

Article

Decision Tree-Based Pilot Workload Prediction Through Optimized HRV Features Selection

Carmelo Rosario Vindigni * , Giuseppe Iacolino , Antonio Esposito , Calogero Orlando  and Andrea Alaimo 

Department of Engineering and Architecture, Kore University of Enna, 94100 Enna, Italy;
giuseppe.iacolino@unikorestudent.it (G.I.); antonio.esposito@unikore.it (A.E.);
calogero.orlando@unikore.it (C.O.); andrea.alaimo@unikore.it (A.A.)

* Correspondence: carmelorosario.vindigni@unikore.it

Abstract

This research explores the use of physiological signals derived from heart activity to assess mental effort during flight-related tasks. Data were collected through wearable sensors during simulations with varying cognitive demands. Specific indicators related to heart rate variability (HRV) were extracted and tested in different combinations to identify those most relevant for distinguishing levels of mental workload (WL). A Random Forest (RF) ensemble method is applied to classify two conditions, and its performance is examined under various settings, including model complexity and data partitioning strategies. Results showed that certain feature pairs significantly enhanced classification accuracy. The best features settings obtained from the RF are then used to train the other two decision trees-based classifiers, namely the AdaBoost and the XGBoost. Moreover, the decision trees models output is compared with predictions from a Kriging spatial interpolation technique, showing superior results in terms of reliability and consistency. This study highlights the potential of using heart-based physiological data and advanced classification techniques for developing intelligent support systems in aviation.

Keywords: pilots mental workload monitoring; HRV; decision trees classification; aviation human factors

1. Introduction

While approximately 75% of aviation accidents are attributed to human error [1,2], it is important to recognize that human error is not a monolithic phenomenon; rather, it originates from multiple, distinct sources. Pilot errors may stem from fatigue, which contributes to 15–20% of accidents [3], inadequate training, lapses in situational awareness, or poor communication, among other factors [4]. Among these, high workload (WL) has emerged as a particularly significant causal factor. According to NASA's Aviation Safety Reporting System (ASRS), high workload is cited in approximately 80% of pilot incidents and accidents associated with crew error [5,6]. Workload encompasses a combination of mental and neural states that influence human performance across perceptual, cognitive, and sensorimotor domains [7]. Many incidents can therefore be linked to lapses in attention and a general reduction in operational performance resulting from fatigue and cognitive overload. Consequently, monitoring fatigue and WL levels in real time is crucial, as it enables swift corrective actions that can help avoiding critical safety failures [8].

Pilot WL is usually assessed through a combination of methods, including self-reported evaluations, which serve as subjective metrics, and the analysis of physiological



Academic Editor: Miroslav Kelemen

Received: 4 November 2025

Revised: 18 December 2025

Accepted: 6 January 2026

Published: 9 January 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

signals, which serve as objective metrics [9]. Among the various self-reported evaluations, the NASA Task Load Index (NASA-TLX) [10], along with the Subjective Workload Assessment Technique (SWAT) and the Cooper–Harper scale, is one of the most widely adopted subjective metrics for evaluating perceived WL [11]. According to refs. [12,13], subjective metrics may be influenced by the pilot's psychological state failing to represent the actual WL level [14]. For this reason, they are often complemented by objective measures, which are typically less affected by personal factors. Objective measures involve collecting and interpreting biometric data such as heart activity [15], eye tracking [16], body temperature [17], respiration rate [18], and brainwave patterns (EEG) [19] to presume the level of WL experienced by pilots [20].

Among the various physiological signals, Electrocardiogram (ECG) data have been frequently employed to detect pilot WL. In particular, heart rate variability (HRV), a well-established indicator of autonomic nervous system activity [21], has proven to be a reliable metric for assessing WL conditions [22]. In addition, advances in Machine Learning (ML) techniques have directed attention toward HRV-based models leading to significant progress in automated classification of the WL level. Specifically, ML algorithms learn to map HRV measurements to workload categories by training on pilot data. Several studies in the literature (e.g., [23,24]) have employed advanced surrogate models for classification tasks. However, the Random Forest (RF) model has attracted particular interest due to its flexibility and robustness. In this regard, ref. [25] employed the RF model within an ensemble learning approach to classify pilot WL into three cognitive levels: low, medium, and high. The method identified five key predictors of simulator performance, demonstrating its suitability for exploring multidimensional human factors in aviation safety research. Similarly, Rajendran et al. [26] collected, processed and classified ECG signals from eighteen subjects to determine whether the pilots were in a state of Channelised Attention (CA), Diverted Attention (DA), or Startle/Surprise (SS). Among the various ML models evaluated, the RF classifier achieved the highest performance, reaching an accuracy of 90.62%. Moreover, in ref. [27], the performance of RF, AdaBoost, and XGBoost was compared for classifying the same mental states using NASA high-fidelity flight simulation data, comprising features extracted from ECG, EEG respiration, and GSR, recorded on a sample of 18 pilots [28,29]. The study reported that XGBoost achieved the best overall performance; however, the dataset was highly imbalanced across classes, which may have biased the results. In ref. [30], HRV features were employed to classify pilots' mental workload using data collected from 7 pilots. Several Machine Learning classifiers were evaluated, among which Random Forest was included. For the binary classification of WL (Low WL vs. High WL), the Random Forest model achieved the best results among the studied classifiers with an overall accuracy of 76%. According to these studies, the present work aims to provide a more robust comparison of tree-based classifiers using a balanced dataset of pilots' WL levels.

Generally, decision trees-based models stand out for their robustness, interpretability, and ability to handle heterogeneous and noisy data [31]. Their strong balance between predictive accuracy and practical applicability made them a suitable choice for the present research work.

In light of the above, employing flight simulators (FS) becomes essential for evaluating the psycho-physical readiness of pilots in critical situations, offering valuable opportunities for both training [32] and stress assessment [13]. Nowadays, a variety of FS are available, ranging from basic setups [33] to Full-FS (FFS) [34] which are characterized by a six degrees of freedom movement capability and high reliability [35]. Although the utilization of an FFS may pose some limitations, including purchase and operational costs [36], it still represents the most effective and reliable option currently available for realistic pilot

WL assessment. Several studies centered on FFS can be found in the literature. For instance, in ref. [37] an FFS was employed to replicate the complex and dynamic conditions of a nocturnal flight mission. The simulator allowed for the precise manipulation of WL variables and continuous monitoring of pilot performance without introducing real-world risks. The results demonstrated that increased WL significantly impaired sustained attention capabilities. Furthermore, the FFS provided objective data, revealing the cognitive limitations imposed by elevated task demands and pointing out the need for effective WL and fatigue management strategies in aviation settings. In refs. [38,39], pilot WL was assessed across FFS and real-flight scenarios using subjective and physiological measures, demonstrating high correlation between environments and supporting the validity of simulator-based WL prediction using HRV data. Similarly, in ref. [15], HRV data of pilots obtained using an FFS were related to WL levels, demonstrating the validity of the approach.

Therefore, the present study contributes to the assessment of pilots' cognitive workload as a means to enhance operational safety. By exploiting physiological signals, specifically HRV, in combination with validated Machine Learning techniques, this work aims to support the development of intelligent cockpit systems capable of detecting when a pilot's mental workload reaches levels that may impair performance, and, eventually, adapting automation and task presentation accordingly. In particular, this study focuses on comparing the classification performance of decision tree-based algorithms, which are selected as strong candidates for potential real-world implementation due to their high interpretability, especially when compared to more complex neural network models. This choice aligns with the European Union Aviation Safety Agency (EASA) Artificial Intelligence roadmap [40], which emphasizes that AI systems must be highly interpretable to be integrated safely into the cockpit. In fact, according to [40], the integration of Artificial Intelligence (AI) technologies in aviation must adhere to principles of trustworthiness, robustness, and human-centric design. Within this framework, monitoring pilots' mental WL is crucial for maintaining situational awareness and ensuring appropriate levels of human-AI interaction. By recognizing pilots' cognitive states, AI-driven decision support systems can adapt their behavior accordingly, supporting operational safety without overwhelming or under-loading the human operator. Aligned with the objectives of the International Civil Aviation Organization (ICAO) Global Aviation Safety roadmap [41], this work contributes to the proactive mitigation of operational safety risks through intelligent systems capable of adapting to pilot states. By predicting mental WL using physiological signals, specifically HRV, this approach supports safety management systems and complements safety enhancement initiatives aimed at improving situational awareness and decision-making under varying cognitive demands. Furthermore, the study reflects the principles of Industry 5.0, which highlight the need for a more human-centric, resilient, and sustainable industrial paradigm. It embraces the notion of the Healthy Operator exploiting non-invasive wearable technologies to model cognitive fatigue and WL. Addressing key challenges such as inter-individual variability and the dynamic nature of mental WL, the proposed methodology employs decision trees models for WL detection, proposing a roadmap for integrating it into a unified framework aimed at supporting the development of intelligent cockpit systems that enhance pilot performance and flight safety.

In this framework, the present work aims to assess which HRV indices are most effective for decision tree-based classification in dynamic flight conditions, while prior studies have used various HRV metrics, systematic evaluation of their performance under different combinations within ensemble Machine Learning frameworks on simulator data from experienced pilots remains limited. This study analyzes five complementary HRV features, time domain (SDNN, HRV_{ti}), frequency domain (LF, TOT_{pow}), and nonlinear

(SD2) indices, to determine the optimal combination for workload classification using Random Forest (Version 4.7-1.2), AdaBoost (scikit-learn version 1.8.0), and XGBoost (version 3.1.2). Data were collected from experienced pilots (mean 633 flight hours) performing standardized tasks in a full-flight simulator. A balanced dataset with equal representation of low and high workload conditions allows fair assessment of algorithm sensitivity to the WL states considered.

The paper is organized as follows: Section 2 presents the methodology, including the experimental setup, participant details, flight task description, and HRV feature selection. Section 3 presents the study's principal findings, detailing the statistical evaluation of physiological indices and the performance of the ML models. It also offers a comparison between the RF, AdaBoost, and XGBoost classifier developed in the present work and a prior Kriging-based study [42], highlighting distinctions in experimental methodology, and predictive efficacy. Section 4 outlines future research directions, focusing on system integration and real-time implementation within intelligent cockpit frameworks. Finally, Section 5 concludes the paper, summarizing key findings and their implications for aviation safety and human–machine interaction.

2. Methodology

This section outlines the experimental setup employed to generate the dataset used for training and evaluating the selected decision trees-based ensemble algorithms and the Kriging model. A detailed explanation of the rationale behind the selection of key HRV features, which were chosen based on their relevance to stress and WL assessment, is also provided. Moreover, the flight tasks performed by the participants are described, providing context for the physiological data collected. Finally, a summary of the resulting dataset is presented, highlighting its structure and key characteristics for subsequent modeling and analysis.

2.1. Test Sample and Experimental Setup

In this study, HRV metrics were used to compare WL levels across different flight phases, classified as category C according to Military Specification 8785C [43]. The experimental data were collected during exercises performed by 34 experienced pilots using a ground-based Full Flight Simulator (FFS) located at the Mediterranean Aeronautics Research and Training Academy (MARTA) Centre, a facility of the Kore University of Enna (see Figure 1). Although the size of the experimental dataset is relatively limited, it is consistent with sample sizes commonly adopted in simulator-based studies on pilot workload assessment, as reported in the literature (e.g., [44–46]). For the sake of completeness the statistical characterizations of the test sample is briefly reported in the following; for a more detailed description about the experimental procedure and related statistical analyses the interested reader is referred to a previous work by the authors [15]. The anthropometric characterization of the test sample is shown in Table 1, indicating a homogeneous physical profile among the pilots. None of the individuals reported any medical conditions, illnesses, or specific factors that could potentially affect physiological measurements.

Table 1. Summary of participants' demographic and physical characteristics (29 males, 5 females).

Variable	Mean	SD	Median	Range/Notes
Age (years)	31.8	8.1	29	23–52
Flight experience (hours)	633.1	146.2	285	–
Height (m)	1.74	0.06	1.75	–
Weight (kg)	75.4	11.7	75.0	–
Body Mass Index (males)	25.18	2.71	–	Healthy range
Body Mass Index (females)	22.96	1.96	–	Healthy range



Figure 1. CESSNA Citation C560 XLS FFS at the Kore University of Enna, Italy.

Objective data were derived from the analysis of each pilot's ECG signal, focusing specifically on HRV. A key parameter used in this analysis is the Inter Beat Interval (IBI), which represents the time interval between two consecutive R-peaks in the ECG waveform. This metric is widely recognized as one of the most informative indicators for assessing stress levels [9,47]. ECG signals were recorded during the flight tasks using the EcgMove 3 sensor, developed by Movisens GmbH [15,48]. The device was worn using a chest strap, with electrodes placed directly on the pilots' chest to ensure reliable signal acquisition during the simulated flight missions. The EcgMove 3 provides raw data at a sampling rate of 1024 Hz for ECG, 64 Hz for 3D acceleration, and 8 Hz for barometric altitude. The Kubios HRV Premium software version 3.5.0 was used for data visualization, selection of relevant time intervals, and extraction of HRV features from the ECG signal [49,50].

2.2. Flight Tasks

Each test included phases typical of flight and every pilot followed the same protocol and gave informed consent. The flight segments were labeled using acronyms aligned with the ICAO ADREP Taxonomy. Specifically, the following:

- Standing (STD): The aircraft remains stationary at the runway threshold. During this phase, pilots are seated at rest for five minutes to establish a physiological baseline;
- Take-Off (TOF): The aircraft initiates and completes the takeoff procedure;
- Maneuvering (MNV): This phase involves two steady turns, with the second turn in the opposite direction of the first, performed consistently by all pilots;
- Landing (LDG): The aircraft performs the final descent and landing.

Each of these segments was standardized to a five-minute duration, in accordance with established guidelines for reliable HRV analysis [51]. Moreover, the handling quality classification [43] provides context regarding the complexity of each flight phase. According to this framework, TOF and LDG are classified as category C terminal phases, typically requiring high-precision flight path control.

2.3. Dataset

To define a reliable dataset for training and evaluating the classification models, physiological data were collected from the 34 individuals. The mean scores μ of selected physiological indices across the sample were used as reference to study the statistical robustness of the dataset. Specifically, to examine variations across the flight phases, a one-way repeated measures ANOVA was conducted. The Greenhouse–Geisser correction was applied to adjust the degrees of freedom when the assumption of sphericity was

violated. Descriptive statistics, F -values, and effect sizes (η^2) are reported in Table 2. All statistical tests were conducted with a significance threshold of 5%. The statistical analysis demonstrated significant main effects of flight phases on all five physiological indices under investigation, with a p -value of $p < 0.001$, validating the inclusion of these indices in ML model development.

Table 2. Statistics of the physiological indices across flight phases and ANOVA results.

Index	STD		TOF		MNV		LDG		F	η^2
	μ	SD	μ	SD	μ	SD	μ	SD		
SDNN	41.70	15.01	21.37	12.06	20.72	13.54	20.48	13.98	(2.207, 72.840) = 47.955	0.592
HRV_{ti}	9.51	3.28	4.62	2.22	4.73	2.66	4.43	2.32	(1.796, 59.277) = 75.783	0.696
LF	1206.1	807.1	316.3	335.28	337.01	376.39	320.83	379.88	(1.553, 51.255) = 41.053	0.554
TOT_{pow}	1816.7	1319.9	493.1	599.3	529.8	672.4	486.1	594.7	(1.682, 55.494) = 33.889	0.507
SD2	54.46	18.43	27.28	14.38	26.49	15.94	25.88	17.03	(2.173, 71.713) = 61.823	0.652

To define a reliable binary classification scheme, the TOF, LDG, and MNV phases were grouped into a single High WL category; this choice was supported by three main considerations: (i) according to Military Specification 8785C, these phases fall under Category C, which requires high-precision flight-path control; (ii) their physiological profiles are highly consistent; in ref. [44], it is shown that these phases elicit comparable autonomic nervous system activation and similarly degraded flight performance, indicating a shared cognitive state; (iii) from an operational standpoint, ref. [52] demonstrates that machine-learning classifiers achieve superior performance when trained on multi-task and heterogeneous datasets. Based on these considerations, the data were organized into two distinct WL levels, namely Low for the standing baseline data and High for the remaining flight phases. To evaluate the significance of the WL levels across the five physiological measures, a new ANOVA analysis was performed, revealing that the combined set effectively differentiates between the two WL levels, as reported in Table 3. For the sake of completeness and clarity in the reader's interpretation of the data provided, Table 3 lists the low-level data again in terms of means and Standard Deviation (SD).

Table 3. Statistics of the physiological measures for Low and High WL levels.

Index	Low		High		F	η^2
	μ	SD	μ	SD		
SDNN	41.696	15.013	20.854	11.708	93.591	0.739
HRV_{ti}	9.509	3.283	4.590	2.230	107.978	0.766
LF	1206.147	807.103	324.716	313.126	52.337	0.613
TOT_{pow}	1813.685	1319.878	502.981	522.558	46.208	0.583
SD2	54.458	18.425	26.550	14.186	113.644	0.775

3. Results and Discussion

Defining the dataset for training and testing of the selected ML algorithms, the analyses first focused on the pre-processing steps of the dataset and the various aspects involved in training the classifier, ensuring that the data are properly prepared for optimal model performance. The first aim was to determine whether certain combinations of features can improve the predictive accuracy of the RF algorithm. To explore this, all possible combinations of the five features were systematically generated and assessed. For each combination, two procedures have been implemented and compared, namely a Leave-One-Out Cross-Validation (LOOCV) and a Stratified k -fold cross-validation with $k = 5$, which was identified in this study as the best trade-off between generalization performance and computational cost. The primary motivation for these analyses derives from the observation that using multiple physiological signals can introduce challenges such as data sparsity in

high-dimensional spaces. This, in turn, may reduce model performance and increase the risk of overfitting when too many features are included [22].

Figure 2 illustrates the classification performance for various feature combinations using LOOCV and stratified k -Fold Cross-Validation. In both validation schemes, the highest classification performance is achieved using the combination of SDNN and SD2, yielding a mean accuracy of 80.88% in both LOOCV and k -Fold CV. The inclusion of the LF feature (i.e., using SDNN, LF, and SD2) does not lead to any improvement in performance, suggesting that LF does not provide additional discriminative information in this context.

The lowest performance under LOOCV is observed for the SDNN and HRV_TI combination, with an average accuracy of 63.24% and a standard deviation of 48.58%. In contrast, the lowest result in the k -Fold CV setting is obtained using the HRV_TI and SD2 combination, with a mean accuracy of 64.40% and a standard deviation of 15.38%. Although the specific combinations differ, both include HRV_TI, indicating that this feature introduces noise.

A notable distinction between the two cross-validation methods lies in the variability of the results. LOOCV generally produces higher standard deviations, reflecting a greater sensitivity to individual data points. In contrast, stratified k -Fold CV yields consistently lower standard deviations across feature combinations, providing more stable and reliable performance estimates. Despite this difference, the overall average accuracies are nearly identical, namely 73.82% for LOOCV and 73.55% for k -Fold CV. The standard deviations of the mean accuracies across all feature combinations are similarly low (3.86% and 3.88%, respectively), highlighting the robustness of the RF model.

Based on these findings, the combination of SDNN and SD2 was selected to develop the WL classifier, as it offers the highest accuracy while minimizing the number of required features. From a physiological perspective, the emergence of SDNN and SD2 as the most informative predictors is consistent with their established roles as indices of autonomic regulation. SDNN, defined as the standard deviation of all normal-to-normal intervals, reflects overall heart rate variability across both low- and high-frequency components and is widely regarded as a global marker of autonomic flexibility and stress-related modulation of cardiac control. SD2, the long-axis descriptor of the Poincaré plot, quantifies long-term variability and the temporal structure of RR-interval dynamics along the line of identity, and is strongly influenced by sustained shifts in a sympathetic–parasympathetic balance. In the present dataset, both indices exhibited large effect sizes and a consistent decrease from the STD phase to the higher-demand phases (TOF, MNV, LDG), indicating a robust and monotonic sensitivity to workload (see Table 2). In contrast, power and HRV_{ti} showed greater intra-phase variability and overlap across flight segments, suggesting that frequency domain metrics and histogram-based indices are more susceptible to influences from respiratory fluctuations, non-stationarities, and limited segment duration in dynamic operational tasks [53].

Following dataset refinement, a comprehensive performance analysis was carried out. Specifically, the influence of two key parameters, the number of trees in the RF and the training/testing data split, was systematically evaluated using a stochastic approach. The assessment employed standard classification metrics, namely out-of-bag (OOB) error [54], precision, recall, and F1-score [55]. For each combination of training set proportion and number of trees, 100 RF models were trained and evaluated. The resulting average values of the OOB error and performance metrics are reported in Figure 3. The lowest value of OOB error $OOB_{min} = 0.022$ has been registered for a train percentage of the 90% and a number of trees of 500. On the other hand, from the analysis of the performance metrics it is noted that the highest values are registered for a number of trees of 300, especially for the class “High”, which is the most important to be recognized by the WL sensing

system object of this project. These results can be explained by considering the trade-off between model complexity and overfitting. In fact, increasing the number of trees generally reduces variance and improves generalization; however, beyond a certain point, adding more trees yields diminishing returns in performance metrics, particularly on specific classes. The peak in performance metrics at 300 trees, especially for the “High” class, suggests that this configuration provides a better balance between model accuracy and generalization. Increasing to more than 300 trees, the model begins to overfit to less relevant patterns, impacting its effectiveness in the classification. In fact, only the SDNN and SD2 features were used resulting in a constrained two-dimensional feature space. In such low-dimensional settings, the law of diminishing returns for increasing the number of trees becomes evident. In ref. [56], it is shown that when the feature dimensionality is low (<5), model convergence typically occurs between 250 and 500 trees, as bootstrap diversity is driven mainly by sample variation rather than by feature subsetting. Similarly, in ref. [57], it is reported that with highly constrained feature sets, approximately 90% of the achievable performance gain is captured within the first 150/200 trees, while additional improvements beyond 300 trees are minimal, contributing on average less than a 0.5% increase in accuracy per additional 100 trees. The trends observed in Figure 3 are consistent with these findings; in fact, the performance curves flatten substantially after 300 trees, and the OOB error decreases by only 0.006 between 300 and 500 trees, indicating that the model is already close to convergence.

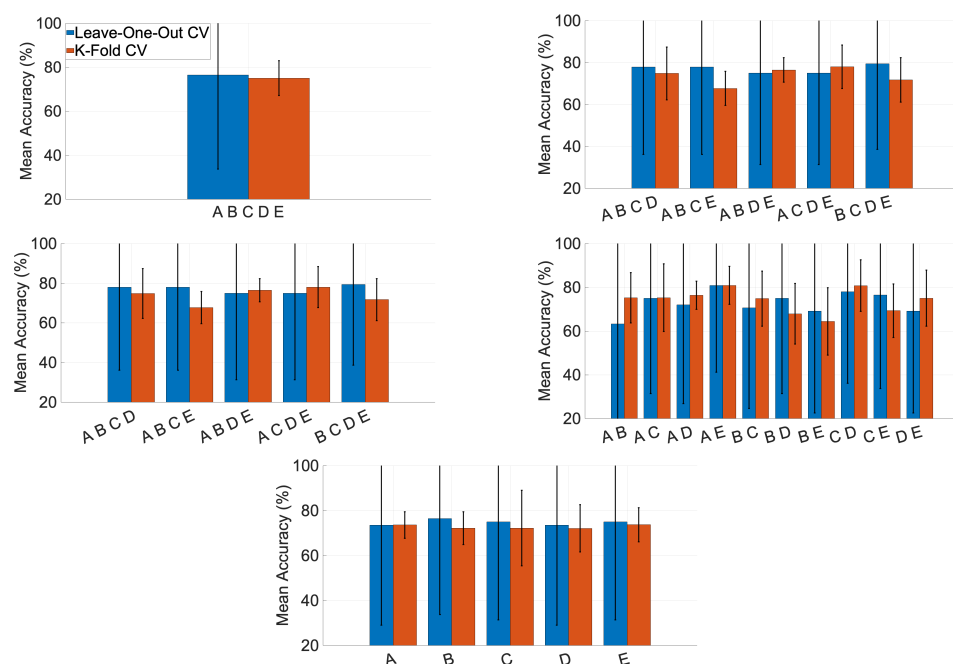


Figure 2. Features combination study results. A = SDNN, B = HRV_{ti} , C = LF, D = TOT_{pow} , E = SD2.

After defining the training set percentage and the number of trees in the forest, validation curves were studied for two key algorithm hyper-parameters: maximum depth and minimum samples per split. Specifically, the maximum depth determines the number of levels in each decision tree; in fact, shallower trees help reduce overfitting by limiting model complexity, while deeper trees can capture more intricate patterns in the data. The minimum samples per split parameter sets the minimum number of samples required to split an internal node; higher values make the model more conservative by preventing the formation of overly specific branches. A k-fold cross-validation with $k = 10$, namely the best test set size, of the RF have been carried out for each combination of these hyper-parameters and the obtained results, in terms of mean metrics values, are reported in Figure 4. The

variability observed in performance metrics across different tree depths and minimum samples per split values suggests that the model is highly sensitive to hyper-parameters tuning, due to the limited dataset size. This motivates future research using larger datasets to better understand the influence of model complexity on generalization performance, as discussed in Section 4. Although no clear trend can be identified in the performance curves with respect to the variation in the hyperparameters, it is nevertheless evident that a specific combination yields the highest values across all evaluated performance metrics. The optimal configuration is obtained with a maximum tree depth of 2 and a minimum number of samples required to split an internal node equal to 8. The classification performance corresponding to this best hyper-parameters setting is summarized in Table 4. As shown, the model achieves balanced and satisfactory results across both classes, with an overall test accuracy of 83%.

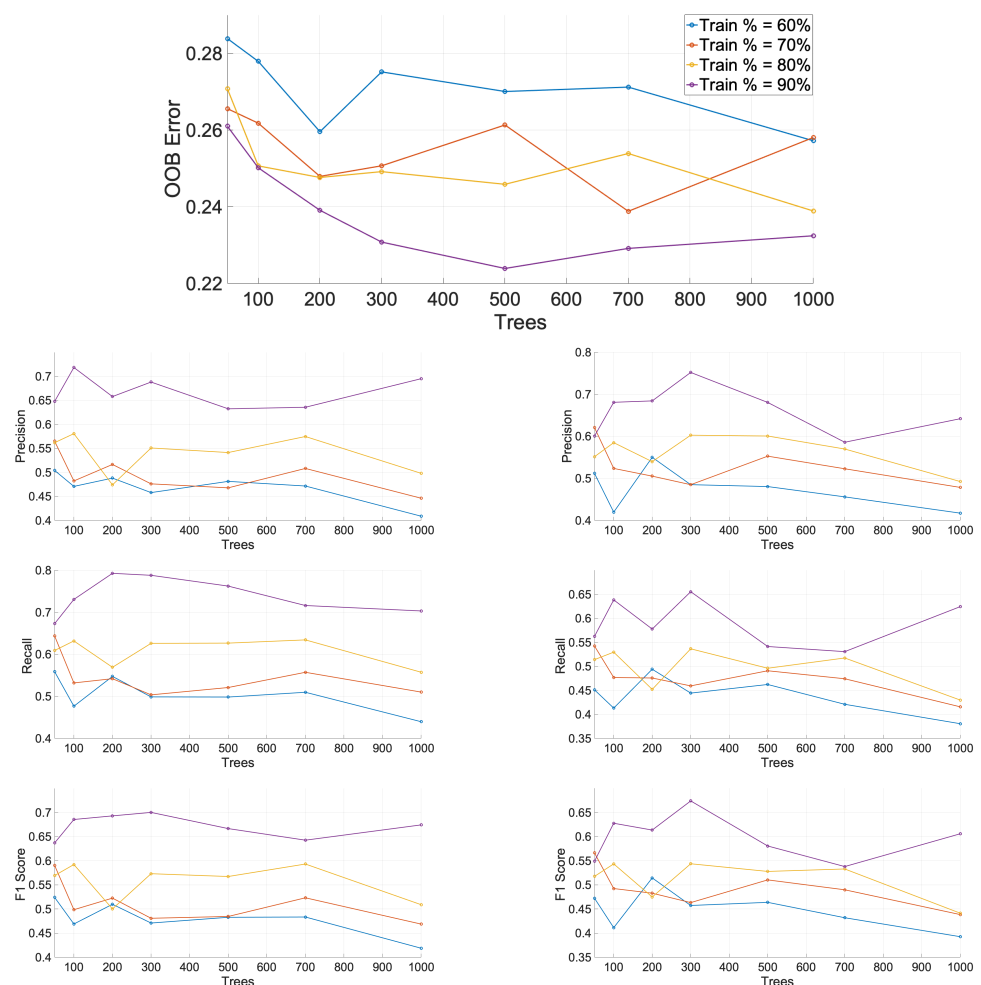


Figure 3. Performance metrics results—number of trees and train percentage.

The outcomes from the baseline RF model are then considered as a starting point to design two enhanced decision tree models; these are AdaBoost [58] and XGBoost [59]. Both methods, like RF, rely on decision trees as base learners, but instead of building them independently, they employ a boosting strategy in which trees are added sequentially to correct the errors of previous iterations. AdaBoost achieves this by reweighting misclassified instances to focus subsequent trees on the harder samples, whereas XGBoost generalizes this approach through gradient-based optimization with a regularized objective, generally improving efficiency, scalability, and overfitting control.

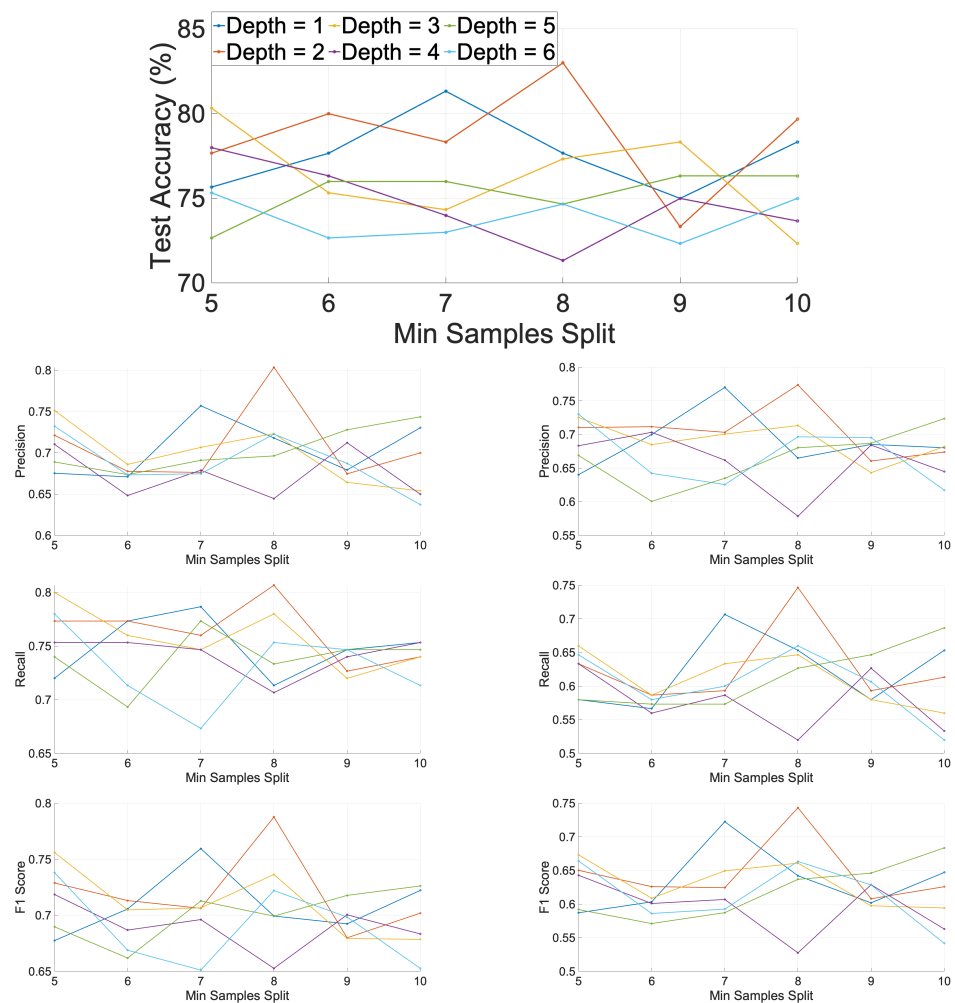


Figure 4. Performance metrics results—trees depth and samples per split.

Both AdaBoost and XGBoost models were trained using the same SDNN and SD2 feature pair identified as optimal in the Random Forest analysis. To explore the AdaBoost model's behavior, a grid search was conducted over four hyperparameters: the number of boosting cycles ($\{50, 100, 300\}$), the learning rate ($\{0.1, 0.5, 1.0\}$), the maximum number of splits in each weak learner ($\{1, 2, 3\}$), and the minimum leaf size ($\{1, 5, 10\}$). All parameter combinations were evaluated through stratified k-fold cross-validation. A notable outcome was that six distinct configurations converged to the same mean accuracy of 80%, indicating that AdaBoost exhibits highly stable convergence for these dataset and feature representation. From these best performing solutions, the selected AdaBoost configuration corresponded to the least complex, consisting of 50 boosting cycles, a learning rate of 0.1, a maximum of 2 splits per weak learner, and a minimum leaf size of 1. Compared with Random Forest, AdaBoost achieved a slightly lower overall accuracy (80% vs. 83%) but substantially improved the recall of the High-workload class (83% vs. 75%), suggesting enhanced sensitivity to high workload conditions, which is particularly relevant in safety-critical contexts.

XGBoost was evaluated following the same methodology, using a grid search over a comparable hyperparameter space: number of boosting rounds ($\{50, 100, 300\}$), learning rate ($\{0.1, 0.5, 1.0\}$), maximum tree depth ($\{1, 2, 5\}$), and minimum child weight ($\{1, 5, 10\}$). As observed with AdaBoost, several parameter combinations yielded the same mean accuracy (81%), confirming the robustness of the SDNN+SD2 feature space across different boosting architectures. Among the best performing configurations, the simplest was

selected, corresponding to 50 boosting cycles, a learning rate of 1, 2 splits per weak learner, and a minimum child weight of 1. The resulting XGBoost model achieved a test accuracy of 81%, slightly lower than Random Forest but higher than AdaBoost. Importantly, XGBoost achieved the highest F1-score for the High-workload class (0.82), with balanced precision and recall, making it the most effective model for detecting workload increases without compromising classification reliability.

The results comparison of the three ensemble methods is provided in Table 4. Random Forest achieved the highest overall accuracy but exhibited reduced sensitivity to High-workload instances. In contrast, both AdaBoost and XGBoost showed improved class-specific performance for the High-workload condition, trading a small reduction in global accuracy for better identification of safety-critical events. The consistency observed across multiple hyperparameter configurations for both boosting methods indicates that the SDNN+SD2 feature combination provides a stable and physiologically meaningful representation for the selected decision trees-based ML models. The results in Table 4 are eventually compared with the ones obtained in ref. [30] where HRV has also been used to classify between Low and High workload states. The present results demonstrate substantial improvements across all three Machine Learning algorithms compared to [30]. Using a significantly larger dataset of 34 pilots, compared to 7 pilots, the trained models achieved superior performance metrics. Random Forest yielded an accuracy of 83% (vs. 76%), AdaBoost achieved 80% accuracy, and XGBoost attained 81% accuracy. Notably, our RF and XGBoost models also improved precision scores, with RF achieving 81% and 77% for the Low and High classes, respectively, vs. 77% overall precision of [30], while maintaining competitive recall values.

Table 4. Comparison of performance metrics for RF, AdaBoost and XGBoost.

Method	Class	Precision	Recall	F1 Score	Accuracy
RF	Low	0.81	0.81	0.79	0.83
	High	0.77	0.75	0.74	
AdaBoost	Low	0.84	0.81	0.80	0.80
	High	0.82	0.81	0.78	
XGBoost	Low	0.83	0.80	0.81	0.81
	High	0.81	0.83	0.82	
RF [30]	Overall	0.74	0.85	–	0.76

For further comparison, the Kriging model is also applied. It is worth mentioning that, while decision trees are based on ensemble learning methods that excel at handling complex, high-dimensional data, Kriging provides probabilistic predictions and models uncertainty effectively [60]. Kriging's strength lies in its ability to model the spatial correlation between the features, which is particularly useful when the relationship between input variables is not entirely linear and involves significant interaction. Since Kriging is a regression-based technique, the class labels Low and High were encoded as binary values: Low was mapped to 0, and High was mapped to 1. This conversion allows Kriging to predict continuous values that reflect the intensity of WL, which have been thresholded with respect to the median value 0.5 to classify the data into Low (<0.5) or High (>0.5) categories. When no a priori information exists regarding asymmetric error costs, a threshold of 0.5 serves as the point of maximum decision entropy. At this value, the classifier is required to commit to one class precisely when the continuous prediction provides no inherent bias toward either category. This choice is appropriate for the balanced dataset used in this work, where both classes carry equivalent operational significance. Moreover, maintaining the threshold at 0.5 avoids introducing a data-dependent bias, as would occur with median-

based thresholds, and instead treats both workload states as symmetrically important. This symmetry is essential because both types of misclassification are undesirable; in fact, false alarms lead to unnecessary and potentially risky interventions, while missed alarms are inherently dangerous since the system fails to anticipate a high-workload condition that may compromise safety. For these reasons, the 0.5 threshold represents the most neutral and operationally justified decision point of the Kriging model designed in this work. Different proportions of training and test datasets have been considered within a stochastic approach, as already presented for the RF model. Specifically, 100 trainings have been performed and the obtained mean values of the performance metrics are reported in Table 5. Increasing the training set percentage from 50% to 90% results in a general improvement in performance metrics. Specifically, both the F1 scores and precision/recall values for the Low and High classes show positive trends. The best test accuracy (70.57%) is achieved when 90% of the data are used for training. This configuration also yields the highest F1 score for the High class (0.71), along with high precision and recall values across both classes. Therefore, using 90% of the data for training provides the best model performance.

Table 5. Performance metrics results for the Kriging model.

Train %	Accuracy	F1 Low	F1 High	Precision Low	Precision High	Recall Low	Recall High
50	66.47	0.63	0.68	0.69	0.68	0.63	0.71
60	66.21	0.61	0.69	0.73	0.67	0.58	0.77
70	67.14	0.64	0.68	0.74	0.68	0.63	0.76
80	67.14	0.63	0.68	0.75	0.66	0.61	0.77
90	70.57	0.65	0.71	0.77	0.69	0.65	0.82

The performance metrics of the decision trees-based and Kriging models are eventually compared to highlight the model's ability to classify WL states. In terms of precision, all three decision tree methods demonstrate superior precision compared to Kriging. Random Forest achieves 0.81 for Low class and 0.77 for High, AdaBoost reaches 0.84 and 0.82, respectively, while XGBoost attains 0.83 and 0.81. In contrast, Kriging achieves only 0.77 for Low class and 0.69 for High class. The substantial gap in High-class precision indicates that decision trees more reliably identify true high-workload instances, reducing false positive rates in safety-critical classifications. The F1 score synthesis of precision and recall shows substantial superiority of decision tree methods, too. Random Forest achieves an F1 of 0.79 (Low) and 0.74 (High), AdaBoost reaches 0.80 and 0.78, while XGBoost attains 0.81 and 0.82. Kriging yields notably lower F1 scores (0.65 for Low, 0.71 for High), indicating poor overall class-specific accuracy. Therefore, the decision trees-based models provides more balanced and reliable classification performance across all considered metrics, making them a better choice for the implementation of a WL monitoring system.

4. Strategic Roadmap

In light of the encouraging results obtained in this preliminary study, a comprehensive strategic roadmap is proposed to consolidate the effectiveness and operational reliability of workload prediction systems (Figure 5). This roadmap is structured in three progressive phases, each addressing specific scientific, technical, and operational challenges while maintaining focus on transitioning from proof-of-concept validation toward real-world aviation deployment.

The first phase is dedicated to establishing robust scientific and operational foundations. A central objective is the expansion and diversification of the experimental dataset. The dataset employed in this study consisted of recordings from 34 pilots with relatively homogeneous demographic characteristics. This limitation primarily reflects the practical

constraints inherent to simulator-based experimental research. Notably, comparable sample sizes are commonly reported in the literature on this topic: for instance, ref. [44] involved 20 subjects, ref. [45] included 40 participants, and [46] was conducted with 20 subjects. This homogeneity provides internal validity but limits generalizability to the diverse pilot populations in modern aviation. Increasing the number of participants, among both genders, spanning various expertise levels and demographic profiles, and incorporating a wider array of flight scenarios, including routine and emergency conditions, will allow the model to better capture the complexity of human cognitive responses in aviation, reducing the risk of bias or overfitting.

Concurrently, this initial phase addresses a critical finding emerging from the present study: significant inter-individual physiological variability in cardiac autonomic responses to cognitive demands, while the Machine Learning models used achieved 81–83% accuracy at the population level, individual pilot HRV responses to identical flight tasks vary considerably due to factors including baseline fitness level, age, circadian state, and prior flight experience. This variability suggests that a one-size-fits-all generic workload classifier, even though well-optimized, may misclassify workload states for certain pilot subpopulations. Consequently, the first phase includes a systematic investigation of personalized adaptation of the workload classifier to individual pilots. Rather than applying fixed decision thresholds derived from population-level training, personalization should involve: (i) establishing individual pilot baseline HRV profiles during known low-workload conditions (e.g., seated rest, routine cruise flight); (ii) adjusting classification boundaries to reflect each pilot's unique autonomic response curve, allowing the system to detect abnormal deviations from that individual's normal range rather than comparing against population norms; (iii) incorporating pilot-specific data collected over multiple flight operations to refine personalized models and account for training effects, fatigue accumulation, and other temporal dynamics. By establishing individual reference profiles, the system gains sensitivity to detect workload-induced deviations more reliably than population-wide comparisons. Additionally, as a necessary intermediate step, implementation and validation in high-fidelity flight simulators has to be addressed to ensure system robustness and safety. The FFS controlled environment allows for thorough testing of the workload classifier's performance under realistic operational conditions, while avoiding risks associated with direct deployment in live flight operations.

During the second phase, devoted to the development and optimization, the focus is on building operationally ready systems with enhanced technical robustness. Specifically, it is needed to explore edge-computing solutions capable of processing physiological signals directly onboard the aircraft, ensuring low latency and timely feedback. The adoption of lightweight, non-intrusive wearable devices have also to be refined, promoting user comfort and measurement stability even in high-stress or prolonged missions, thus enabling practical implementation in real-world aviation scenarios. Moreover, the integration of minimally invasive physiological sensors, including those for electro-dermal activity [61], respiration [62], face temperature [63], and eye activity [64], would enhance the reliability and robustness of the workload evaluation system. This multi-modal sensor integration expands beyond the HRV-only approach, providing complementary physiological markers of cognitive load. From a methodological standpoint, the development phase will investigate hybrid approaches that combine the interpretability and robustness of classical Machine Learning algorithms like Random Forest with the representational power of deep learning architectures such as recurrent and convolutional neural networks. These techniques may offer enhanced ability to capture temporal dependencies, physiological trends, and nonlinear patterns, leading to more precise classifications. Equally crucial during this phase is the development of the human-machine interface. Designing adaptive cockpit

systems capable of modulating automation levels, adjusting task allocation, or tailoring information presentation based on the pilot's inferred workload represents a promising path to reduce cognitive overload, improve situational awareness, and support decision-making. These systems should align with the principles of trustworthy, human-centered AI as outlined by EASA and ICAO [40,41], promoting transparency, ethical alignment, and operational safety.

Finally, the third phase, devoted to validation and deployment, focuses on operationalizing the system in real-world aviation environments. A central objective is the integration of the decision trees-based classifiers into intelligent cockpit environments, enabling continuous, real-time monitoring of pilots' cognitive workload. This integration involves full validation in commercial flight operations across diverse flight scenarios, routine operation, and abnormal or emergency conditions. Long-term studies will be crucial to evaluate the impact of workload monitoring tools on training effectiveness, pilot well-being, and safety outcomes. Such investigations will support empirical validation and contribute to the development of more resilient aviation ecosystems where humans and intelligent technologies collaborate in a balanced and ethically responsible manner. Regulatory validation and compliance with EASA and FAA certification requirements will ensure operational readiness and airworthiness. The three phases illustrated in Figure 5 represent an integrated progression toward operational deployment, with parallel advancement of scientific research, algorithm development, sensor technology, human-machine interface design, and regulatory compliance. This structured approach ensures that each phase builds on validated foundations from prior phases, reducing technical and operational risk while maintaining the rigor and integrity of aviation safety standards.

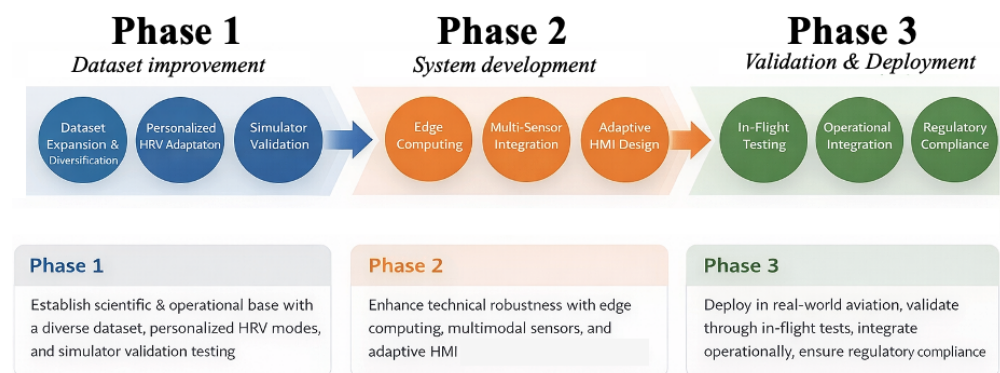


Figure 5. Strategic roadmap.

5. Conclusions

The present study investigated the potential of HRV features, derived from non-invasive ECG signals, as reliable indicators for monitoring pilots' mental WL during flight operations. Within a rigorously controlled experimental framework, data were collected from a sample of certified pilots using an FFS. Among the key findings, tree-based ensemble classifiers demonstrated strong performance in distinguishing between low and high workload (WL) conditions. In particular, the Random Forest model achieved the highest overall accuracy (83%), while AdaBoost and XGBoost provided more balanced per-class performance, with improved precision and recall for the High-WL class. This comparative analysis highlights how boosting-based strategies can complement the robustness of Random Forest by offering improved control over class-specific trade-offs. Overall, the results confirm that tree-based ensemble methods are well suited to model the complex and nonlinear autonomic responses associated with cognitive workload.

Furthermore, the analysis demonstrated that time-domain (SDNN) and nonlinear (SD2) HRV indices are particularly effective for WL classification, as they capture auto-

onomic nervous system dynamics while remaining robust to respiratory artifacts inherent in dynamic flight conditions. The use of a balanced dataset resulted in more conservative and realistic performance estimates compared to previous studies relying on larger but imbalanced datasets, while the accuracy metrics obtained are higher than similar works dealing with binary classification. These findings suggest that effective HRV-based workload classification depends less on increasing algorithmic complexity or dataset features size and more on the alignment between physiological interpretability and operationally meaningful evaluation metrics. Finally, the results emphasize the need for further investigations into personalized WL classifiers capable of adapting to pilots' continuous training and experience, which progressively shape individual physiological responses to external stimuli.

Author Contributions: Conceptualization, A.E., C.O. and A.A.; methodology, C.R.V. and A.E.; software, C.R.V., G.I. and A.E.; formal analysis, A.E.; investigation, C.R.V.; resources, A.E.; data curation, C.R.V., G.I. and A.E.; writing—original draft, C.R.V. and G.I.; writing—review and editing, A.E., C.O. and A.A.; visualization, C.R.V.; supervision, C.O. and A.A.; project administration, C.O. and A.A.; funding acquisition, C.O. and A.A. All authors have read and agreed to the published version of the manuscript.

Funding: Authors gratefully acknowledge support of the “SiciliAn MicronanOTech Research And Innovation Center—SAMOTHRACE” project (MUR, PNRR-M4C2, ECS_00000022).

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Dismukes, R.K.; Berman, B.A.; Loukopoulos, L. *The Limits of Expertise: Rethinking Pilot Error and the Causes of Airline Accidents*; Routledge: London, UK, 2017.
2. Koskelo, J.; Lehmusaho, A.; Laitinen, T.P.; Hartikainen, J.E.; Lahtinen, T.M.; Leino, T.K.; Huttunen, K. Cardiac autonomic responses in relation to cognitive workload during simulated military flight. *Appl. Ergon.* **2024**, *121*, 104370. [[CrossRef](#)] [[PubMed](#)]
3. Goode, J.H. Are pilots at risk of accidents due to fatigue? *J. Saf. Res.* **2003**, *34*, 309–313. [[CrossRef](#)] [[PubMed](#)]
4. Helmreich, R.L. On error management: Lessons from aviation. *BMJ* **2000**, *320*, 781–785. [[CrossRef](#)] [[PubMed](#)]
5. Kharoufah, H.; Murray, J.; Baxter, G.; Wild, G. A review of human factors causations in commercial air transport accidents and incidents: From to 2000–2016. *Prog. Aerosp. Sci.* **2018**, *99*, 1–13. [[CrossRef](#)]
6. McElhatton, J.; Drew, C. Hurry-Up Syndrome: Time Pressure as a Causal Factor in Aviation Safety Incidents. *ASRS Directline* **1993**, *5*, 1–8.
7. Kramer, A.F.; Parasuraman, R. Neuroergonomics: Applications of neuroscience to human factors. In *Handbook of Psychophysiology*, 3rd ed.; Cacioppo, J.T., Tassinary, L.G., Berntson, G.G., Eds.; Cambridge University Press: New York, NY, USA, 2007; pp. 704–722.
8. Wilson, N.; Guragain, B.; Verma, A.; Archer, L.; Tavakolian, K. Blending human and machine: Feasibility of measuring fatigue through the aviation headset. *Hum. Factors* **2020**, *62*, 553–564. [[CrossRef](#)]
9. Charles, R.L.; Nixon, J. Measuring mental workload using physiological measures: A systematic review. *Appl. Ergon.* **2019**, *74*, 221–232. [[CrossRef](#)]
10. Hart, S.G.; Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in Psychology*; Elsevier: Amsterdam, The Netherlands, 1988; Volume 52, pp. 139–183.
11. Rubio, S.; Díaz, E.; Martín, J.; Puente, J.M. Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Appl. Psychol.* **2004**, *53*, 61–86. [[CrossRef](#)]
12. Wong, L.; Meyer, G.; Timson, E.; Perfect, P.; White, M. Objective and subjective evaluations of flight simulator fidelity. *Seeing Perceiving* **2012**, *25*, 91. [[CrossRef](#)]
13. Iacolino, G.; Esposito, A.; Orlando, C.; Alaimo, A. A brief review of pilots' workload assessment using flight simulators: Subjective and objective metrics. *Mater. Res. Proc.* **2023**, *37*, 754–757. [[CrossRef](#)]
14. Hu, L.; Yan, X.; Yuan, Y. Study on the evaluation method of pilot workload in eVTOL aircraft operation. *Heliyon* **2024**, *10*, e37970. [[CrossRef](#)] [[PubMed](#)]

15. Alaimo, A.; Esposito, A.; Faraci, P.; Orlando, C.; Valenti, G.D. Human heart-related indexes behavior study for aircraft pilots allowable workload level assessment. *IEEE Access* **2022**, *10*, 16088–16100. [\[CrossRef\]](#)
16. Peißl, S.; Wickens, C.D.; Baruah, R. Eye-tracking measures in aviation: A selective literature review. *Int. J. Aerosp. Psychol.* **2018**, *28*, 98–112. [\[CrossRef\]](#)
17. Pereira, E.; Sigcha, L.; Silva, E.; Sampaio, A.; Costa, N.; Costa, N. Capturing Mental Workload Through Physiological Sensors in Human–Robot Collaboration: A Systematic Literature Review. *Appl. Sci.* **2025**, *15*, 3317. [\[CrossRef\]](#)
18. Grassmann, M.; Vlemincx, E.; von Leupoldt, A.; Van den Bergh, O. The role of respiratory measures to assess mental load in pilot selection. *Ergonomics* **2016**, *59*, 745–753. [\[CrossRef\]](#)
19. Hernández-Sabaté, A.; Yauri, J.; Folch, P.; Piera, M.À.; Gil, D. Recognition of the mental workloads of pilots in the cockpit using EEG signals. *Appl. Sci.* **2022**, *12*, 2298. [\[CrossRef\]](#)
20. Wang, L.; Gao, S.; Tan, W.; Zhang, J. Pilots' mental workload variation when taking a risk in a flight scenario: A study based on flight simulator experiments. *Int. J. Occup. Saf. Ergon.* **2023**, *29*, 366–375. [\[CrossRef\]](#)
21. Sztajzel, J. Heart rate variability: A noninvasive electrocardiographic method to measure the autonomic nervous system. *Swiss Med. Wkly.* **2004**, *134*, 514–522. [\[CrossRef\]](#)
22. Wang, P.; Houghton, R.; Majumdar, A. Detecting and predicting pilot mental workload using heart rate variability: A systematic review. *Sensors* **2024**, *24*, 3723. [\[CrossRef\]](#)
23. Patel, M.; Lal, S.K.; Kavanagh, D.; Rossiter, P. Applying neural network analysis on heart rate variability data to assess driver fatigue. *Expert Syst. Appl.* **2011**, *38*, 7235–7242. [\[CrossRef\]](#)
24. Mohanavelu, K.; Poonguzhali, S.; Janani, A.; Vinutha, S. Machine learning-based approach for identifying mental workload of pilots. *Biomed. Signal Process. Control* **2022**, *75*, 103623. [\[CrossRef\]](#)
25. Bauer, H.; Nowak, D.; Herbig, B. Helicopter simulator performance prediction using the random forest method. *Aerosp. Med. Hum. Perform.* **2018**, *89*, 967–975. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Rajendran, A.; Kebria, P.M.; Mohajer, N.; Khosravi, A.; Nahavandi, S. Machine learning based prediction of situational awareness in pilots using ecg signals. In Proceedings of the 2021 IEEE Symposium Series on Computational Intelligence (SSCI), Orlando, FL, USA, 5–7 December 2021; pp. 1–6.
27. Alreshidi, I.; Yadav, S.; Moulitsas, I.; Jenkins, K. A comprehensive analysis of machine learning and deep learning models for identifying pilots' mental states from imbalanced physiological data. In Proceedings of the AIAA AVIATION 2023 Forum, San Diego, CA, USA, 12–16 June 2023; p. 4529.
28. Harrivel, A.R.; Liles, C.; Stephens, C.L.; Ellis, K.K.; Prinzel, L.J.; Pope, A.T. Psychophysiological sensing and state classification for attention management in commercial aviation. In Proceedings of the AIAA Infotech@ Aerospace, Grapevine, TX, USA, 9–13 January 2017; p. 1490.
29. 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. Prediction of cognitive states during flight simulation using multimodal psychophysiological sensing. In Proceedings of the AIAA Information Systems–AIAA Infotech@ Aerospace, Grapevine, TA, USA, 9–13 January 2017; p. 1135.
30. Carlsen, V.; Manzi, R.; Dellinger, S.; Craig, T.; Koban, D. Predicting Pilot Workloads Using Physiological Measures. *Ind. Syst. Eng. Rev. (ISER)* **2025**, *12*. [\[CrossRef\]](#)
31. Salman, H.A.; Kalakech, A.; Steiti, A. Random forest algorithm overview. *Babylon. J. Mach. Learn.* **2024**, *2024*, 69–79. [\[CrossRef\]](#)
32. Cameron, B.; Rajaei, H.; Jung, B.; Langlois, R. Development and implementation of cost-effective flight simulator technologies. In Proceedings of the International Conference of Control, Dynamic Systems, and Robotics, Setubal, Portugal, 29–31 July 2016; Volume 126, p. D01.
33. Vidakovic, J.; Lazarevic, M.; Kvrjic, V.; Vasovic Maksimovic, I.; Rakic, A. Flight simulation training devices: Application, classification, and research. *Int. J. Aeronaut. Space Sci.* **2021**, *22*, 874–885. [\[CrossRef\]](#)
34. Baarspul, M. A review of flight simulation techniques. *Prog. Aerosp. Sci.* **1990**, *27*, 1–120. [\[CrossRef\]](#)
35. Oberhauser, M.; Dreyer, D. A virtual reality flight simulator for human factors engineering. *Cogn. Technol. Work* **2017**, *19*, 263–277. [\[CrossRef\]](#)
36. Myers, P.L., III; Starr, A.W.; Mullins, K. Flight simulator fidelity, training transfer, and the role of instructors in optimizing learning. *Int. J. Aviat. Aeronaut. Aerosp.* **2018**, *5*, 6. [\[CrossRef\]](#)
37. Hörmann, H.J.; Gontar, P.; Haslbeck, A. Effects of workload on measures of sustained attention during a flight simulator night mission. In Proceedings of the 18th International Symposium on Aviation Psychology, Dayton, OH, USA, 18–21 May 2015.
38. Zheng, Y.; Lu, Y.; Jie, Y.; Fu, S. Predicting workload experienced in a flight test by measuring workload in a flight simulator. *Aerosp. Med. Hum. Perform.* **2019**, *90*, 618–623. [\[CrossRef\]](#)
39. Fuentes-García, J.P.; Clemente-Suárez, V.J.; Marazuela-Martínez, M.Á.; Tornero-Aguilera, J.F.; Villafaina, S. Impact of real and simulated flights on psychophysiological response of military pilots. *Int. J. Environ. Res. Public Health* **2021**, *18*, 787. [\[CrossRef\]](#)
40. MLEAP Consortium. EASA research—Machine Learning Application Approval (MLEAP): Final report. In *Horizon Europe Research and Innovation Programme Report*; European Union Aviation Safety Agency: Cologne, Germany, 2024.

41. International Civil Aviation Organization. Global Aviation Safety Roadmap. 2007. Available online: <https://www.icao.int/safety/GASP/Pages/Roadmaps.aspx> (accessed on 9 June 2025).
42. Esposito, A.; Iacolino, G.; Orlando, C.; Alaimo, A. Metamodelling of the workload assessment in simulated flights using the Kriging method. In Proceedings of the ICAS Proceedings, Chaing Mai, Thailand, 24–25 October 2024.
43. Standard, M. *Flying Qualities of Piloted Aircraft*; US Department of Defense MIL-STD-1797A: Arlington County, VA, USA, 1990.
44. Yuan, J.; Jia, B.; Zhang, C.; Tian, L.; Yi, H.; Wei, L. Pilot mental workload analysis in the A320 traffic pattern based on HRV features. *Front. Neuroergonomics* **2025**, *6*, 1672492. [[CrossRef](#)] [[PubMed](#)]
45. Wilson, J.C.; Nair, S.; Scielzo, S.; Larson, E.C. Objective measures of cognitive load using deep multi-modal learning: A use-case in aviation. *Proc. ACM Interact. Mob. Wear. Ubiquitous Technol.* **2021**, *5*, 1–35. [[CrossRef](#)]
46. Liu, Y.; Gao, Y.; Yue, L.; Zhang, H.; Sun, J.; Wu, X. A Real-Time Detection of Pilot Workload Using Low-Interference Devices. *Appl. Sci.* **2024**, *14*, 6521. [[CrossRef](#)]
47. Henelius, A.; Hirvonen, K.; Holm, A.; Korpela, J.; Muller, K. Mental workload classification using heart rate metrics. In Proceedings of the 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Minneapolis, MN, USA, 3–6 September 2009; pp. 1836–1839.
48. Movisens GmbH. *EcgMove 3 User Manual*; Movisens GmbH: Karlsruhe, Germany, 2018.
49. KUBIOS OY. *Kubios HRV Premium 3.5.0l*; KUBIOS OY: Kuopio, Finland, 2025.
50. Tarvainen, M.P.; Niskanen, J.P.; Lipponen, J.A.; Ranta-Aho, P.O.; Karjalainen, P.A. Kubios HRV—heart rate variability analysis software. *Comput. Methods Programs Biomed.* **2014**, *113*, 210–220. [[CrossRef](#)]
51. Malik, M. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use: Task force of the European Society of Cardiology and the North American Society for Pacing and Electrophysiology. *Ann. Noninvasive Electrocardiol.* **1996**, *1*, 151–181. [[CrossRef](#)]
52. Matuz, A.; van der Linden, D.; Darnai, G.; Csathó, Á. Generalisable machine learning models trained on heart rate variability data to predict mental fatigue. *Sci. Rep.* **2022**, *12*, 20023. [[CrossRef](#)]
53. Gu, Z.; Zarubin, V.; Martsberger, C. The effectiveness of time domain and nonlinear heart rate variability metrics in ultra-short time series. *Physiol. Rep.* **2023**, *11*, e15863. [[CrossRef](#)]
54. Breiman, L. *Out-of-Bag Estimation: Technical Report*; Department of Statistics, University of California: Oakland, CA, USA, 1996.
55. Naidu, G.; Zuva, T.; Sibanda, E.M. A review of evaluation metrics in machine learning algorithms. In Proceedings of the Computer Science Online Conference, Online, 3–5 April 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 15–25.
56. Probst, P.; Boulesteix, A.L. To tune or not to tune the number of trees in random forest. *J. Mach. Learn. Res.* **2018**, *18*, 1–18.
57. Oshiro, T.M.; Perez, P.S.; Baranauskas, J.A. How many trees in a random forest? In Proceedings of the International Workshop on Machine Learning and Data Mining in Pattern Recognition, New York, NY, USA, 15–19 July 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 154–168.
58. Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [[CrossRef](#)]
59. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
60. Ankenman, B.; Nelson, B.L.; Staum, J. Stochastic kriging for simulation metamodeling. *Oper. Res.* **2010**, *58*, 371–382. [[CrossRef](#)]
61. Masters, M.; Schulte, A. Physiological Sensor Fusion for Real-Time Pilot Workload Prediction in a Helicopter Simulator. In Proceedings of the AIAA SCITECH 2022 Forum, San Diego, CA, USA, 3–7 January 2022; p. 2344.
62. Hebbar, P.A.; Bhattacharya, K.; Prabhakar, G.; Pashilkar, A.A.; Biswas, P. Correlation between physiological and performance-based metrics to estimate pilots’ cognitive workload. *Front. Psychol.* **2021**, *12*, 555446. [[CrossRef](#)]
63. Alaimo, A.; Esposito, A.; Milazzo, A.; Orlando, C. An aircraft pilot workload sensing system. In *Proceedings of the European Workshop on Structural Health Monitoring*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 883–892.
64. Veltman, J.; Gaillard, A. Physiological workload reactions to increasing levels of task difficulty. *Ergonomics* **1998**, *41*, 656–669. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.