an approach based on a 10-layers 1D-CNN to classify four motor imagery (MI) and baseline states, which showed promising results in terms of performance. Similarly, Tabar et al. (Tabar & Halici, 2017) model based on CNN and stacked autoencoders to classify EEG MI signals, demonstrating the effectiveness of DL models compared to ML models. Furthermore, Zorzos et al. (Zorzos et al., 2023) extracted time-frequency domain characteristics from the EEG signal to train on a shallow CNN model, with three convolutional layers, for the detection of mental fatigue, which showed promising results in terms of performance and interpretability.

In the context of aviation, Wu et al. (Wu et al., 2021; Wu et al., 2019) addressed the problem of obtaining the representation of the fatigue status feature and detecting the fatigue behaviour status of pilots through EEG signals. The authors decomposed the EEG signals of pilots into four frequency bands, namely δ, θ, α, β, and used them to train a deep contractive autoencoder network, achieving 91.67 % performance accuracy. Furthermore, Cui et al. (Cui et al., 2022) developed a CNN model to detect driver drowsiness using EEG data, achieving an average accuracy of 78.35% in 11 participants. They also employed an interpretation tool to recognize the biological features of drowsiness states. Han et al. (Han et al., 2020) proposed a multimodal approach to classify four mental states, namely distraction, workload, fatigue and normal, using EEG, electrocardiogram (ECG), respiration (Resp.) and electrodermal activity (EDA). They extracted PSD features from the EEG signals which were used to train a CNN model and trained a long-short temporal memory (LSTM) model on the other non-brain signals, achieving an average accuracy of 85.2%. Johnson et al. (Johnson et al., 2015) used the

average power of the frequency bands as features to detect task complexity levels in flight simulator experiment. In addition, Roza et al. (Cesar Cavalcanti Roza & Adrian Postolache, 2019; Roza et al., 2019) focused on detecting pilot's emotions, namely happy, sad, angry, scared, surprised, and disgust, using artificial neural network (ANN) in simulated flights. Binias et al. (Binias et al., 2018) proposed an ML approach to discriminate between states of brain activity related to idle but focused anticipation of visual signals and reaction to them.

6.3.2 Gaps in the Existing Literature

While these studies have made significant contributions to the field, this study aims to address. Firstly, many previous studies have focused on binary or multiclass classification problems, such as distinguishing between rest and task states or between different levels of cognitive load. However, less research has been conducted on the detection of specific mental states, such as CA, DA, SS, and NE states, particularly in the context of aviation.

Second, while DL models have shown promise in EEG analysis, understanding the decision-making process of these models remains a challenge. Many previous studies have focused on improving the performance of these models, but less attention has been paid to their interpretability. This is a significant gap, as understanding the features that these models consider important can provide valuable insights into the underlying cognitive processes associated with different mental states.

6.3.3 Previous Research on Detecting CA, DA, SS, and NE States

Several studies have explored the detection of specific mental states using EEG data. For instance, Harrivel et al. (Harrivel et al., 2016; Harrivel et al., 2017) recorded brain signals (i.e., EEG) and non-brain signals (i.e., ECG, R, and galvanic skin response (GSR)), capturing the attention related pilot performance limiting states, including CA, DA, SS, and NE. The authors employed various ML techniques to perform binary and multiclass classification tasks in two different studies. Terwilliger et al. (Terwilliger, 2020) attempted to discriminate between the normal state and an event state, where the CA, DA and SS states

are combined and named an event state. In previous research (Alreshidi et al., 2022), we performed a multiclass classification task to attempt to predict the CA, DA, SS, and NE states. The main purpose of the study was to measure the impact of applying different preprocessing techniques on the performance of the ML model and to investigate the feasibility of concatenating EEG datasets recorded from different environments settings. It was found that the employing of pre-processing techniques has an impact on classification performance. Thus, an automated preprocessing approach was proposed to improve the signal-to-noise ratio and classification performance (I. Alreshidi et al., 2023). The proposed approach demonstrated the importance of preprocessing EEG data before training them in ML models. The attention-related pilot performance limiting states is heavily class imbalanced. In a recent study (Ibrahim Alreshidi et al., 2023), we evaluated the impact of employing various data resampling techniques on classification performance. It was discovered that the use of a combination of downsampling and oversampling techniques improves the performance of the ML models.

However, less research has been conducted on the detection of the specific mental states of CA, DA, SS, and NE states. These states are particularly relevant in the context of aviation, where pilots need to rapidly switch between different mental states in response to varying task demands. Understanding and detecting these states could have significant implications for safety and performance in aviation and other high-stakes environments.

6.3.4 Positioning of the Current Work

This study builds on the existing literature in several ways. Firstly, it focusses on the detection of specific mental states that are relevant in the context of aviation, addressing a gap in the existing literature. Second, it uses a 1D-CNN model, which has been shown to be effective in handling high-dimensional data such as EEG signals. This model is trained on PSD features of EEG data, which represent the distribution of signal power over various frequency bands and are associated with different cognitive processes and mental states.

Furthermore, this study addresses the need for model interpretability in EEG analysis by employing SHAP values. SHAP values provide a measure of the contribution of each feature to the model's prediction, offering insights into the model's decision-making process. This approach not only leverages the power of CNN models for EEG analysis but also addresses the critical need for model interpretability in this domain.

The EEG data used in this study are sourced from the AHPLS dataset, a rich and diverse dataset that has been used in several recent studies to understand and model human cognitive states. The AHPLS dataset is unique in its inclusion of data from pilots under various mental states, providing a robust and realistic dataset for training and testing our model. The use of such a dataset is a significant contribution to the field, as it allows for the exploration of EEG-based mental state detection in a real-world, high-stakes environment such as aviation.

In summary, this study extends the existing literature by focussing on the detection of specific mental states in pilots using a 1D-CNN model trained on PSD features of EEG data. It also addresses the need for model interpretability by employing SHAP values to identify the important features of each mental state. The use of the AHPLS dataset further enhances the relevance and applicability of this research in the field of aviation.

6.4 The Proposed Approach

In this section, we describe the methods utilised to preprocess the EEG data, extract meaningful features from the EEG data, and handle the data imbalance issue. In addition, we explain the proposed 1D-CNN model and the interpretability method (i.e., SHAP) used to identify the most important features for each mental state. Figure 6-1 illustrates an overview of the proposed approach.

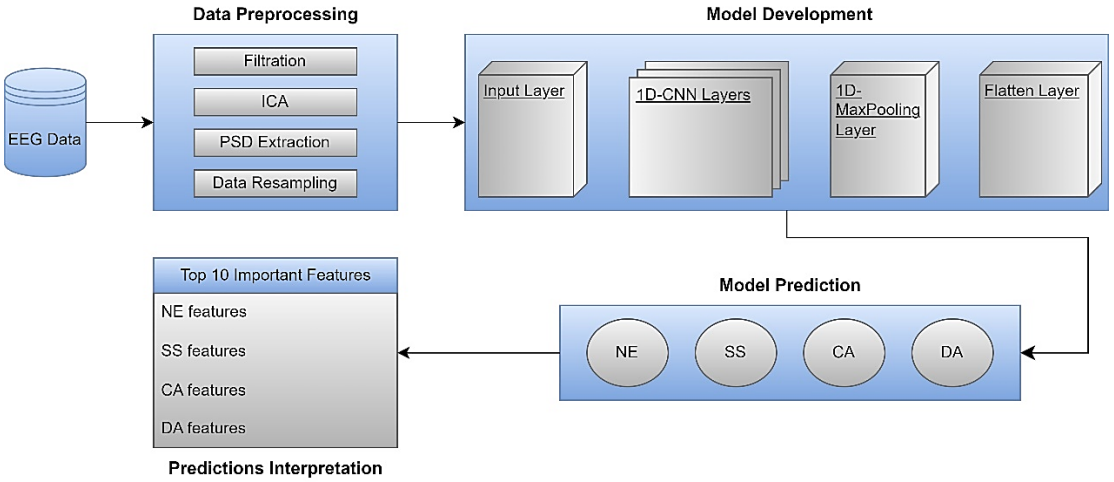

Figure 6-1 An overview of the proposed approach

6.4.1 Data Preprocessing

EEG data was initially segmented into 1 second and filtered using a finite impulse response (FIR) filter with a frequency range of 1 to 40 Hz (Widmann et al., 2015). This step was instrumental in attenuating extraneous noise and enhancing the signal-to-noise ratio of the EEG data, thereby improving the quality of the data for subsequent analysis. Subsequent to the filtering process, the data were subjected to an artefact removal procedure to address ocular-related artefacts, a common occurrence in EEG data. This was achieved utilizing the Independent Component Analysis (ICA) algorithm, as delineated by Aapo Hyvärinen in his seminal work (Hyvarinen, 1999). The ICA algorithm, renowned for its robustness in the separation of independent sources, was employed to isolate and subsequently remove components of the EEG data that were indicative of ocular movements.

Upon successful removal of artefacts, spectral analysis was performed on the sensor data. This was facilitated by the “multitaper” method, a technique that employs discrete prolate spheroidal sequences (DPSS) tapers (SLEPIAN, 1978). This method was selected due to its ability to provide robust spectral estimates with minimised variance. The lower and upper-bound frequencies of interest were set to 1 and 40 Hz, respectively. This frequency range was strategically chosen to focus on the frequency bands that were pertinent to the

study, while concurrently excluding frequencies that could potentially introduce noise into the analysis. The rigorous preprocessing steps outlined above ensured the optimal preparation of the EEG data for the ensuing stages of the study.

Following the preprocessing of the EEG data, the dataset was partitioned into training and testing datasets, with proportions of 80% and 20% respectively. Then, we split the partitioned training dataset into training and validation datasets, with proportions of 70% and 30% respectively. This division was carried out to facilitate the model's learning process and to ensure a robust evaluation of its performance. To address the issue of data imbalance, the SMOTEENN method was used. This hybrid resampling technique, which combines the synthetic minority oversampling technique (SMOTE) (Fernández et al., 2018) and the Edited Nearest Neighbours (ENN), is highly effective in handling imbalanced data.

The SMOTEENN method first applies SMOTE to generate synthetic samples from the minority class, thereby balancing the class distribution. Mathematically, for each minority class sample x, it chooses one of its k nearest neighbors x' and generates a new sample at a random point between x and x', i.e., $x_{new} = x + \lambda . (x' - x)$, where λ is a random number between 0 and 1. Subsequently, the ENN method is applied to remove any instances of the majority class that are surrounded by minority class instances and any instances of the minority class that are misclassified by its three nearest neighbours. This cleaning process ensures that the oversampling does not overgeneralise the minority class by creating noisy samples. The application of SMOTEENN in this study ensured a balanced representation of classes, thereby improving the model's ability to generalise from the training data to unseen data.

6.4.2 The One-Dimensional Convolutional Neural Network (1D_CNN)

The 1D Convolutional Neural Network (1D-CNN) model is a variant of the traditional Convolutional Neural Network that is specifically designed for sequence data (Abdoli et al., 2019; Kiranyaz et al., 2019). The model is

composed of five 1D-CNN layers, followed by a MaxPooling1D layer and a flattened layer. The mathematical operation performed by a 1D-CNN layer can be described as follows.

Given an input sequence $x = [x_1, x_2, \dots, x_n]$, a filter $w = [w_1, w_2, \dots, w_k]$ of length k is applied to the sequence to produce a new sequence $y = [y_1, y_2, \dots, y_{n-k+1}]$, where each element y_i is computed as:

$$y_i = b + \sum_{j=1}^{k} w_j \, . \, x_{i+j-1} \quad \textbf{(6-1)}$$

Here, b is a bias term. This operation is applied for each filter in the layer, and the results are typically passed through a nonlinear activation function, such as the Rectified Linear Unit (ReLU) function.

After passing through the five 1D-CNN layers, a MaxPooling1D layer is applied. This layer reduces the dimensionality of its input by applying a max operation over sliding windows of a specified size. If the window size is p, then the output $z = [z_1, z_2, \dots, z_{n/p}]$ is computed as:

$$z_i = \max_{j=(i-1)\,p+1} y_i \quad \textbf{(6-2)}$$

This operation helps to make the model more robust to shifts and distortions in the input data and reduces the computational complexity of subsequent layers. After passing through the five 1D-CNN and MaxPooling1D layers, the output is flattened into a one-dimensional vector. This flattened output can then be passed through one or more fully connected layers, which perform the final classification or regression task. The 1D-CNN model's strength lies in its ability to effectively capture local dependencies in the input data, making it particularly well suited for tasks involving time series or sequence data. Its architecture allows it to learn both short- and long-term patterns in the data, which can be crucial for many prediction tasks.

6.4.3 SHapley Additive exPlanations (SHAP)

SHAP is a unified measure of feature importance that assigns each feature an importance value for a particular prediction. The concept of SHAP is based on Shapley values, a concept from cooperative game theory that assigns payouts to players depending on their contribution to the total payout (Shapley, 1953). SHAP values interpret the output of the ML and DL models using a game-theoretic approach, attributing the prediction of each instance to its features (Lundberg & Lee, 2017).

The SHAP value for a feature is the average marginal contribution of that feature across all possible combination of features. Mathematically, the SHAP value φi for a feature i is given by:

$$\varphi i = \sum_{S \subseteq M \setminus (i)} \frac{(|S|\,!\,(|M| - |S| - 1))!}{|M|!} [f\,(S \cup (i)) - f(S)] \quad \textbf{(6-3)}$$

where M is the set of all features, S is a subset of M without feature (i), $|S|$ is the number of features in S, $|M|$ is the total number of features, and f is the prediction function. The term $(|S|!\,(|M| - |S| - 1)!/|M|!)$ is the weight representing the number of times a subset S of size $|S|$ appears in all possible subsets of M.

In practice, SHAP values for a model are computed for each feature across all instances in the dataset. This process involves aggregating the contributions of each feature across all the possible coalitions of features for every data instance, providing a comprehensive view of feature importance across the entire dataset. In this study, we have employed SHAP values to interpret the predictions of our CNN model trained on PSD features of EEG data. By calculating SHAP values for each feature across all instances, we are able to identify the most influential features for each mental state prediction. This approach not only allows us to determine the top 10 most influential features for each mental state but also offers insight into the neural correlates of these states. Utilizing SHAP values in this manner enhances the model's

interpretability, contributing to a deeper understanding of the neural mechanisms underlying the mental states of interest in aviation.

6.5 Experimental setup

This section elaborates on the experimental setup employed in this study, encompassing the dataset, Python libraries, PC specifications, and hyperparameter tuning. Each component plays a crucial role in the overall research design and contributes to the validity and reliability of the results.

6.5.1 Dataset

In the present study, we employed a publicly released EEG dataset, extracted from the AHPLS dataset. This dataset encompasses psychophysiological data derived from 20 EEG channels, collected from 17 pilots operating within a flight simulation environment. The data were annotated with labels corresponding to different mental states, namely the CA, DA, SS, and NE states.

The EEG channels are denoted as follows: FP1, F7, F8, T4, T6, T5, T3, FP2, O1, P3, Pz, F3, Fz, F4, C4, P4, POz, C3, Cz, and O2. Each channel was sampled at a frequency of 256 Hz, ensuring a high-resolution temporal dataset.

A noteworthy characteristic of this dataset is its class imbalance. The NE class constitutes the majority of the dataset, accounting for 83% of the total instances. This is followed by the CA class, which comprises 14% of the dataset. The DA and SS classes are significantly underrepresented, making up 2% and 1% of the dataset, respectively. This class imbalance poses a challenge for conventional ML models, necessitating the use of specialised techniques to ensure robust and generalisable performance.

6.5.2 Python Libraries and PC Specifications

The computational experiments were conducted on a PC equipped with an Intel(R) Core(TM) i7-10700 CPU @ 2.90GHz. The PC boasts a RAM capacity of 32.0 GB, ensuring efficient handling of large datasets and complex computations.

Python, renowned for its simplicity and powerful libraries, was used for all computational tasks. We utilised several Python libraries, each serving a distinct purpose. NumPy and Pandas were used for efficient data handling and manipulation, providing robust structures for dataset operations. MNE-Python, version 1.2, a library dedicated to processing electrophysiological signals, was used to handle the specific data types present in the AHPLS dataset. Scikit-Learn was used for various machine learning tasks, including data preprocessing and model evaluation. TensorFlow, version 2.4, a powerful library for creating and training DL models, was used to construct and train our neural network models. Lastly, the SHAP library was used to interpret the predictions of the proposed model.

6.5.3 Hyperparameter Tuning

In the process of model development, hyperparameter tuning was performed to optimise the performance of the 1D-CNN model. The hyperparameters were fine-tuned based on the specific requirements of the task and the characteristics of the dataset. Table 6-1 summarises the hyperparameters that were fine-tuned for the 1D-CNN model:

Table 6-1 Hyperparameters of the layers of the 1D-CNN model

Layer	Filter	Kernel Size	Activation Function	Padding	Kernel Initializer
1	64	3	Relu	Same	GlorotNormal
2	128	3	Relu	Same	GlorotNormal
3	256	3	Relu	Same	GlorotNormal
4	128	3	Relu	Same	GlorotNormal
5	64	3	Relu	Same	GlorotNormal

The architecture of the model included five 1D-CNN layers, followed by a MaxPooling1D layer and a flattened layer. The convolutional layers were designed with a kernel size of 3 and a stride of 1. For padding, the 'same' mode was used, which means that the output of each convolutional layer has the same length as the input by padding the input with zeros if necessary. This

approach ensures that edge information is not lost and allows for consistent feature detection across the entire length of the input sequence.

The MaxPooling1D layer had a pool size of 2. The cumulative structure of these layers resulted in a receptive field of 12 data points in the final convolutional layer. Given the EEG data's sampling rate of 256 Hz, this receptive field corresponds to approximately $\frac{12}{256}$ seconds (around 47 milliseconds). This duration was deemed suitable for capturing the relevant EEG features, particularly those associated with frequency-specific activity within short time windows. Regarding the training of the model, an optimizer plays a critical role in the learning process. For our 1D-CNN model, we employed the Adaptive Moment Estimation (Adam) optimizer. Adam is widely used due to its efficiency in handling sparse gradients and its adaptive learning rate capabilities. This optimizer adjusts the learning rate throughout training, which helps in navigating the complex optimization landscapes typically encountered in neural network training.

The output layer of the model was a dense layer with 4 neurons and a softmax activation function, which is suitable for multiclass classification tasks. In addition to the layer-specific parameters, several global parameters were also set for the training process. The learning rate was set to 1e-4, which determines the step size at each iteration while moving toward a minimum of a loss function. The model was trained for 100 to 150 epochs, where an epoch is an iteration over the entire dataset. The batch size was set to 32, which is the number of samples processed before the model is updated.

The GlorotNormal initializer was used with different seed numbers for initialising the kernel's weights. This initializer draws samples from a truncated normal distribution centred on 0, with $\boldsymbol{stddev} = \boldsymbol{sqrt}(\mathbf{2} / (\boldsymbol{fan\_in} + \boldsymbol{fan\_out}))$, where $\boldsymbol{fan\_in}$ is the number of input units in the weight tensor and $\boldsymbol{fan\_out}$ is the number of output units. This initializer is also known as the Xavier normal initializer.

The selection of these hyperparameters, alongside the receptive field design and the use of the Adam optimizer, was crucial in optimizing the model's performance on the validation set. This approach ensured that the model could effectively learn from the training data and generalize, as evidenced by its subsequent evaluation on the test set.

6.6 Results

The results section of this study is organised into several subsections, each addressing a distinct aspect of the analysis. The first subsection investigates the impact of different mental states on the power distribution across EEG frequency bands. This analysis further elucidates the neural mechanisms underlying the mental states of interest and their manifestation in EEG data. The second subsection presents the performance metrics of the 1D-CNN model trained on the PSD features of EEG data from 17 pilots along with the training accuracy and loss function curves, and the confusion matrix. These metrics include accuracy, precision, recall, and F1-score, providing a comprehensive evaluation of the model's ability to detect four mental states: CA, DA, SS, and NE states. Lastly, the model interpretation subsection delves into the feature importance analysis, using SHAP values to identify the top 10 most important features for each mental state. This analysis provides insights into the EEG frequency bands and channels that are most influential in the model's decision-making process, offering a deeper understanding of the neural correlates of the mental states under study.

Together, these subsections provide a comprehensive evaluation of the model's performance, an exploration of the key features driving its predictions, and an examination of the neural underpinnings of the mental states it is designed to detect. The results presented in this section not only demonstrate the effectiveness of the proposed approach, but also contribute to our understanding of the neural correlates of mental states in the context of aviation.

6.6.1 Examining the Effects of Mental States on EEG Frequency Bands

The PSD of the EEG signals was analysed across different mental states and frequency bands. The average power in each frequency band (delta δ, theta θ, alpha α, beta β, and gamma γ) was calculated for each mental state (NE, SS, CA, and DA) and visualized using a bar plot shown in Figure 6-2 and a heatmap as depicted in Figure 6-3.

In Figure 6-2, the bar plot shows the average power in each frequency band for each mental state. The height of each bar represents the average power in that band for that state. It can be observed that there are distinct differences in the power across different frequency bands for each mental state. This suggests that the power in different frequency bands may be a useful feature for distinguishing between different mental states. However, there is also considerable variability within each band and state. This suggests that there may be individual differences or other factors that are not captured by the average power alone.

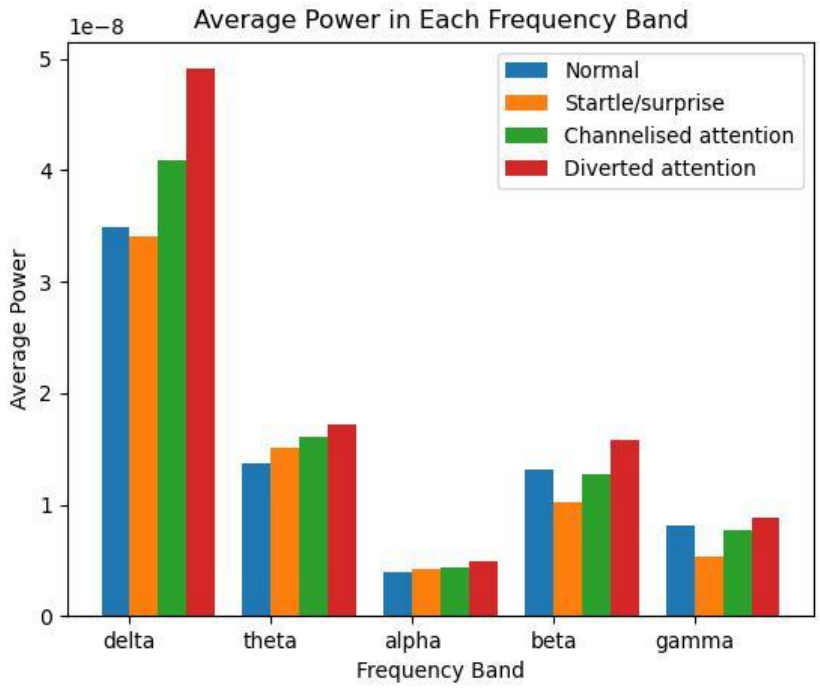

Figure 6-2 The average power in each frequency band across pilots

In Figure 6-3, the heatmap shows the average power in each frequency band for each EEG channel. The colour of each cell represents the average power in that band for that channel. It can be seen that there are distinct patterns of power across different channels and frequency bands. This suggests that the spatial distribution of power in different frequency bands may also be a useful feature for distinguishing between different mental states. However, it is also apparent that there is considerable variability across different channels, suggesting that the power in different frequency bands may be influenced by the location of the electrodes and the underlying brain regions.

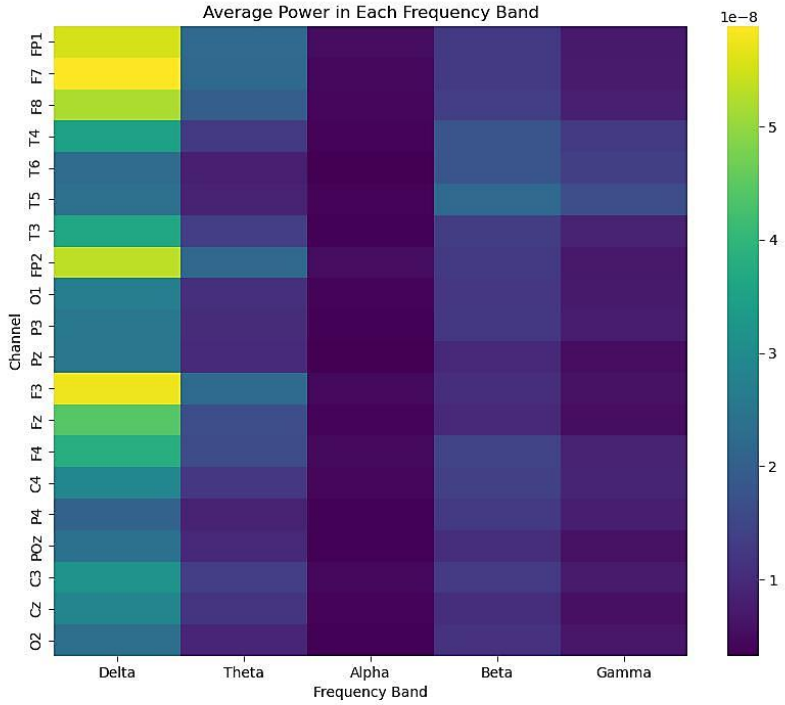

Figure 6-3 Heatmap for the average power in each frequency band for EEG channels

Together, these results suggest that the power in different frequency bands and the spatial distribution of the power across different channels may be useful features to distinguish between different mental states. However, further analysis is needed to determine the statistical significance of these differences and investigate the potential influence of other factors such as individual differences and electrode placement. Future research could also investigate the

temporal dynamics of power in different frequency bands, as the current analysis only considers the average power over the entire recording period.

6.6.2 Classification Results

The study utilised the proposed model to identify four distinct mental states of the pilots. CA, DA, SS, and NE states. The model was trained on PSD features derived from 5 frequency bands across 20 EEG channels. This approach allowed for a comprehensive representation of the EEG data, capturing the complex interplay of different frequency bands across multiple channels.

To comprehensively evaluate the model's performance and its adaptability to individual variations, a two-phase training approach was employed. Initially, the model was trained individually on each of the 17 pilots, resulting in the creation of 17 distinct models. This individualized training aimed to capture any unique EEG patterns and responses to different mental states specific to each pilot. Subsequently, to assess the model's generalizability and performance across a more diverse dataset, a single combined model was trained on the aggregated data encompassing all 17 pilots.

The performance of each of these models, both individual and combined, was evaluated using four key metrics: Accuracy, Precision, Recall, and F1-Score. These metrics provide a holistic view of the model's performance, capturing its ability to make correct predictions (Accuracy), its precision in correctly identifying positive instances (Precision), its effectiveness in identifying all positive instances (Recall), and the balance between its Precision and Recall (F1-Score). The use of these metrics across both individual and combined models offers insights into the model’s ability to adapt to individual differences as well as its robustness in a broader, more varied context.

As presented in Table 6-2, the results showed a high degree of consistency across all metrics for each pilot. The accuracy, precision, recall, and F1-scores all fell within a relatively narrow range of 94% to 99%. The highest accuracy and precision of 99% were achieved by Pilot 2. The lowest scores across all metrics

were observed for Pilot 12, with an accuracy of 94%, precision of 91%, and an F1-score of 90%.

Table 6-2 Classification results of individual and combined pilots

Pilot ID	Accuracy	Precision	Recall	F1-score
1	97.21	97.15	94.76	96.09
2	98.68	98.60	97.59	97.55
3	97.31	97.94	95.42	96.30
4	96.22	93.29	93.30	94.11
5	96.52	97.61	96.06	95.99
6	97.13	95.90	94.42	95.26
7	97.27	97.68	95.76	96.10
8	96.39	95.02	92.20	93.18
9	96.45	96.11	96.80	96.94
10	95.34	95.59	89.88	92.04
11	94.55	94.52	88.82	91.21
12	94.41	90.94	89.33	90.30
13	95.84	95.18	92.95	93.77
14	96.48	96.52	92.03	93.27
15	97.12	94.85	96.46	95.14
16	98.31	97.81	95.19	95.93
17	97.87	96.59	96.10	96.99
All	96.28	95.66	94.22	95.11

When the model was trained on the combined PSD features of all pilots, it achieved an accuracy, precision of 96%, recall of 94%, and F1-score of 95%. This suggests that the model was able to generalise well from individual pilots to a larger population.

In addition to the proposed model performance metrics, the training process was also evaluated by examining the accuracy and loss curves for the training and validation datasets as depicted in Figure 6-4. For the training dataset, the accuracy curve demonstrated a consistent upward trend, indicating a steady improvement in the model's ability to correctly predict the mental states as the training progressed. This consistent improvement suggests that the model was effectively learning the patterns in the training data and adapting its parameters accordingly. Simultaneously, the loss curve for the training dataset showed a consistent downward trend, indicating that the model was successfully reducing the error in its predictions over time. This is a positive sign of the model's learning capability as it shows that the model was able to progressively minimise the discrepancy between its predictions and the actual values.

In contrast, the accuracy and loss curves for the validation dataset showed slight irregularities. Despite these irregularities, the overall trend of the validation accuracy curve was positive, and the validation loss curve generally showed a decreasing trend. This indicates that, despite the fluctuations, the model was able to apply what it learnt from the training data to unseen data, demonstrating a good level of generalisation.

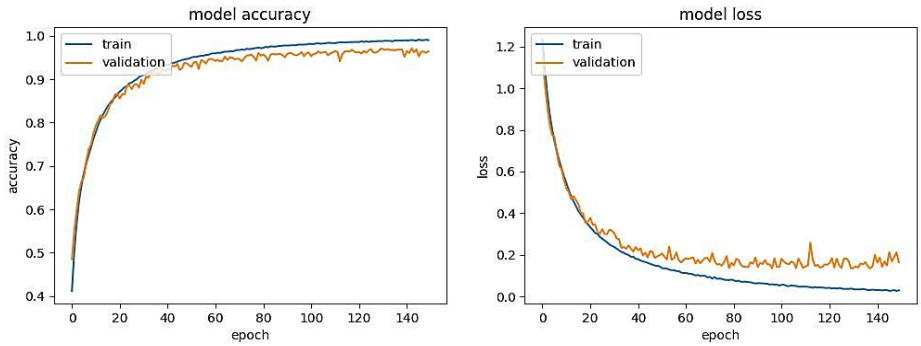

Figure 6-4 Training accuracy and loss curves of the proposed model

The performance of the proposed model was evaluated in a more detailed manner using a confusion matrix. This matrix provides a comprehensive view of the model's ability to correctly classify each of the four mental states: NE, SS, CA, and DA. The matrix is structured such that each row represents the instances in an actual class while each column represents the instances in a predicted class. The confusion matrix is depicted below in Figure 6-5.

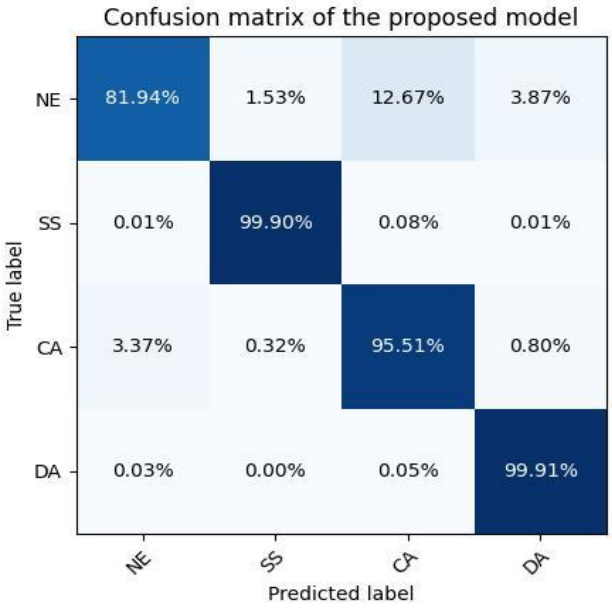

Figure 6-5 Confusion matrix of the proposed approach

The diagonal elements of the confusion matrix, which represent the percentage of correct predictions for each mental state, show that the model achieved high accuracy rates for each of the four mental states, with the lowest being 81.94% for the NE and the highest being 99.91% for the DA state.

6.6.3 Model Interpretation using SHAP

The study employed SHAP values to identify the top 10 most important features for each mental state class: NE), SS, CA, and DA. The SHAP values provide a measure of the contribution of each feature to the model's prediction for each class, allowing for an understanding of which features are most influential in determining the mental state.

For the NE class, the top 10 features were primarily delta and beta frequency bands from various EEG channels. The mean absolute SHAP values for these characteristics, as shown in Figure 6-6, ranged between 0.15 and less than 0.25. The SS class showed a similar pattern, with the top 10 features being predominantly delta and beta frequency bands. As shown in Figure 6-7, the mean absolute SHAP values for these characteristics ranged between a little less than 0.08 and a little bit less than 0.11. The CA class also showed a predominance of delta and beta frequency bands in the top 10 features. Figure 6-8 illustrates that the mean absolute SHAP values for these features ranged between 0.127 and approximately 0.225. Lastly, for the DA class, the top 10 features were primarily delta and beta frequency bands. Figure 6-9 shows that the mean absolute SHAP values for these features ranged between a little bit less than 0.08 and a little bit more than 0.12.

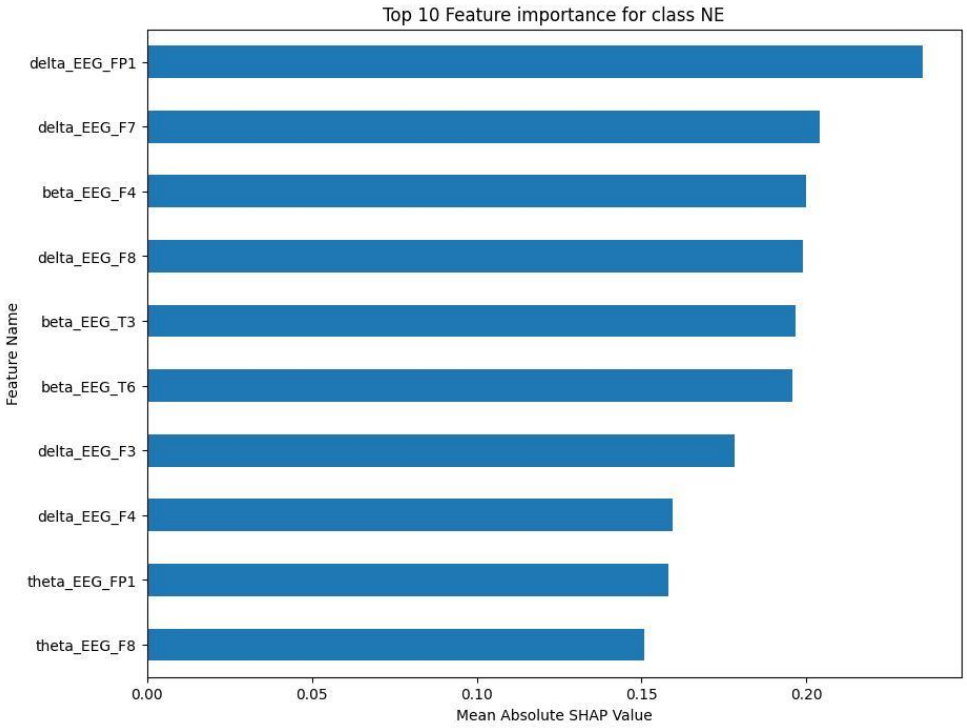

Figure 6-6 Top 10 important features for NE class

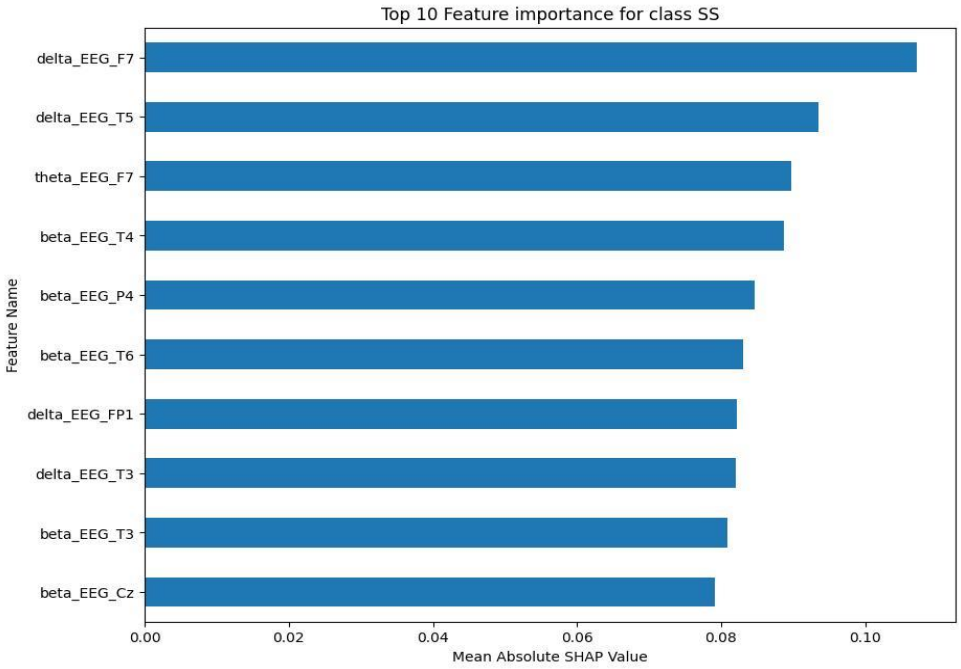

Figure 6-7 Top 10 important features for SS class

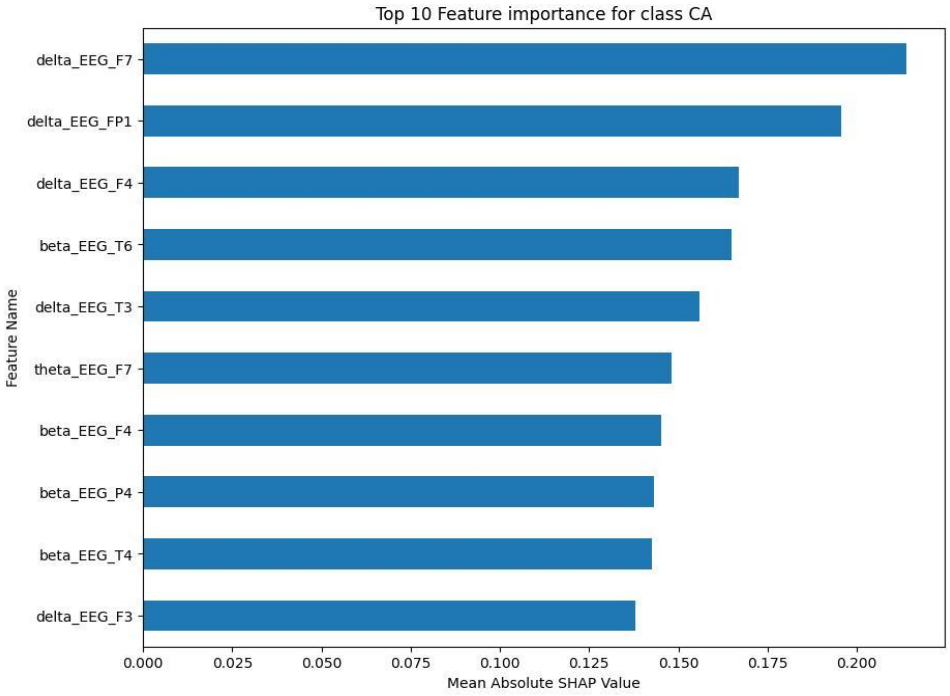

Figure 6-8 Top 10 important features for CA class

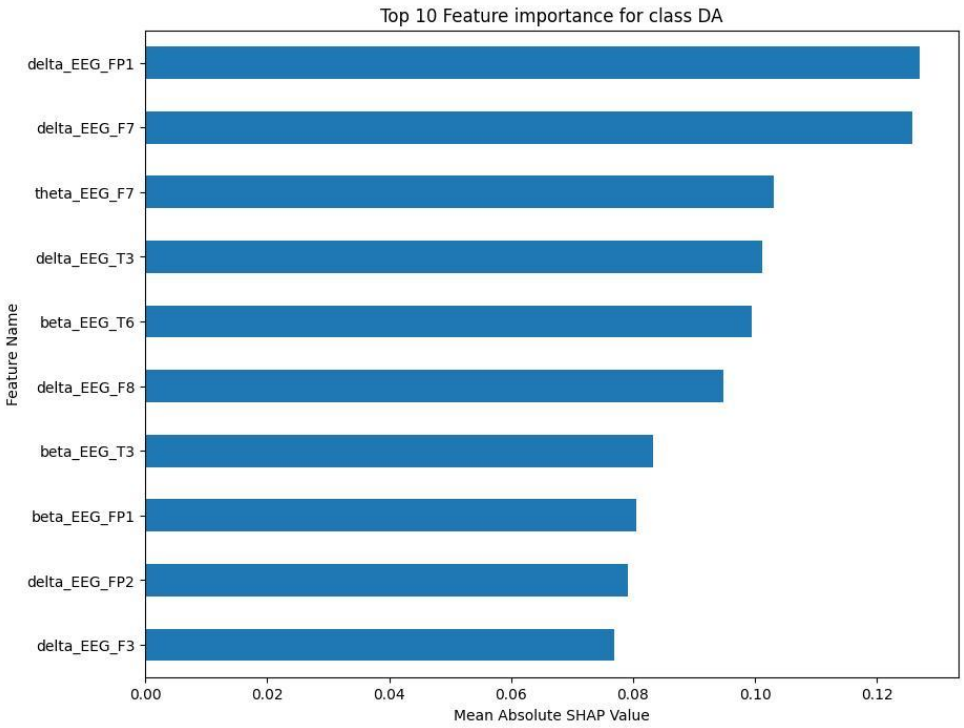

Figure 6-9 Top 10 important features for DA class

6.7 Discussion

The results of this study demonstrate the potential of using the proposed approach to detect mental states based on PSD features of the EEG data. The high-performance metrics across all pilots suggest that the model is effective in distinguishing between the four mental states: CA, DA, SS, and NE.

The use of PSD features from 5 frequency bands (i.e., delta, theta, alpha, beta, and gamma) across 20 EEG channels likely contributed to the model's high performance. These features provide a rich representation of the EEG signals, capturing important frequency-specific information that is relevant for distinguishing between different mental states. PSD features encapsulate the power distribution over various frequency bands, which is a crucial aspect of EEG signals that are often linked to different mental states. However, the variation in the model's performance across different pilots indicates that individual differences may have influenced the results. Each pilot may have unique EEG patterns and responses to different mental states, which could affect the model's performance. For example, the model achieved the highest

performance metrics with Pilot 2, suggesting that the features of this pilot's EEG data were particularly well-suited to the model. On the other hand, the model's performance was lowest with Pilot 12, indicating that there may be unique aspects of this pilot's EEG data that were not as effectively captured by the model.

The fact that the model performed well on the combined data of all pilots is promising. It suggests that the model is capable of generalising across different individuals, which is crucial for its potential application in real-world settings. However, the slightly lower recall score in comparison to the other metrics indicates that there is still room for improvement in the model's ability to correctly identify all instances of the different mental states. Future work could explore ways to further improve the model's performance. This could include refining the model's architecture, experimenting with different methods of preprocessing the EEG data, or incorporating additional features that capture more information about the pilots' mental states. For instance, exploring different types of feature extraction methods or incorporating temporal information could potentially enhance the model's performance. Additionally, additional validation with larger datasets and in real-world settings would be beneficial to confirm these findings and further refine the model.

This study provides valuable insights into the potential of using the proposed model for detecting mental states based on PSD features of EEG data. The high-performance metrics achieved by the model suggest that it could be a valuable tool in fields such as aviation, where monitoring pilots' mental states could contribute to safety and performance. However, it is important to note that while the model's performance is promising, the interpretation and application of these results should be done with caution. The model's performance is based on the specific dataset used in this study, and its performance may vary with different datasets. Therefore, further research and validation are necessary to fully understand the model's capabilities and limitations.

Analysis of the accuracy and loss curves provides valuable insights into the learning process and its ability to generalise from the training data to unseen

data. The steady increase in the training accuracy and the consistent decrease in the training loss demonstrate that the model was effectively learning from the PSD features extracted from the EEG data. This suggests that the model's architecture and the preprocessing steps taken, including the use of PSD features, were well-suited for the task of detecting the four mental states.

The slight irregularities observed in the validation accuracy and loss curves suggest that the model's performance varied when applied to different subsets of data. These fluctuations could be due to a variety of factors, including inherent variability in the EEG data, individual differences between pilots, or the specific division of the data into training and validation sets. Despite these irregularities, the overall positive trend in the accuracy of the validation and the loss of general decrease in the validation loss indicate that the model was able to generalise its learning to new data, which is a crucial aspect of its performance.

However, the presence of these irregularities also suggests potential areas for improvement in the model. Future work could explore different strategies for managing these irregularities, such as adjusting the model's architecture, experimenting with different methods of data preprocessing, or using different strategies for dividing the data into training and validation sets.

The accuracy and loss curves provide additional evidence of the potential of the proposed model to detect mental states based on PSD features extracted from the EEG data. Despite some irregularities in the validation curves, the overall trends suggest that the model is capable of learning effectively from the data and generalising its learning to new data. This holds promise for the model's application in real-world settings, such as aviation, where accurate detection of pilots' mental states could contribute to safety and performance.

The confusion matrix provides a deeper understanding of the model's performance across the four mental states. It is evident that the model performs exceptionally well in classifying the SS and DA states, with almost perfect accuracy rates of 99.90% and 99.91%, respectively. This high level of accuracy suggests that the model is highly effective in distinguishing these states, likely

due to the distinct PSD features associated with these mental states in the EEG data. The CA state also saw a high accuracy rate of 95.51%, indicating that the model is also capable of effectively identifying this state. However, the NE state had a noticeably lower accuracy rate of 81.94%. This could be due to the inherent complexity in distinguishing the NE state from the other mental states, as the NE state might not exhibit as distinct PSD features as the other states.

The off-diagonal elements of the confusion matrix, which represent the instances where the model made incorrect predictions, provide further insights into the model's performance. For instance, the model misclassified the NE state as the CA state in 12.67% of instances. This could suggest that the EEG features of these two states might share some similarities, causing the model to confuse between them.

Despite these challenges, the overall performance of the model, as demonstrated by the confusion matrix, is highly promising. The model's ability to achieve high accuracy rates across the four mental states suggests that it is capable of effectively using PSD features extracted from EEG data to detect different mental states. Future work could focus on improving the model's ability to distinguish the normal state, potentially by incorporating additional features or refining the model's architecture. Furthermore, additional validation with larger datasets and in real-world settings would be beneficial to confirm these findings and to further refine the model.

The results of the SHAP analysis, as visualised in Figure 6-6, Figure 6-7, Figure 6-8 and Figure 6-9, provide valuable insight into the most important features to predict each mental state. The predominance of delta and beta frequency bands in the top 10 features for each class suggests that these frequency bands may be particularly important for distinguishing between different mental states. Delta waves are typically associated with sleep or deep relaxation, while beta waves are associated with active thinking or focus. This aligns with the nature of the mental states being predicted, as channelised and diverted attention would likely involve more active thinking (beta waves), while a normal state might be more relaxed (delta waves).

However, it is interesting to note that the specific EEG channels that were most important varied between the classes. This suggests that different mental states may be associated with activity in different regions of the brain, which is captured by the different EEG channels. For example, the F7 channel (frontal lobe) was important for the NE, SS, and CA classes, while the T3 channel (temporal lobe) was important for the SS and DA classes. This could potentially provide insights into the neural mechanisms underlying these mental states.

The range of the mean absolute SHAP values for the top 10 features in each class also provides information about the relative importance of these features. The NE and CA classes had higher SHAP values compared to the SS and DA classes, suggesting that the top features for NE and CA may have a stronger influence on the model's predictions.

Building upon the results and insights gained from the SHAP analysis, an important consideration for future enhancements to our model involves exploring the integration of residual connections or residual block structures. Commonly used in deep CNN architectures for their ability to facilitate feature learning, residual connections could potentially improve our model’s capacity to capture complex patterns within EEG data. However, this approach is not without challenges, especially when considering the interpretability aspect using SHAP values.

Integrating residual connections increases the model's complexity and might obscure the direct relationship between input features and the output, complicating the attribution of SHAP values and potentially reducing the interpretability of the model. Additionally, the increased complexity requires more computational resources for SHAP analysis, impacting the efficiency of the interpretability process. These considerations are crucial in high-stakes fields like aviation, where understanding the rationale behind a model's predictions is as important as the accuracy of the predictions themselves. The decision to forgo residual connections in the current design was made with an emphasis on maintaining a balance between model performance, computational efficiency, and interpretability.

In future iterations of this research, exploring the use of more complex architectures, including those with residual connections, could be beneficial, especially if larger datasets become available, or if there are advancements in interpretability methods for complex models. Adjusting the model's architecture and experimenting with advanced features like residual connections might enable the model to learn more nuanced patterns within the EEG data, possibly enhancing its predictive performance. However, any such modifications would require careful consideration of the trade-offs between model complexity, performance, and interpretability, particularly in the context of EEG-based mental state detection.

6.8 Conclusion

The present study has made significant strides in demonstrating the potential of the proposed approach for the classification of distinct mental states, specifically CA, DA, SS and NE states, based on PSD features derived from EEG data. The model's performance, as assessed by accuracy, precision, recall, and F1-score metrics, was consistently high across all pilots, indicating its robustness and generalisability. This is a promising finding, suggesting that the model can effectively learn from individual pilots and apply this learning to a broader population.

The use of SHAP values in this study has provided a deeper understanding of the model's decision-making process. By identifying the most influential features for each mental state, we have gained insights into the importance of both frequency-specific and spatial information in EEG data for mental state classification. The predominance of delta and beta frequency bands in the top features for each class suggests that these frequency bands play a crucial role in differentiating between various mental states.

However, the study also revealed the complexity of the task at hand. The variation in the model's performance across different pilots and the range of SHAP values across classes underscore the influence of individual differences and the complexity of the mental states being predicted. These findings suggest

that while the model is effective, there is still room for improvement and refinement. Future work could focus on enhancing the model's architecture, exploring different preprocessing methods, or incorporating additional features that capture more nuanced aspects of the pilots' mental states.

Furthermore, the findings need to be validated with larger datasets and in real world settings to confirm their applicability and further refine the model. The high performance metrics achieved by the model suggest its potential utility in fields such as aviation, where monitoring pilots' mental states could contribute to safety and performance. However, the interpretation and application of these results should be done with caution, considering the specific dataset used in this study and the potential variability in the model's performance with different datasets.

In conclusion, this study contributes valuable insights into the potential of CNN models for mental state detection based on PSD features of EEG data. It underscores the importance of comprehensive feature representation, the influence of individual differences, and the need for further research and validation to fully realise the potential of this approach. The findings of this study pave the way for future research in this area, with the ultimate goal of enhancing safety and performance in high-stakes fields such as aviation.

6.9 Appendices

6.9.1 Appendix A: Data and Reproducibility Code

In the interest of promoting transparency and reproducibility, the data utilised in this chapter, along with the associated code for analyses, have been made publicly accessible. The dataset and the code for replicating the analyses can be found under the Digital Object Identifier (DOI):

https://doi.org/10.17862/cranfield.rd.24155832

REFERENCES

Abdoli, S., Cardinal, P., & Koerich, A. L. (2019). End-to-end environmental sound classification using a 1D convolutional neural network. *Expert Systems with Applications*, *136*, 252-263. https://doi.org/10.1016/j.eswa.2019.06.040

Alreshidi, I., Moulitsas, I., & Jenkins, K. W. (2023). Multimodal Approach for Pilot Mental State Detection Based on EEG. *Sensors (Basel)*, *23*(17). https://doi.org/10.3390/s23177350

Alreshidi, I., Yadav, S., Moulitsas, I., & Jenkins, K. (2023). A Comprehensive Analysis of Machine Learning and Deep Learning Models for Identifying Pilots' Mental States from Imbalanced Physiological Data. AIAA AVIATION 2023 Forum,

Alreshidi, I. M., Moulitsas, I., & Jenkins, K. W. (2022). Miscellaneous EEG Preprocessing and Machine Learning for Pilots' Mental States Classification: Implications. 2022 The 6th International Conference on Advances in Artificial Intelligence,

Basar, E., Basar-Eroglu, C., Karakas, S., & Schurmann, M. (2001). Gamma, alpha, delta, and theta oscillations govern cognitive processes. *Int J Psychophysiol*, *39*(2-3), 241-248. https://doi.org/10.1016/s0167-8760(00)00145-8

Bashivan, P., Rish, I., Yeasin, M., & Codella, N. (2016). Learning representations from EEG with deep recurrent-convolutional neural networks. 4th International Conference on Learning Representations (ICLR),

Binias, B., Myszor, D., & Cyran, K. A. (2018). A Machine Learning Approach to the Detection of Pilot's Reaction to Unexpected Events Based on EEG Signals. *Comput Intell Neurosci*, *2018*, 2703513. https://doi.org/10.1155/2018/2703513

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., & Babiloni, F. (2014). Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neurosci Biobehav Rev*, *44*, 58-75. https://doi.org/10.1016/j.neubiorev.2012.10.003

Buzsaki, G., Anastassiou, C. A., & Koch, C. (2012). The origin of extracellular fields and currents--EEG, ECoG, LFP and spikes. *Nat Rev Neurosci*, *13*(6), 407-420. https://doi.org/10.1038/nrn3241

Cesar Cavalcanti Roza, V., & Adrian Postolache, O. (2019). Multimodal Approach for Emotion Recognition Based on Simulated Flight Experiments. *Sensors (Basel)*, *19*(24). https://doi.org/10.3390/s19245516

Chen, J. B., Song, L., Wainwright, M. J., & Jordan, M. I. (2018). Learning to Explain: An Information-Theoretic Perspective on Model Interpretation.

International Conference on Machine Learning, Vol 80, *80*. <Go to ISI>://WOS:000683379200091

Cohen, M. X. (2017). Where Does EEG Come From and What Does It Mean? *Trends Neurosci*, *40*(4), 208-218. https://doi.org/10.1016/j.tins.2017.02.004

Cui, J., Lan, Z., Liu, Y., Li, R., Li, F., Sourina, O., & Muller-Wittig, W. (2022). A compact and interpretable convolutional neural network for cross-subject driver drowsiness detection from single-channel EEG. *Methods*, *202*, 173-184. https://doi.org/10.1016/j.ymeth.2021.04.017

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*.

Giudice, M. L., Varone, G., Ieracitano, C., Mammone, N., Bruna, A. R., Tomaselli, V., & Morabito, F. C. (2020). *1D Convolutional Neural Network approach to classify voluntary eye blinks in EEG signals for BCI applications* 2020 International Joint Conference on Neural Networks (Ijcnn), <Go to ISI>://WOS:000626021404101

Han, S. Y., Kwak, N. S., Oh, T., & Lee, S. W. (2020). Classification of pilots' mental states using a multimodal deep learning network. *Biocybernetics and Biomedical Engineering*, *40*(1), 324-336. https://doi.org/10.1016/j.bbe.2019.12.002

Harmony, T. (2013). The functional significance of delta oscillations in cognitive processing. *Front Integr Neurosci*, *7*, 83. https://doi.org/10.3389/fnint.2013.00083

Harrivel, A. R., Liles, C., Stephens, C. L., Ellis, K. K., Prinzel, L. J., & Pope, A. T. (2016). Psychophysiological Sensing and State Classification for Attention Management in Commercial Aviation. AIAA Infotech @ Aerospace,

Harrivel, A. R., Stephens, C. L., Milletich, R. J., Heinich, C. M., Last, M. C., Napoli, N. J., Abraham, N., Prinzel, L. J., Motter, M. A., & Pope, A. T. (2017). Prediction of Cognitive States during Flight Simulation using Multimodal Psychophysiological Sensing. AIAA Information Systems-AIAA Infotech @ Aerospace,

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans Neural Netw*, *10*(3), 626-634. https://doi.org/10.1109/72.761722

Johnson, M. K., Blanco, J. A., Gentili, R. J., Jaquess, K. J., Oh, H., & Hatfield, B. D. (2015). Probe-Independent EEG Assessment of Mental Workload in Pilots. 7th Annual International IEEE EMBS Conference on Neural Engineering,

Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., & Inman, D. J. (2019). 1D Convolutional Neural Networks and Applications: A Survey. http://arxiv.org/abs/1905.03554

Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Res Brain Res Rev*, *29*(2-3), 169-195. https://doi.org/10.1016/s0165-0173(98)00056-3

Lundberg, S. M., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions* 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA.

Makeig, S., Kothe, C., Mullen, T., Bigdely-Shamlo, N., Zhang, Z. L., & Kreutz-Delgado, K. (2012). Evolving Signal Processing for Brain-Computer Interfaces. *Proceedings of the Ieee*, *100*(Special Centennial Issue), 1567-1584. https://doi.org/10.1109/Jproc.2012.2185009

Mattioli, F., Porcaro, C., & Baldassarre, G. (2022). A 1D CNN for high accuracy classification and transfer learning in motor imagery EEG-based brain-computer interface. *J Neural Eng*, *18*(6). https://doi.org/10.1088/1741-2552/ac4430

Mihajlovic, V., Grundlehner, B., Vullers, R., & Penders, J. (2015). Wearable, wireless EEG solutions in daily life applications: what are we missing? *IEEE J Biomed Health Inform*, *19*(1), 6-21. https://doi.org/10.1109/JBHI.2014.2328317

Oehling, J., & Barry, D. J. (2019). Using machine learning methods in airline flight data monitoring to generate new operational safety knowledge from existing data. *Safety Science*, *114*, 89-104. https://doi.org/10.1016/j.ssci.2018.12.018

Prinzel, L. J., Freeman, F. G., Scerbo, M. W., Mikulka, P. J., & Pope, A. T. (2000). A closed-loop system for examining psychophysiological measures for adaptive task allocation. *Int J Aviat Psychol*, *10*(4), 393-410. https://doi.org/10.1207/S15327108IJAP1004_6

Roza, V. C., Postolache, O., Groza, V., & Pereira, J. M. D. (2019, 26-28 June 2019). Emotions Assessment on Simulated Flights. 2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA),

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

Schomer, D. L., & Lopes da Silva, F. H. (2017). *Niedermeyer's Electroencephalography* (Vol. 1). Oxford University Press. https://doi.org/10.1093/med/9780190228484.001.0001

Shapley, L. S. (1953). 17. A Value for n-Person Games. In *Contributions to the Theory of Games (AM-28), Volume II* (pp. 307-318). Princeton University Press. https://doi.org/10.1515/9781400881970-018

SLEPIAN, D. (1978). Prolate spheroidal wave functions fourier analysis and uncertainty-V: The discrete case. *THE BELL SYSTEM TECHNICAL JOURNAL*.

Tabar, Y. R., & Halici, U. (2017). A novel deep learning approach for classification of EEG motor imagery signals. *J Neural Eng*, *14*(1), 016003. https://doi.org/10.1088/1741-2560/14/1/016003

Teplan, M. (2002). Fundamentals of EEG measurement. *MEASUREMENT SCIENCE REVIEW*, *2*.

Terwilliger, P. S., Jack; Walker, Shannon; Harrivel, Angela. (2020). *A ResNet Autoencoder Approach for Time Series Classification of Cognitive State* MODSIM,

Widmann, A., Schroger, E., & Maess, B. (2015). Digital filter design for electrophysiological data--a practical approach. *J Neurosci Methods*, *250*, 34-46. https://doi.org/10.1016/j.jneumeth.2014.08.002

Wu, E. Q., Deng, P. Y., Qiu, X. Y., Tang, Z. R., Zhang, W. M., Zhu, L. M., Ren, H., Zhou, G. R., & Sheng, R. S. F. (2021). Detecting Fatigue Status of Pilots Based on Deep Learning Network Using EEG Signals. *IEEE Transactions on Cognitive and Developmental Systems*, *13*(3), 575-585. https://doi.org/10.1109/Tcds.2019.2963476

Wu, E. Q., Peng, X. Y., Zhang, C. Z. Z., Lin, J. X., & Sheng, R. S. F. (2019). Pilots' Fatigue Status Recognition Using Deep Contractive Autoencoder Network. *Ieee Transactions on Instrumentation and Measurement*, *68*(10), 3907-3919. https://doi.org/10.1109/Tim.2018.2885608

Zorzos, I., Kakkos, I., Miloulis, S. T., Anastasiou, A., Ventouras, E. M., & Matsopoulos, G. K. (2023). Applying Neural Networks with Time-Frequency Features for the Detection of Mental Fatigue. *Applied Sciences-Basel*, *13*(3). https://doi.org/ARTN 1512

10.3390/app13031512

7 Discussion

The realm of aviation, while being a marvel of human achievement, is not devoid of challenges. Particularly in the cockpit, the mental states of pilots play an instrumental role in ensuring safe and efficient flight operations. With the convergence of sophisticated ML and DL methodologies, there emerges an unprecedented opportunity to enhance our understanding and prediction of pilots' cognitive states, potentially revolutionizing safety protocols in aviation. This discussion chapter encapsulates a series of pioneering studies undertaken to illuminate pilots' intricate neural and cognitive dynamics, leveraging advanced computational techniques and data-driven insights. While distinct in its approach, each chapter collectively contributes to the overarching goal of elevating aviation safety through cognitive state detection and understanding. Below, this thesis provides a discussion of the key findings and the intellectual contributions of each chapter.

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

This systematic literature review offers an exhaustive survey of existing research on the utility of machine learning and psychophysiological data, particularly EEG, in the study of pilot behaviour. The study underscores the predominance of various ML models, with Ensemble models standing out significantly. It also places emphasis on the importance of multiple performance metrics like accuracy, precision, and recall, beyond just accuracy.

A critical examination of the existing literature reveals a pronounced focus on easily quantifiable behavioural aspects, such as 'Workload' and 'Fatigue.' While these aspects are undoubtedly important, this focus has led to a relative neglect of other performance-limiting states like 'CA,' 'DA,' and 'SS.' This lack of representation in the literature indicates a significant research gap and suggests the need for a more comprehensive approach in future studies.

The review identifies several methodological constraints that merit immediate attention. One of the most significant of these is the under-exploration of the effects of preprocessing techniques on model performance. Given the high-dimensional and intricate nature of psychophysiological data, the relevance of sophisticated preprocessing techniques becomes even more pressing. Additionally, the review underscores the limitations associated with the prevalent focus on traditional statistical and frequency-domain features, raising questions about the comprehensive nature of these methods. Furthermore, the review identifies an insufficient focus on the critical issue of data imbalance, revealing a pressing need for future research to examine the impact of various balancing techniques on model performance. The adoption of more advanced models, such as 1D-CNN, and the inclusion of interpretative or explainable models stand out as promising avenues for future research.

The collective findings from this review have practical and academic ramifications. They offer a roadmap for future methodological innovation, inform the design of more effective human-machine interfaces, and provide insights that could be used to develop real-time monitoring systems. These findings thus lay the groundwork for future research initiatives aimed at bolstering both aviation safety and operational efficiency.

7.2 Miscellaneous EEG Preprocessing and Machine Learning for Pilots' Mental States Classification: Implications

The complex interplay between EEG data preprocessing and its subsequent ramifications on ML models tailored for discerning pilots' cognitive states is meticulously examined in this chapter. EEG signals, while encapsulating a rich tapestry of cognitive information, concurrently pose challenges owing to their intrinsic vulnerability to both internal and external noise perturbations. Such noise interferences, if not meticulously addressed, have the potential to undermine the veracity and precision of ensuing analytical endeavours. Many contemporary methodologies either overlook this noise or employ an arbitrary amalgamation of conventional preprocessing techniques to mitigate it. This lack

of standardised preprocessing protocols poses a significant challenge, making it arduous to compare the efficacy of different ML models.

In light of this challenge, the chapter embarks on a mission to assess the impact of various conventional preprocessing techniques on the performance of ML models, namely SVM and ANN. It zooms in on fundamental preprocessing strategies, notably the band-pass filter and ICA. The influence of preprocessing methods on two ML models, namely SVM and ANN, is scrutinized. Utilizing a publicly accessible dataset encompassing EEG recordings from a pilot subjected to diverse cognitive stimuli, the study seeks to establish associations between preprocessing techniques and the performance metrics of the ML models.

This chapter contributes to the understanding of how traditional EEG preprocessing techniques impact the performance of ML models. A pivotal revelation of the investigation is that the chosen traditional preprocessing techniques appear not to significantly influence the performance of the ML models. One of the salient contributions is the empirical evidence suggesting that ML models, when trained with EEG data amalgamated from diverse environments, exhibit commendable predictive prowess. This outcome not only underscores the feasibility of pooling a pilot's EEG data collected across varied settings but also fortifies the premise of employing ML models in real-world, heterogeneous scenarios. Concurrently, despite the successes with traditional preprocessing, the findings also underscore an emergent imperative for the development and integration of advanced EEG preprocessing methodologies to further enhance data integrity and model reliability.

7.3 Multifaceted Approach for Pilot Mental State Detection Based on EEG

This chapter addresses the pivotal role of pilots' cognitive abilities in ensuring flight safety. Over recent years, concerns have surged regarding potential accidents triggered by the deteriorating mental states of pilots. In response, a

unique multifaceted method for discerning mental states in pilots, rooted in EEG signal analysis, was innovated.

A standout feature of this research is its robust preprocessing pipeline tailored for EEG data. This pipeline meticulously filters out artefacts, ensuring the purest form of EEG signals for further analysis. Subsequent to this cleansing process, a sophisticated feature extraction method, underpinned by Riemannian geometry analysis, is employed on the sanitized EEG data. The final layer of this approach harnesses a hybrid ensemble learning technique, which amalgamates the results from multiple ML classifiers. Such a multifaceted strategy has yielded impressive outcomes, boasting an accuracy rate of 86% when applied to the artefact-free EEG data. It's noteworthy to mention that this EEG dataset has its origins in flight and non-flight experiments involving 18 pilots and has been generously made public via NASA's open portal.

Beyond its immediate results, this research carves a niche in the broader academic discourse. It not only furnishes a dependable method to identify pilots' mental states but also underscores the immense potential lying at the intersection of EEG signals and ensemble learning algorithms. This nexus is seen as a beacon of hope in crafting advanced cognitive cockpit systems. The integration of an automated preprocessing mechanism, Riemannian geometry-based feature extraction, and a hybrid ensemble learning model represents an advancement in this field. The study offers a novel approach to addressing the challenges of detecting pilot mental states and introduces methods that could be influential in future cognitive state analysis research.

7.4 A Comprehensive Analysis of Machine Learning and Deep Learning Models for Identifying Pilots' Mental States from Imbalanced Physiological Data

Understanding and monitoring the mental states of pilots is a matter of paramount importance, especially given the critical nature of their roles. In environments where split-second decisions can make the difference between safety and catastrophe, it's essential to have systems in place that can alert or

intervene when a pilot's mental state deviates from the norm. The presented research is a leap in this direction.

A salient feature of this study is its integration of EEG data with non-brain signals like ECG, GSR, and Resp. data. This integration forms the basis of a novel DL architecture that combines the power of one-dimensional CNN and LSTM models. To address the inherent challenge posed by the imbalanced dataset, specific resampling techniques were employed. This involved downsampling the data based on CS and oversampling through SMOTE, which in turn led to the creation of balanced datasets. This rebalancing was pivotal in enhancing the performance of the models in use. It also showcases the adaptability and resilience of these techniques in the face of real-world data challenges.

The study stands out in its broad evaluation of a variety of ML and DL models. This list includes models like XGBoost, AdaBoost, RF, FFNN, standalone 1D-CNN, and standalone LSTM. The primary aim was to gauge their effectiveness in accurately detecting pilots' mental states. The core findings of the research pinpoint the XGBoost algorithm and the proposed 1D-CNN+LSTM model as the superior in this domain. These models not only exhibit high efficacy but also hold significant potential in enhancing safety and performance measures across the aviation sector. Moreover, the implications of these findings extend beyond aviation, suggesting the utility of these models in various industries where monitoring mental states is of paramount importance.

7.5 Illuminating the Neural Landscape of Pilot Mental States: A Convolutional Neural Network Approach with SHAP Interpretability

This study delves into the challenge of predicting pilots' mental states, a key concern in aviation safety and performance. The EEG data emerges as a potent tool for this detection process. While EEG data offers a promising avenue for detecting pilots' mental states, it isn't devoid of challenges. Traditional ML and

DL models, although powerful, often function as black boxes, with their decision-making processes remaining elusive.

The chapter's primary objective is to craft an interpretable model capable of discerning four distinct mental states in pilots: CA, DA, SS, and NE states, all through EEG data. The research strategy employs a 1D-CNN model trained on PSD features derived from pilots' EEG data.

However, the true novelty of the study lies in its commitment to interpretability. The SHAP values, derived from cooperative game theory, were incorporated to unravel the decision-making process of the model. In essence, SHAP values provide a clear breakdown of how each feature influences the model's prediction. Through this, the researchers could spotlight the top 10 most influential features for each mental state, demystifying the neural underpinnings and offering a transparent view into the model's inner workings.

The results attained by this research are nothing short of exemplary. A staggering average accuracy of 96% stands as a testament to the model's efficacy. Coupled with a precision of 96%, recall of 94%, and an F1-score of 95%, it's evident that the model is both precise and robust. Beyond these metrics, the study provides a deep dive into the effects of various mental states on EEG frequency bands, weaving a tapestry of insights into how different cognitive states manifest in brain activity.

This research's standout contributions are manifold. Firstly, it pioneers an approach that synergizes high-performance model development with enhanced interpretability. Secondly, it offers a comprehensive analysis of the neural mechanisms tied to different mental states. Lastly, by focusing on interpretability, the study addresses a longstanding challenge in the ML domain, ensuring that models are not just black boxes but tools that can be understood and refined. In essence, this work lays a solid foundation for the future of mental state detection in aviation, emphasizing both accuracy and transparency.

7.6 Implications and Broader Significance of ML in Pilot Cognitive Analysis

The preceding sections have meticulously dissected methodologies and findings that utilise ML, DL, and EEG data analysis. This section seeks to extrapolate the broader implications and real-world significance of these studies, transcending academic confines to highlight their transformative potential in practical scenarios.

7.6.1 Safety and Accident Prevention

The paramount benefit derived from these studies is the enhancement of aviation safety. By accurately gauging a pilot's mental state, whether it's fatigue, cognitive overload, or momentary distraction, an early warning system is established. This proactive approach can be pivotal in averting potential mishaps. Human errors, frequently stemming from cognitive challenges or distractions, contribute substantially to aviation accidents. These research findings, therefore, present tools that could drastically mitigate such incidents.

7.6.2 Training and Skill Enhancement

These studies hold transformative potential for flight schools and training programs. By harnessing these models, real-time feedback can be offered to trainee pilots, illuminating areas of cognitive strain or attention lapses. This immediate feedback mechanism allows training modules to be customized, addressing individual weaknesses. The outcome is a cadre of pilots equipped to adeptly navigate high-pressure scenarios, ensuring smoother flights and fewer error-induced challenges.

7.6.3 Enhanced Cockpit Systems

The cockpit of the future may look markedly different, courtesy of these studies. Future aviation systems could seamlessly integrate these models, providing real-time monitoring of pilots. If a pilot's mental state exhibits any anomalies, these advanced systems could intervene. This intervention could manifest as

task simplification, specific function automation, or even alerts to co-pilots, ensuring that the cockpit environment remains conducive to optimal decision-making.

7.6.4 Interdisciplinary Applications

While aviation stands to gain immensely, the methodologies extracted from these studies have broader implications. The tools and techniques, especially those pertaining to EEG data interpretation and ML models, have versatile applications. Fields that demand acute concentration, such as surgical operations, space missions, or even the operation of intricate heavy machinery, can benefit. This cross-disciplinary potential ensures heightened safety and efficiency standards across a myriad of sectors.

7.6.5 Transparent and Explainable AI

One of the foundational challenges in AI is its often inscrutable nature. One of the reviewed studies have emphasized the significance of model interpretability. In an age where trust in AI systems is crucial, especially when regulatory clearances and stakeholder confidence are on the line, this focus on transparency is invaluable. This study, therefore, not only address immediate research questions but also pave the way for a more transparent AI landscape.

7.6.6 Personalized Pilot Well-being Programs

Beyond the confines of the cockpit, these studies have implications for holistic pilot well-being. By decoding the stressors and triggers that affect a pilot's cognitive state, airlines can architect personalized well-being initiatives. These programs can be geared towards ensuring that pilots remain at their cognitive zenith, balancing the rigours of flying with adequate rest and mental recuperation.

7.6.7 Research and Development Catalyst

Lastly, these studies serve as a beacon for the broader research community. By demarcating benchmarks, elucidating best practices, and spotlighting existing

challenges, they invite researchers worldwide to innovate further. This collaborative spirit ensures that the findings from these studies are just the beginning, with more ground-breaking insights on the horizon.

7.7 Limitations and Future Directions

In this thesis, a comprehensive examination of pilots' cognitive states has been undertaken, leveraging the potency of ML and EEG-based techniques. The insights derived hold immense value; however, it is crucial to discern the limitations inherent within the studies and propose future research trajectories that can address these constraints.

Data Constraints: A recurring limitation across the studies is the dependency on EEG datasets from a finite group of pilots. While the depth of analysis is commendable, the breadth, given the limited cohort, might introduce challenges in extrapolating the results to the broader pilot community. To address this, future endeavours could focus on amassing and analysing data from a larger and more varied sample of pilots. There is potential in forging collaborative research partnerships across multiple flight schools or airlines, ensuring a more representative dataset.

Multimodality: While some studies commendably meld EEG data with other sensors, there remains an expanse of uncharted territory in fully exploring the synergies between diverse data sources. Delving into this domain, future research can champion a truly integrated multimodal approach, intertwining EEG data with other physiological metrics.

External Factors: The analytical lens of the studies is predominantly fixated on cognitive states as deciphered from EEG data. Some external variables, pivotal in the cockpit environment, might have been overshadowed. A promising research avenue here would be the integration of environmental sensors. By capturing and correlating external factors like cabin pressure, ambient noise, or even interpersonal dynamics in the cockpit, with EEG data, a more holistic understanding of pilots' cognitive states, influenced by both internal and external stimuli, could be achieved.

Standardisation of Preprocessing: The absence of a uniform preprocessing methodology for EEG data across the studies poses challenges for direct comparative analyses. This divergence underscores the need for future research to architect a standardised preprocessing framework. Collaborative endeavours with peers in the field could be the key, resulting in a cohesive methodology that paves the way for more harmonised research outcomes.

Exploration of Advanced ML Techniques: The potential application of state-of-the-art ML algorithms such as transformers, which have shown promise in other domains for processing sequential data, could be explored in future studies. While the current data structure in the thesis may not be ideally suited for sequential models like transformers, future research could investigate novel approaches to restructure or augment the EEG data to leverage these advanced techniques. Additionally, the exploration of connectivity features, which focus on the interrelationships between different EEG channels, could offer new insights into the neural mechanisms underlying cognitive states. These approaches, however, would require careful consideration of the data structure and preprocessing methods to ensure their applicability and effectiveness.

Real-world Application: The controlled environs of simulated flight scenarios, often employed in the research, might not encapsulate the multifaceted dynamics of actual flights. To bridge this chasm, future studies could venture into real flight environments or refine simulators to more authentically mirror real-world challenges.

Technological Evolution: The realm of ML and AI is perpetually evolving. While the studies employ contemporary models and techniques, the relentless pace of technological advancements means that newer, potentially more effective methodologies could emerge, necessitating continuous updates to the research.

In conclusion, the studies reviewed in this thesis undeniably advance the understanding of pilots' cognitive states, leveraging state-of-the-art ML techniques and EEG data analysis. However, it is essential to interpret the

findings within the ambit of the mentioned limitations. Acknowledging these constraints not only provides a balanced perspective but also offers a roadmap for future research endeavours, aimed at addressing the identified gaps.

8 Conclusion

In the ever-evolving world of aviation, human cognition, particularly that of pilots, stands at the intersection of safety and efficiency. This thesis, through a series of meticulously crafted studies, dives deep into the complexities of pilots' mental states, charting a roadmap for the future of aviation safety. A central theme echoing throughout the research is the irrefutable link between flight safety and a pilot's cognitive state. As aircraft become technological marvels, the human element, with all its cognitive nuances, becomes even more critical. Delving into pilot errors, the research establishes that challenges like fatigue, stress, and attention diversions are more than mere inconveniences; they are potential harbingers of safety breaches.

Harnessing the potential of EEG data emerges as a powerful tool across the studies. This neural electrical symphony, as captured by EEG, offers a window into a pilot's cognitive landscape. The research journey spans from systematic literature reviews, highlighting the current state of ML techniques, to the development of cutting-edge models tailored for EEG data. SVM, ANN, and a blend of CNNs and LSTM models showcase the diverse computational approaches adopted to decode this neural data. Yet, the path to insights isn't straightforward. The research underscores challenges like data imbalances and the inherent noise in EEG readings. Solutions like data resampling and advanced preprocessing pipelines are introduced, emphasizing the research's commitment to accuracy and reliability. Furthermore, the exploration of a multimodal approach, integrating multiple physiological signals, showcases a holistic approach to understanding pilot cognition.

Interpretability, often a missing link in DL models, receives its due importance. The use of SHAP values in one of the studies shines a light on the 'black box' nature of neural networks, ensuring that the models developed are not only accurate but also interpretable, fostering trust and furthering understanding. Individual differences in cognitive states, the need for larger and more diverse datasets, and the potential integration of more advanced neural networks all beckon future exploration. The research also hints at the broader applications of

these findings, from training modules for pilots to real-time monitoring systems in cockpits.

In summation, this thesis stands as a monumental contribution to aviation safety. By intertwining human cognition, advanced computational models, and a relentless pursuit of interpretability and accuracy, it charts a course for a future where the skies are not just busier, but also safer and more attuned to the cognitive well-being of those who navigate them.

8.1 Synopsis of Research Objectives and Key Findings

The research encompasses five comprehensive studies, each meticulously designed to fulfil specific objectives. The pivotal findings from these studies are delineated as follows:

In the foundational study, an exhaustive literature review revealed a pronounced research gap, especially in the domain of ML techniques applied to predict pilots' mental states. The paucity of existing literature underscored the need for the ensuing research, which sought to bridge this gap through novel methodologies.

The subsequent study, addressing the second objective of the thesis, demonstrated that conventional preprocessing techniques did not markedly enhance performance. Intriguingly, the models showcased proficiency in processing data from varied environments. Yet, when EEG signals underwent solely a band-pass filter, a slight decline in classification efficacy was observed. A prominent challenge arose: the models' inability to accurately identify the SS mental state. This limitation, likely rooted in pronounced class imbalances, accentuated the urgency for advanced automated preprocessing methods, with a focus on artefact management and mental state differentiation. Such insights spurred a further exploration into artefact elimination techniques and their influence on ML models.

In alignment with Objective 3, the fusion of cutting-edge automated preprocessing, Riemannian geometry-based feature extraction, and ensemble

learning models emerged as potent in predicting mental states from artefact-free EEG data. With an impressive accuracy of 86%, this approach adeptly addressed EEG artefacts and underscored the potential of EEG systems in mental state detection via spatial features and ensemble models.

The fourth exploration, in tandem with Objective 4, highlighted that of the models employed, XGBoost stood out with exemplary performance, achieving an accuracy of 94.94%. AdaBoost was a close contender, while the proposed 1D-CNN+LSTM model attained an accuracy of 85.76%. Interestingly, the sole use of SMOTE did not significantly bolster the models' capacity to discern specific mental states. To rectify this, the CS method was synergised with SMOTE, pruning instances from the majority class resembling those in the minority class, leading to noticeable performance enhancements. Furthermore, the proposed DL model exhibited amplified efficacy with larger datasets. The introduction of 1D-CNN, renowned in speech recognition realms, marked a pioneering application to EEG data. This phase of research underscored the immense potential of integrating multimodal sensor data with advanced models, like the 1D-CNN+LSTM, for classifying pilots' cognitive states.

The culminating study, aligned with Objective 5, introduced a ground-breaking model leveraging PSD features from EEG data, achieving an impressive 96% performance accuracy across pilots. This model also identified the top 10 salient features for each mental state, with delta and beta frequency bands emerging as crucial determinants.

8.2 Potential Impact

The research has the profound potential to revolutionise the surveillance of pilots' cognitive states, enabling timely predictions and interventions for AHPLS. Such advancements might serve as a vanguard against potential aviation mishaps stemming from AHPLS, fortifying safety protocols within the aviation domain. Additionally, the techniques proposed offer invaluable utility for educational endeavours. Furnishing intricate insights into cognitive landscapes and factors impinging on performance, these methodologies hold promise for

crafting bespoke training curricula. Such programs, tailored to address and mitigate performance inhibitors, could catalyse enhanced pilot efficacy.

The implications of these findings traverse the boundaries of aviation. For instance, the insights unearthed could find resonance in sectors where the surveillance of cognitive states is paramount, spanning domains like air traffic management, nuclear energy plant operations, and even the healthcare sector. Furthermore, the research's initiative to infuse transparency into DL models designated for cognitive state detection amplifies trust. Such endeavours could kindle greater confidence and adoption of AI methodologies in arenas where safety is paramount. This could catalyse the digital metamorphosis of sectors where the comprehension and management of human cognitive states stand central.

To encapsulate, the research carries the potential to instigate transformative shifts not just within the aviation sphere but across diverse sectors. It paves the way for safer, more optimised operational landscapes while championing the judicious application of AI.