

HRV BASED PILOTS WORKLOAD PREDICTION FOR AN INTELLIGENT COCKPIT WARNING SYSTEM

Carmelo Rosario Vindigni[1], Antonio Esposito[1], Calogero Orlando[1] & Andrea Alaimo[1]

[1]Department of Engineering and Architecture, Kore University of Enna, 94100, Enna, Italy

Abstract

This study examines the feasibility of integrating heterogeneous experimental data to enhance dataset dimensionality in cognitive workload assessment, a research field typically constrained by limited sample sizes. The experimental activity was conducted at the Mediterranean Aeronautics Research and Training Academy (MARTA) of Kore University of Enna, using a six-degree-of-freedom full-flight simulator. Electrocardiographic signals are collected during standardized flight simulation tasks from two distinct participant groups: 34 experienced professional pilots and 15 graduate students in aerospace engineering. The simulated missions included representative flight phases and tasks associated with different workload levels, including takeoff, maneuvering, en-route flight, altitude-change maneuvers, and landing. Heart rate variability (HRV) features are systematically extracted from this combined dataset of 49 subjects. By merging physiologically comparable subjects performing equivalent cognitive tasks, the study aims to demonstrate that mixed-population datasets can effectively increase experimental power and the statistical significance of the adopted model. To evaluate whether predictive performance benefits from increased sample size and diversity, two ensemble machine learning models, namely Random Forest (RF) and AdaBoost, have been assessed for binary and ternary workload classification on the heterogeneous dataset. The results indicate that decision tree-based HRV ensemble models can support cognitive workload monitoring in cockpit-oriented applications. However, the ternary classification analysis also shows that the intermediate workload class remains the most difficult to identify, suggesting that additional physiological features may be required to improve the discrimination of intermediate cognitive workload states.

Keywords: Mental Workload Monitoring, Heart Rate Variability, Random Forest Classification, Aviation Human Factors.

1. Introduction

Human error remains a predominant cause of aviation accidents, accounting for approximately 75% of critical occurrences within the aerospace sector [1]. Among the multiple factors contributing to human error, cognitive overload emerges as a particularly critical element for operational safety. According to reports from NASA's Aviation Safety Reporting System, approximately 80% of incidents and malfunctions associated with crew error present cognitive overload as a determining factor. This phenomenon, characterized by a complex combination of mental and neural states that influence human performance at perceptual, cognitive, and sensorimotor levels, represents a significant threat to flight safety, as it causes reduction in attention and degradation of decision-making capabilities [2]. Timely identification of cognitive overload through continuous physiological monitoring is the basis requirement for the implementation of real-time corrective interventions, thereby preventing critical situations. Currently, workload assessment is conducted through a combination of subjective self-evaluation methodologies and objective physiological signal analysis [3]. However, subjective measures frequently suffer from influences originating in the psychological state of pilots, thereby

https://doi.org/10.71945/icas2026_0422

compromising assessment accuracy [4]. Consequently, the scientific community has progressively oriented attention toward approaches grounded in objective biometric data, particularly those derived from cardiac activity, as a reliable indicator of cognitive load states. Heart rate variability represents an established instrument for evaluating autonomic nervous system activity [5]. When coupled with advanced machine learning methodologies, HRV permits automated and reliable classification of cognitive workload levels.

Although HRV-based pilot workload assessment has received increasing attention, the available literature still shows important methodological gaps. A recent systematic review on pilot mental workload detection using HRV identified 29 relevant studies and highlighted substantial heterogeneity in experimental protocols, HRV acquisition procedures, workload labelling strategies, and machine learning pipelines [5]. This variability makes direct comparison across studies difficult and indicates that HRV-based workload prediction has not yet converged toward a standardized modelling framework.

Most previous investigations have focused either on binary workload discrimination or on the statistical characterization of HRV changes across flight phases. Binary approaches are useful for detecting extreme workload conditions, but they provide limited information for adaptive cockpit systems, where intermediate workload states are particularly relevant. In operational aviation, a pilot may progressively move from a nominal workload condition to overload through an intermediate state; therefore, the ability to distinguish low, medium, and high workload levels is essential for developing early-warning and decision-support systems. However, three-level workload classification using HRV features alone remains relatively underexplored.

Recent studies have started to address multi-level workload recognition from cardiac features. Wang et al. [6] proposed a three-level pilot workload evaluation model in a low-visibility flight simulation environment using ECG-derived indices and a Hidden Markov Model, reporting an accuracy of 87.5%. Zhou et al. [7] investigated three task-defined cognitive-load levels during simulated turning manoeuvres using HRV features and machine/deep learning models, with the best performance obtained by a long short-term memory (LSTM) attention architecture. More recently, Yuan et al. [8] analysed pilot workload across the A320 traffic pattern using HRV features and several classifiers, including RF, k-nearest neighbors, gradient boosting decision tree, and extreme gradient boosting (XGBoost); after feature selection, XGBoost achieved the best performance, with an accuracy of 66.67% and an F1-score of 58.33%. These contributions confirm the relevance of HRV for multi-level workload monitoring, but they also show that only a limited number of works have explicitly investigated three-level pilot workload classification with tree-based models. This gap is particularly relevant because decision tree-based models offer a favourable compromise between predictive capability and interpretability. In safety-critical domains such as aviation, model transparency is not only desirable but also necessary to support trust, verification, and human-centred integration of AI-based systems. The European Union Aviation Safety Agency has explicitly emphasized AI explainability, learning assurance, and ethics-based assessment as key concepts for the safe and trustworthy deployment of AI technologies in aviation [9, 10]. In this perspective, decision tree-based algorithms are attractive because their decisions can be analysed through feature importance, split rules, and post-hoc explanation tools such as SHapley Additive exPlanations (SHAP), which are particularly suited to tree ensembles [11]. Compared with deep learning approaches, tree-based models therefore provide a more transparent modelling framework for identifying which HRV features contribute most to workload classification. Within this context, the present study contributes by applying HRV-based pilot workload prediction to a ternary classification problem, distinguishing low, medium, and high cognitive workload levels. Specifically, the analysis focuses on two decision tree-based ensemble classifiers, Random Forest and AdaBoost, selected for their balance between classification performance and explainability. This choice aims to support the development of interpretable physiological monitoring modules for future intelligent cockpit warning systems.

2. Dataset

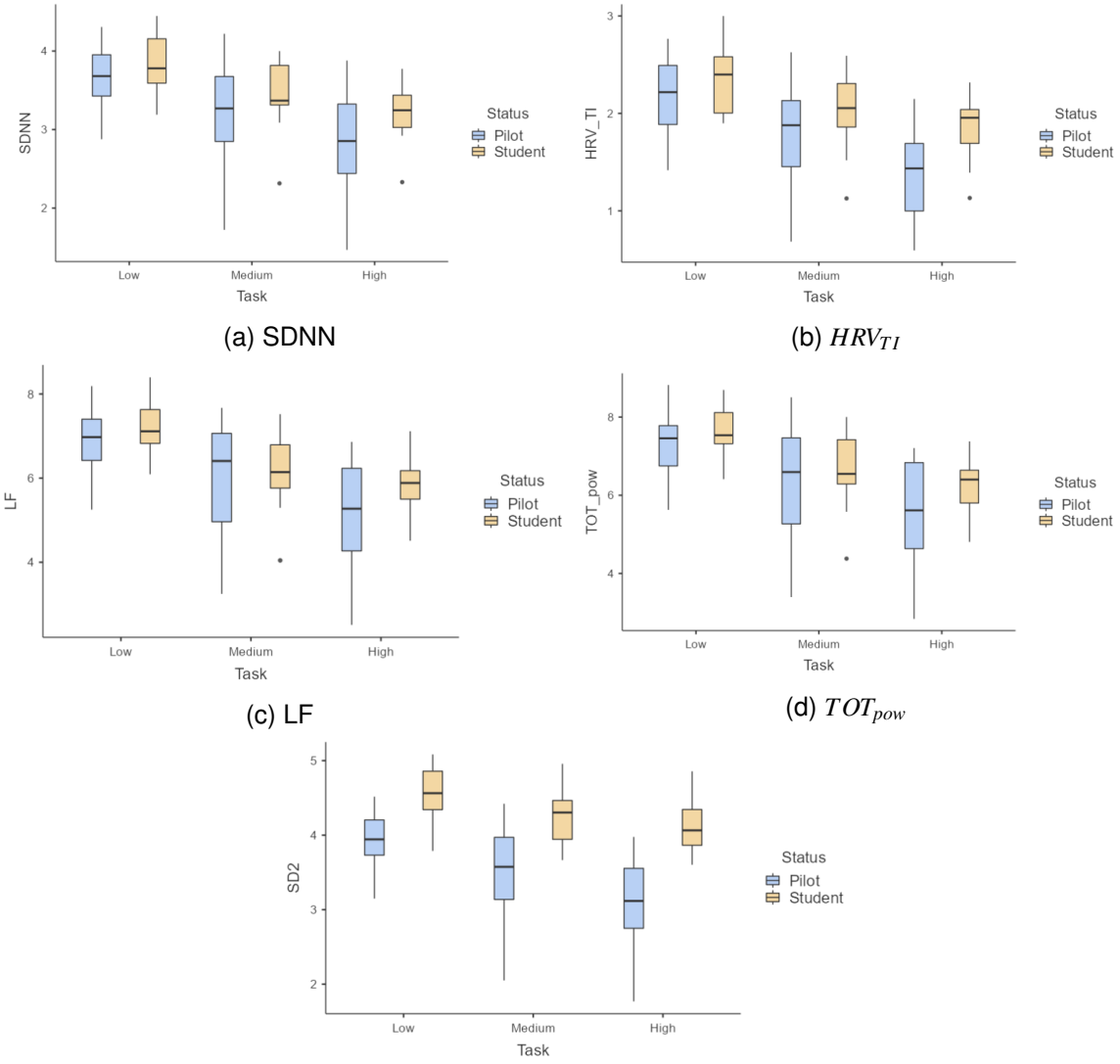
Physiological data were collected at the Mediterranean Aeronautics Research and Training Academy (MARTA) center of the Kore University of Enna, utilizing a full-flight simulator (FFS) equipped with six degrees of freedom movement platform. The first participant sample comprised 34 pilots with substantial operational experience, averaging 633 flight hours and a mean age of 31.8 years [12].

The second participant sample comprised 15 graduate students in aerospace engineering with no flight experience and mean age of 21 years.The group composition demonstrated homogeneity from an anthropometric perspective, with no individual reporting medical conditions or factors capable of influencing physiological measurements.
The experimental procedure involved continuous monitoring of cardiac activity through an EcgMove 3 sensor, positioned on a chest strap to ensure reliable signal acquisition during simulated flight missions. The device recorded electrocardiographic signals at a sampling frequency of 1024 Hz. From these signals, five heart rate variability indices representative of distinct analytical perspectives were extracted: SDNN and HRV_{TI} in the time domain, LF and TOT_{pow} in the frequency domain, and SD2 as a nonlinear index of RR-interval dynamics. The flight protocols were differentiated among the two groups to reflect the different flight experiences.
For the pilot sample, each simulated flight comprised four standardized phases, each lasting five minutes in accordance with consolidated guidelines for reliable HRV analysis [13]. The initial phase, designated "Standing", maintained the aircraft stationary on the runway with pilots in a resting state to establish a physiological baseline reference. The four subsequent phases were: takeoff (TOF), maneuvering (MNV), en-route flight (ENR) and landing (LDG), classified in category C according to military specifications for aircraft handling qualities, requiring precise flight path control. For model development and validation, data were organized into three cognitive workload categories. The low workload condition was defined by metrics acquired during the standing phase, the medium workload condition was identified by the en route phase, while the high workload category comprised aggregated data from the three operational phases (takeoff, maneuvering, and landing).
For the engineering students sample, the experimental protocol consisted of three sequential phases. The first phase was a standing condition, equivalent to that adopted for the pilot group. This was followed by a familiarization phase with the flight simulator, during which participants were instructed to maintain an en-route altitude of 20,000 ft with the motion system deactivated. The final phase involved a controlled altitude-change task during which the participants were required to climb to 22,500 ft within one minute, maintain the target altitude for one minute, and then descend back to the initial altitude within one minute. Workload levels were defined based on these experimental conditions. The low workload condition corresponded to physiological metrics acquired during the standing phase, the medium-workload condition was associated with the familiarization phase, and the high-workload condition included data collected during the controlled climb and descent maneuvers.

Figure 1 presents the boxplot distributions of the five HRV indices after logarithmic transformation, stratified by participant group and workload category. The log transformation was applied to improve graphical comparability across indices with different numerical ranges and to reduce the influence of distributional asymmetry and extreme values. Therefore, the values shown in the figure should be interpreted on the log scale rather than as raw HRV units.
Across the five log-transformed HRV features, both groups show a general decreasing trend from the low to the high workload condition. For SDNN, median values progressively decrease across workload levels, indicating a reduction in time-domain HRV as cognitive demand increases. A similar pattern is observed for HRV_{TI}, with both groups showing lower median values under higher workload conditions. The frequency-domain features, LF and TOT_{pow}, also exhibit a clear downward shift from low to high workload, suggesting a reduction in spectral power with increasing task demand. Finally, the nonlinear index SD2 shows the same overall direction, with lower values in the high condition than in the low condition.
Although the overall trend is broadly consistent across students and pilots, the boxplots also show some group-specific differences in median values and dispersion. In particular, the separation between low and high appears evident in both groups across all features, whereas the distinction between medium and high is more pronounced in the pilot group than in the student group.
The log-transformed boxplots suggest that increasing workload is associated with a generalized reduction in HRV indices, reflecting changes in autonomic regulation under higher cognitive demand.

2.1 ANOVA analysis

To evaluate whether the student and pilot samples showed comparable responses across the three



(a) SDNN

(b) HRV_{TI}

(c) LF

(d) TOT_{pow}

(e) SD2

Figure 1 – HRV data comparison.

experimental conditions, a series of analyses of variance (ANOVAs) were conducted separately within each group. The analyses examined pairwise contrasts between the three levels of the experimental condition (Low, Medium, and High) for each log-transformed HRV feature (SDNN, HRV_{TI}, LF, TOT_{pow}, and SD2). The aim was to assess whether the pattern of condition-related changes was consistent across the two samples and whether the groups could be analyzed jointly.

The post-hoc comparisons conducted separately in the Student and Pilot groups showed a generally consistent pattern across the two samples, particularly for the Low–Medium and Low–High contrasts. In both groups, all HRV features differed significantly between low and medium and between low and high, indicating that the low condition was clearly distinguishable from the higher-intensity conditions. For the Low–Medium comparison, all features were significant in both Students and Pilots. In the Student group, all variables showed p < .05, with stronger effects for LF and TOT_{pow}, which reached $p < .01$. In the Pilot group, all features were significant with $p < .001$. This indicates that the transition from low to Medium led to significant changes in both populations, although the statistical evidence was stronger and more consistent in the Pilot group.

For the low–high comparison, the pattern was even more robust. All HRV features were significant in both groups, with $p < .01$ and, in most cases, $p < .001$. The effect sizes were large in both Students and Pilots, confirming that the extreme conditions, low and high, were clearly differentiated in both samples. This provides the strongest evidence for a broadly similar response to the experimental manipulation across the two groups.

The most critical result concerned the Medium–High comparison. In the Student group, none of the HRV features reached statistical significance, with all p values > .05. In contrast, in the Pilot group, all features were significant, mostly with p < .05 and, for HRV_{TI}, with $p < .01$. This suggests that, in Pilots, the transition from Medium to High continued to produce a significant physiological differentiation, whereas in Students the Medium and High conditions were less clearly distinguishable.
These results indicate that both groups clearly discriminate the low condition from the Medium and High conditions, whereas the distinction between Medium and High appears evident only in the Pilot group. Therefore, although Students and Pilots show a broadly similar response to the experimental manipulation, it would not be appropriate to collapse the two samples while completely ignoring the Status variable.
For the subsequent machine learning analyses, two classification strategies will be considered. First, for a binary classification task based only on the low and high conditions, the ANOVA results support the unification of the Student and Pilot samples. Indeed, the Low–High comparison was significant for all HRV features in both groups, indicating a robust and consistent separation between the two extreme conditions. Therefore, when the classification problem is reduced to distinguishing low from high, the combined dataset can be used without excluding the Student–High observations.
Then, for the ternary classification, a conservative strategy was adopted. Based on the ANOVA results, which showed that the Medium–High contrast was not significant within the Student group, observations belonging to the Student–High subgroup were excluded from both the training and test sets. Therefore, the ternary classifier was trained and evaluated on a restricted unified dataset including Pilot observations across all three conditions and Student observations only for the Low and Medium conditions. This approach avoids introducing a subgroup whose physiological profile did not show a clear separation between Medium and High, while preserving the possibility of evaluating a three-level classification framework. Accordingly, the performance of the ternary model should be interpreted with reference to this restricted dataset rather than the full combined sample.

Table 1 – Post-hoc ANOVA comparisons for Student and Pilot groups across experimental conditions.

Comparison	**Feature**	**Student**			**Pilot**		
		F	**p**	η^2	**F**	**p**	η^2
Low–Medium	SDNN	6.859	< 0.05	0.203	12.744	< 0.001	0.162
	HRV_TI	4.793	< 0.05	0.151	16.156	< 0.001	0.197
	LF	10.371	< 0.05	0.278	11.964	< 0.001	0.153
	TOTpow	9.852	< 0.05	0.267	11.914	< 0.001	0.153
	SD2	4.418	< 0.05	0.141	13.987	< 0.001	0.175
Low–High	SDNN	19.925	< 0.001	0.434	49.650	< 0.001	0.433
	HRV_TI	14.844	< 0.001	0.363	66.349	< 0.001	0.505
	LF	25.123	< 0.001	0.491	46.796	< 0.001	0.419
	TOTpow	24.512	< 0.001	0.485	45.383	< 0.001	0.411
	SD2	8.809	< 0.05	0.253	55.434	< 0.001	0.460
Medium–High	SDNN	2.260	0.144	0.077	5.614	< 0.05	0.080
	HRV_TI	2.097	0.159	0.072	9.842	< 0.05	0.132
	LF	1.061	0.312	0.038	6.042	< 0.05	0.085
	TOTpow	1.035	0.318	0.037	5.272	< 0.05	0.075
	SD2	0.614	0.440	0.022	6.323	< 0.05	0.089

3. Results

3.1 Binary Classification

The comparative analysis of HRV indices across both participant samples groups reveals pronounced physiological similarities, supporting the validity of integrating heterogeneous populations within a unified machine learning framework for cognitive workload classification.
As a preliminary validation of the physiological equivalence between the two participant populations and to assess model generalization capacity, the RF and AdaBoost classifiers [14], originally devel-

oped and optimized on the professional pilot sample in a prior investigation [15], were re-trained and evaluated on the combined Pilots+Students dataset. This validation procedure constitutes a test of model robustness and transferability. The trained RF and AdaBoost models, which had achieved accuracies of 83% and 80%, respectively, on the pilot sample, demonstrated comparable predictive performance when applied to mixed data. Specifically, both algorithms attained an accuracy of 84% on the combined sample.

Table 2 – Comparative performance metrics for Random Forest and AdaBoost classifiers designed in [15] and evaluated on professional pilot and aerospace engineering students test samples.

Population	Method	Class	Precision	Recall	F1 Score	Accuracy
Pilots	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	
Pilots+Students	RF	Low	0.78	0.88	0.85	0.84
		High	0.84	0.78	0.84	
	AdaBoost	Low	0.87	0.74	0.84	0.84
		High	0.76	0.87	0.86	

The comparable results observed demonstrate that HRV-based classifiers can be effectively applied to heterogeneous samples engaged in comparative cognitive tasks within flight simulation environments.

3.2 Ternary Classification

After the binary classification analysis, a ternary classification task was performed on the combined dataset, including both professional pilots and aerospace engineering students, to evaluate the ability of the models to discriminate among low, medium, and high cognitive workload conditions. This represents a more challenging scenario than the binary case, since the introduction of the intermediate workload class increases the overlap among physiological patterns and requires the classifiers to identify more subtle transitions in HRV responses. It is worth noting that, for the ternary classification task, the ML algorithms settings were retained from the optimal configurations obtained for the binary classification [15]. This choice was made to investigate the direct impact of extending the classification problem from two to three workload classes, without introducing further tuning effects. Table 3 reports the average results obtained through stratified 5-fold cross-validation for the Random Forest and AdaBoost classifiers. Overall, AdaBoost achieved comparable performance to the Random Forest, with a mean accuracy of 0.52 compared with 0.51. The improvement given by the AdaBoost is mostly reflected in the class-wise metrics, where it obtained higher precision, recall, and F1-score for all three workload classes.

Table 3 – Average performance metrics for ternary workload classification using stratified 5-fold cross-validation.

Method	Class	Precision	Recall	F1 Score	Accuracy
Random Forest	Low	0.51	0.54	0.60	0.51
	Medium	0.39	0.27	0.30	
	High	0.48	0.51	0.51	
AdaBoost	Low	0.53	0.56	0.64	0.52
	Medium	0.41	0.40	0.42	
	High	0.51	0.54	0.54	

The results show that the low and high workload classes are identified with higher reliability than the medium workload class. This behavior is expected, since low and high workload conditions

correspond to more clearly separated physiological states, whereas the medium class represents a transitional condition between rest-like and highly demanding flight phases. In particular, Random Forest showed a marked limitation in recognizing the medium workload class, with a recall of 0.27 and an F1-score of 0.30. This indicates that a considerable portion of medium workload samples was assigned to either low or high workload categories. AdaBoost partially mitigated this limitation, increasing the medium-class recall to 0.40 and the F1-score to 0.42.

Compared with the binary classification results reported in the previous section, the ternary classification task produced lower overall performance. In the binary case, the models were required to distinguish only between low and high workload states, which are characterized by more pronounced differences in HRV features. Conversely, the ternary formulation introduces the medium workload condition, whose physiological characteristics partially overlap with the adjacent workload levels. Therefore, the reduction in accuracy observed in the ternary task does not necessarily indicate poor model generalization, but rather confirms the increased complexity of discriminating intermediate cognitive workload states using HRV features alone.

Among the two investigated classifiers, AdaBoost provided the most balanced behavior across workload levels. Its higher F1-scores for the low, medium, and high classes suggest a better capability to model the progressive physiological transition among workload states. Nevertheless, the relatively limited performance obtained for the medium class highlights that additional features, larger training samples, or multimodal physiological indicators may be required to improve the discrimination of intermediate cognitive workload conditions.

3.3 Tuned Ternary Classification

The previous ternary classification analysis was performed by retaining the model configurations originally tuned for the binary workload classification task. This preliminary choice allowed the direct effect of extending the problem from two to three workload levels to be assessed, without introducing additional tuning effects. However, the lower performance observed in the ternary case suggested that the hyperparameters selected for the binary discrimination between low and high workload conditions may not be optimal when the intermediate workload class is included. For this reason, a dedicated tuning procedure was carried out for the ternary classification problem. A grid-search strategy was implemented by jointly varying the input feature subset and the main hyper-parameters of each classifier. All possible combinations of the available HRV features were evaluated, in order to identify the subset providing the best discrimination among low, medium, and high workload conditions. For each feature subset, several algorithm-specific configurations were tested using stratified k-fold cross-validation, preserving the class distribution across folds.

For the Random Forest classifier, the tuning involved the number of trees, the maximum number of splits allowed in each tree, and the minimum leaf size. For the AdaBoost classifier, the number of learning cycles, the learning rate, the maximum number of splits of the weak learners, and the minimum leaf size were varied. For each configuration, accuracy, precision, recall, and F1-score were computed for each workload class. In addition, the macro-averaged F1-score was used as the main selection criterion, since it gives equal importance to the three workload classes and is therefore more appropriate than accuracy in the presence of class-dependent performance differences.

The tuned ternary classification results are reported in Table 4. The grid-search procedure identified different optimal configurations for the two ensemble classifiers, both in terms of selected HRV features and hyper-parameters. For Random Forest, the best configuration used only the triangular index of HRV, HRV_{TI}, with 10 trees, a maximum number of splits equal to 2, and a minimum leaf size of 10. This configuration achieved an average accuracy of the 55.6 % and a macro-F1 score of 0.556. For AdaBoost, the best configuration included HRV_{TI} and LF, with 100 learning cycles, a learning rate of 0.1, a maximum number of splits equal to 2, and a minimum leaf size of 1. This configuration achieved the best overall performance, with an average accuracy of 0.608 and a macro-F1 score of 0.609.

The comparison between the two tuned models confirms that AdaBoost provides the best overall ternary classification performance. Compared with Random Forest, AdaBoost increased the average accuracy from 0.556 to 0.608 and the macro-F1 score from 0.556 to 0.609. This corresponds to an

Table 4 – Best configurations and performance metrics for ternary workload classification.

Method	Class	Precision	Recall	F1 Score	Accuracy	Macro-F1
Random Forest	Low	0.774	0.482	0.573		
Trees = 10; Splits = 2	Medium	0.430	0.576	0.483	0.556 ± 0.009	0.556 ± 0.010
Leafs = 10	High	0.618	0.618	0.611		
AdaBoost	Low	0.666	0.673	0.660		
Cycles = 100; Learn Rate = 0.1	Medium	0.568	0.449	0.495	0.608 ± 0.113	0.609 ± 0.113
Splits = 2; Leafs = 1	High	0.611	0.767	0.672		

improvement of approximately five percentage points in both metrics, indicating that the boosting strategy is more effective in modelling the separation among the three workload levels.

An interesting outcome of the tuning procedure is the limited number of HRV features selected by the best configurations. Random Forest achieved its best performance using only HRV_{TI}, whereas AdaBoost improved the classification performance by combining HRV_{TI} with LF. This suggests that the triangular index carries relevant information for discriminating workload levels, while the addition of the frequency-domain LF component provides complementary information that is particularly useful for the boosting classifier.

The class-wise results show different behaviours for the two classifiers. Random Forest obtained the highest precision for the low workload class, equal to 0.774, but this was associated with a relatively low recall of 0.482. This indicates that, when Random Forest assigns a sample to the low workload class, the prediction is often correct, but a relevant portion of low workload samples is missed. Conversely, AdaBoost showed a more balanced behaviour for the low workload class, with precision and recall equal to 0.666 and 0.673, respectively, resulting in a higher F1-score of 0.660.

The medium workload class remained the most difficult condition to classify for both models. However, AdaBoost achieved a slightly higher F1-score for this class, increasing from 0.483 with Random Forest to 0.495. This improvement was mainly due to higher precision, whereas Random Forest showed higher recall for the medium class. This result suggests that Random Forest tends to assign more samples to the medium workload class, while AdaBoost is more conservative but more precise in identifying it.

The largest difference between the two classifiers was observed for the high workload class. AdaBoost achieved an F1-score of 0.672, compared with 0.611 for Random Forest, mainly due to its substantially higher recall. This indicates that AdaBoost was more effective in detecting high workload samples, which is particularly relevant for the development of cockpit warning systems, where missing high workload conditions may represent the most critical classification error.

Therefore, the analysis confirms that tuning the models specifically for the ternary classification task improves the discrimination of workload levels compared with the previous non-tuned ternary analysis. In particular, AdaBoost provided the best compromise between accuracy, macro-F1 score, and class-wise balance, using a compact feature set composed of HRV_{TI} and LF. Nevertheless, the performance obtained for the medium class remains lower than that observed for low and high workload conditions, confirming that intermediate cognitive workload states are characterized by partially overlapping HRV patterns and require further investigation.

4. Conclusion

This study investigated HRV-based cognitive workload prediction using a heterogeneous dataset composed of professional pilots and aerospace engineering students engaged in flight-simulation tasks. The binary classification results showed comparable performance across the two populations, supporting the feasibility of combining physiologically comparable samples to increase dataset dimensionality. The ternary classification analysis confirmed that distinguishing low, medium, and high workload levels is more challenging than the binary low-high formulation, mainly because the medium workload class shows partial physiological overlap with the adjacent classes. After dedicated tuning of feature subsets and hyperparameters, AdaBoost achieved the best performance, with an average accuracy of 0.608 and a macro-F1 score of 0.609 using HRV_{TI} and LF features, outperforming the Random Forest model.

The results suggest that decision tree-based ensemble models can provide a promising basis for HRV-driven workload monitoring in intelligent cockpit warning systems. The binary classification results indicate that HRV features can effectively discriminate between clearly distinct workload levels, particularly when comparing the Low and High conditions. However, the moderate ternary classification performance suggests that HRV alone may not be sufficient to fully characterize intermediate workload states, especially when the physiological separation between Medium and High conditions is weak.

This limitation was particularly evident in the Student sample, where the Medium–High contrast did not show a significant differentiation. Future studies should therefore consider introducing flight tasks with higher level of operational and cognitive demand than the maneuver currently analyzed to elicit stronger physiological activation in the High condition. Future work will also focus on larger datasets, improved workload labeling, and the integration of complementary physiological signals to enhance the reliability and generalizability of HRV-based workload monitoring systems.

5. Contact Author Email Address

mailto: antonio.esposito@unikore.it

6. Copyright Statement

The authors confirm that they, and/or their company or organization, hold copyright on all of the original material included in this paper. The authors also confirm that they have obtained permission, from the copyright holder of any third party material included in this paper, to publish it as part of their paper. The authors confirm that they give permission, or have obtained permission from the copyright holder of this paper, for the publication and distribution of this paper as part of the ICAS proceedings or as individual off-prints from the proceedings.

References

[1] Koskelo J, Lehmusaho A, Laitinen T P, Hartikainen J E K, Lahtinen T M M, Leino T K and Huttunen K. Cardiac autonomic responses in relation to cognitive workload during simulated military flight. *Applied Ergonomics*, Vol. 121, 104370, 2024.

[2] Kramer A F and Parasuraman R. Neuroergonomics: Applications of neuroscience to human factors. In Cacioppo J T, Tassinary L G and Berntson G G, editors, *Handbook of Psychophysiology*, 3rd edition, Cambridge University Press, pp 704–722, 2007.

[3] Charles R L and Nixon J. Measuring mental workload using physiological measures: A systematic review. *Applied Ergonomics*, Vol. 74, pp 221–232, 2019.

[4] Hu L, Yan X and Yuan Y. Study on the evaluation method of pilot workload in eVTOL aircraft operation. *Heliyon*, Vol. 10, No. 18, 2024.

[5] Wang P, Houghton R and Majumdar A. Detecting and predicting pilot mental workload using heart rate variability: A systematic review. *Sensors*, Vol. 24, No. 12, 3723, 2024.

[6] Wang Y, Guo X, Guo S, Jiang F, Liang Z, Peng L and Chai Y. Machine learning evaluation model of pilot workload in a low-visibility environment. *Scientific Reports*, Vol. 15, 20518, 2025.

[7] Zhou W G, Yu P P, Wu L H, Cao Y F, Zhou Y and Yuan J J. Pilot turning behavior cognitive load analysis in simulated flight. *Frontiers in Neuroscience*, Vol. 18, 1450416, 2024.

[8] Yuan J, Jia B, Zhang C, Tian L, Yi H and Wei L. Pilot mental workload analysis in the A320 traffic pattern based on HRV features. *Frontiers in Neuroergonomics*, Vol. 6, 1672492, 2025.

[9] European Union Aviation Safety Agency. Artificial intelligence concept paper issue 2: Guidance for level 1 and level 2 machine learning applications. European Union Aviation Safety Agency, 2024.

[10] European Union Aviation Safety Agency. Artificial intelligence roadmap 2.0: A human-centric approach to AI in aviation. European Union Aviation Safety Agency, 2023.

[11] Lundberg S M, Erion G, Chen H, DeGrave A, Prutkin J M, Nair B, Katz R, Himmelfarb J, Bansal N and Lee S I. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, Vol. 2, pp 56–67, 2020.

[12] Alaimo A, Esposito A, Faraci P, Orlando C and Valenti G D. Human heart-related indexes behavior study for aircraft pilots allowable workload level assessment. *IEEE Access*, Vol. 10, pp 16088–16100, 2022.

[13] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *Circulation*, Vol. 93, No. 5, pp 1043–1065, 1996.

[14] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M and Duchesnay E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, Vol. 12, pp 2825–2830, 2011.

[15] Vindigni C R, Iacolino G, Esposito A, Orlando C and Alaimo A. Decision tree-based pilot workload prediction through optimized HRV features selection. *Aerospace*, Vol. 13, No. 1, 73, 2026.