Journal of Open Aviation Science (2025), Vol.**3**
doi:10.59490/joas.2025.8034

JOAS

ARTICLE

Predicting Air Traffic Controller Workload from Eye-Tracking Data with Machine Learning

Anastasia Lemetti ,*,[1] Lothar Meyer ,[2] Maximilian Peukert ,[2] Tatiana Polishchuk ,[1] Christiane Schmidt,[1] and Helene Alpfjord Wylde[2]

[1]Linköping University (LiU), Norrköping, Sweden
[2]Air Navigation Services of Sweden (LFV), Research & Innovation, Norrköping, Sweden
*Corresponding author: Anastasia.Lemetti@liu.se

(Received: 18 Feb 2025; Revised: 10 Oct 2025 and 17 Dec 2025; Accepted: 17 Dec 2025; Published: 18 Dec 2025)
(Editor: Junzi Sun; Reviewers: Maik Friedrich and Timothé Krauth)

Abstract
In this paper, we examine the feasibility of assessing air traffic controller (ATCO) workload using non-intrusive eye-tracking measures and machine learning algorithms. A total of $N = 18$ ATCOs participated in simulator runs with tasks inducing three task-load levels: light, moderate, and heavy. Task load was modulated through traffic load and the associated increase in complexity. We collected eye-tracking data (statistical summaries of which serve as features) and obtained subjective workload assessments using self-reported Cooper-Harper Scale scores, which act as label variables. We evaluate the performance of eight classical machine learning models, with the k-nearest neighbors and support vector classifier models emerging as the most promising. To optimize performance, we apply feature selection techniques, focusing on these best-performing models. Feature selection via recursive feature elimination (RFE) based on permutation importance reduces the original 42 features while maintaining or improving performance. The outcomes yield promising results in workload-level estimation, achieving an *F1 score* of 0.870 for low/high workload prediction and an *F1 score* of 0.788 for predicting three different levels of workload. The RFE process identifies optimal feature sets ranging from 7 to 13 features for different tasks, with minimal impact on performance. A "knee point" is observed, representing the optimal balance between model performance and dimensionality. Adding more features beyond this point contributes little to performance improvement while increasing model complexity. These findings indicate that even a few features can be sufficient for accurate workload prediction. We show that head-movement features provide valuable information. Comparable performance is achieved using only ocular features, but this requires more features. Asymmetry in left and right eye metrics holds workload-related information but transforming them into averages and differences reduces performance. Retaining the original features separately is the most effective approach, incorporating their absolute differences may provide slight benefits in certain models.

Keywords: ATCO; workload; eye tracking; machine learning

Abbreviations: ATCO: air traffic controller, CHS: Cooper-Harper Scale, ECG: electrocardiography, EEG: electroencephalography, kNN: k-nearest neighbors classifier, RFE: Recursive Feature Elimination, SVC: support vector classifier

1. Introduction

The work of ATCOs, like that of any other profession, produces a subject-dependent mental workload, defined as "a hypothetical construct that represents the cost incurred by a human operator to achieve a particular level of performance" [1]. Workload can impact safety and capacity, as ATCOs must halt operations when they cannot ensure the ability to perform their assigned tasks according to an EU regulation [2]. Thus, it is crucial that the working conditions enable the ATCO to keep a reasonable workload—in line with the inverted-U curve, which describes the dependency of performance on workload [3]. However, the workload depends non-linearly on various factors, for example, the traffic load and complexity, and the age and experience of the ATCO. Hence, monitoring workload levels, while essential for ensuring safety and capacity, is challenging. Various scales are employed for self-assessment of workload, but these have been identified as unreliable, for example, because of deficient self-awareness [4]. Consequently, numerous studies have explored the potential of (objective) physiological or behavioral indicators to assess operators' workload levels [5, 6, 7, 8, 9, 10, 11] and to identify thresholds for workload levels that allow for safe and efficient performance [12]. Such research entails the recording of potential indicators and necessitates a method to gauge the capability of the indicator, or more often of a set of indicators, to estimate or project the workload. Ideally, the identified measures should be non-intrusive to allow application during operations.

Despite progress in workload assessment, important shortcomings remain. Many existing studies continue to rely on subjective ratings, which are limited by bias and intrusiveness (e.g. [13, 14, 15]). More recently, eye-tracking data combined with machine learning has shown promise for workload prediction [5, 16, 17, 18]. However, these efforts have predominantly focused on domains such as piloting, unmanned aerial system operations, or controlled laboratory settings involving cognitive tasks, and have not been validated in the air traffic control context, where tasks, operational demands, and cognitive processes differ substantially. At the same time, while ATCO workload has been studied using physiological measures such as electroencephalography (EEG), electrocardiography (ECG), and electrodermal activity [19, 20], these investigations have rarely incorporated eye-tracking metrics. Consequently, the potential of eye-tracking–based machine learning models for estimating ATCO workload in realistic operational settings remains largely unexplored. The present study addresses this gap by training and evaluating machine learning models on eye-tracking data collected from licensed ATCOs in high-fidelity simulator scenarios. We further contribute by introducing a reclassification of the Cooper–Harper Scale (CHS) to obtain more meaningful workload categories, systematically comparing multiple machine learning models, and demonstrating that eye-tracking features can predict ATCO workload with high performance.

In this paper, we expand on our recent conference paper [21], applying an approach used in less detail in [22]. In both papers, we explored the validity of predicting ATCO workload from eye-tracking and head-movement data using machine learning techniques. In [22], we evaluated five classical models, demonstrating their potential based on *accuracy* and *F1 scores*. In [21], we analyzed performance using a reduced feature set, comparing predictions based on all recorded metrics to those using only a subset. We employed a hold-out validation strategy and developed a method to identify a reduced feature set that maintains or improves model performance.

Here, we aim to further highlight the prediction performance of machine learning techniques while investigating the impact of different feature sets on model performance—ultimately assessing the feasibility and reliability of the applied methods. In this sense, our contribution is threefold:

1. We enhance the reliability of our previous results by providing a more robust evaluation of machine learning techniques for ATCO workload prediction. Through improved methodological rigor, our findings offer stronger evidence of model performance, making them more trustworthy for practical application.
2. We deepen our investigation of feature reduction, with the motivation that constant recording of a large number of features may prove impractical during operation. While additional features are often assumed to improve classification accuracy, too many can lead to overfitting, increased

computational complexity, and multicollinearity, ultimately reducing reliability—a phenomenon known as the "curse of dimensionality" or "Hughes phenomenon" [23]. Therefore, we focus on identifying the most relevant features for workload prediction, which enhances both model performance and interpretability.

3. We evaluate the role of metrics separately recorded from the left and right eye: in the full dataset, the most important features identified by different machine learning models across workload classes often come from a specific eye. We aim to investigate the contribution of left- and right-eye metrics and compare their respective roles in workload classification. Additionally, we provide a more robust assessment of head-movement features, improving the reliability of their evaluation in workload prediction.

To achieve this, we introduce additional eye-tracking metrics, specifically the number of blinks and blink duration; we improve data pre-processing; we test a broader range of machine learning models; we enhance the model evaluation and the feature reduction method by incorporating stratified k-fold validation and additional performance metrics; we analyze feature aggregation for left and right eye metrics; we further analyze the importance of head-movement features; and we examine the composition of the most important feature sets, identifying the most frequently selected features across the datasets for both binary and multi-class workload prediction, along with their source and statistical properties.

Roadmap. In Section 2, we review related work; in Section 3, we describe our data-collection process; and in Section 4, we present the applied machine learning techniques, detail the feature-extraction process, the model development, and the evaluation methods, and report the corresponding results. We provide a detailed discussion of our results in Section 6 and conclude our paper in Section 7.

2. Related Work

In this section, we present an overview of research on the assessment of workload from physiological measures and the application of machine learning techniques for this assessment in air traffic control and other domains.

Assessment of Workload in Air Traffic Control. Frutos et al. [13] developed the Cognitive Complexity computerized model—the Cognitive ModEl for aTco workload Assessment (COMETA) model. This model processes air traffic control events alongside the ATCO Task Model, which consists of air traffic control actions, to estimate the cognitive demand or mental workload associated with these tasks. The algorithm was calibrated and validated using CHS values. Ibáñez-Gijón et al. [14] extended and validated the COMETA model by incorporating the real-time effects of controllers' actions on airspace dynamics. Their validation process involved recording Instantaneous Self-Assessment and NASA-TLX scores, as well as physiological measures such as electrodermal activity and heart rate.

Pagnotta et al. [19] conducted a review of 39 peer-reviewed studies investigating physiological measures of mental workload in air traffic control, with brain and heart-rate metrics being the most commonly examined. They concluded that "positive relations between measures of mental workload and task difficulty were observed most frequently, indicating that the measures indeed allowed the assessment of mental workload".

Meyer et al. [9] investigated ocular and head-yaw measures for their potential as workload (and fatigue) indicators for ATCOs. Based on data recorded in human-in-the-loop simulations of remote towers, they performed qualitative and quantitative comparisons with established reference measures and identified selected ocular and head-yaw measures as potential indicators.

Lemetti et al. [10] analyzed ocular measures as potential workload (and fatigue) indicators using the Fast Fourier Transform to test whether "humans respond to increasing fatigue with harmonic oscillations in the eye movement, while they respond to increasingly high workload with disruptions to these harmonic oscillations". Based on a proof-of-concept study (using the same data as [9]) they verified their hypothesis in some cases and identified a promising ocular indicator for changes in workload with the fixation duration. We refer also to [8] for other references on mental workload measurement and workload in air traffic control.

Assessment of Workload from Physiological Measures in Different Domains. Charles and Nixon [24] reviewed 58 peer-reviewed journal articles across various domains, including but not limited to air traffic control, that focused on measuring and predicting mental workload using ECG, EEG, respiratory rate, dermal activity, blood pressure, and ocular measurements. They concluded that no single measure reliably differentiates mental workload but highlighted specific indicators, such as pupil diameter and blink rate, that can distinguish between low and high workload levels. Additionally, they identified key factors influencing these physiological measures. More recently, Das Chakladar and Roy [25] conducted a review on estimating cognitive workload levels based on physiological measures.

Russo et al. [17] conducted a systematic review of 26 studies on the use of eye tracking to assess mental workload of unmanned aerial system operators. Their findings highlight the limitations of traditional self-assessment methods, like NASA-TLX, and emphasize the advantages of eye tracking in providing real-time insights into cognitive load, visual attention, and operator fatigue. Key eye-tracking metrics, such as pupil dilation, fixation duration, and gaze dispersion, were found to correlate with workload levels. However, the review also identified methodological variations across studies, limiting comparability and generalization. The authors suggest that standardizing eye-tracking methodologies and integrating multimodal physiological data could enhance its application in optimizing human-machine interfaces, training, and operational efficiency in unmanned aerial system operations.

Machine-Learning-Based Assessment of Workload and Other Cognitive States from Physiological Measures in Different Domains. Lobo et al. [18] aimed to classify pilots' cognitive workload based on eye-tracking and EEG data. For this task, they designed a machine learning tool (the Pilot's Pattern Classifier) and set up an experiment in which pilots need to perform a primary and a secondary, interfering task. As they solve a multi-class problem, they employed the k-nearest neighbor (kNN) and the random forest classifier technique in their tool and identified the kNN technique as "really appropriate" to deal with EEG and eye-tracking data.

Zak et al. [26] suggested a real-time workload assessment method using the subjective workload assessment technique and machine learning models based on joystick interaction data from experienced military unmanned-aerial-vehicle operators. In the study, Zak et al. found a high correlation between joystick movement patterns and assessment scores, demonstrating the feasibility of this approach for real-time workload assessment.

Bixler and D'Mello [27] investigated mind wandering, a common phenomenon where attention involuntarily shifts from task-related processing to unrelated thoughts. They developed an automated gaze-based detector for mind wandering using eye-gaze data from 178 users reading instructional texts. The authors trained machine learning models on gaze-fixation features, and detected mind wandering with 74% accuracy.

Vortmann et al. [28] investigated the use of imaging time series for classifying attentional states based on eye-tracking data, finding that imaging time series significantly improves classification accuracy over traditional statistical methods. Their results suggest imaging time series retains more detailed temporal information, leading to better performance in both screen-based and augmented reality tasks, achieving a classification accuracy of 78.3%.

Sims and Conati [29] introduced a deep learning architecture combining recurrent neural networks and convolutional neural networks to detect user confusion using eye-tracking data. Experiments using data from user interactions with a visualization tool demonstrated that this approach achieved an accuracy of 62% in detecting confusion, outperforming a traditional random-forest classifier.

Ktistakis et al. [15] introduced COLET, a publicly available dataset aimed at advancing research in eye-tracking-based cognitive workload estimation. The dataset comprises eye-movement recordings from 47 participants who completed visual search puzzles under varying levels of complexity and time pressure. Cognitive workload was assessed using the well-established NASA-TLX procedure. The authors applied machine learning models to predict workload levels using the COLET dataset, achieving up to 88% accuracy in distinguishing between low and high workload states. Their results indicated that multitasking and time pressure significantly increase cognitive workload.

Kaczorowska et al. [16] classified workload levels based on eye-tracking data collected from 29 volunteers performing a computerized cognitive test. They trained eight machine learning models for three-class classification and achieved a high accuracy of 97% using only a reduced set of seven key features. However, their study assessed workload in terms of task load, defined by objective task difficulty, in a controlled environment.

Machine-Learning-Based Assessment of Workload from Physiological Measures and Other Metrics in Air Traffic Control. As early as 1999, Chatterji and Sridhar [30] used a multilayer neural network to relate statistical measures derived from traffic data to ATCOs' self-assessed workload. They classified workload into three levels: low, medium, and high.

Sciaraffa et al. [20] aimed to classify three levels of ATCO workload using EEG measurements, comparing five machine learning techniques. They noted, "However, most of the work in literature is limited to classify two levels of workload, the low and high. In these cases, the levels of accuracy reached are generally very high, greater than 80%. Much less are the examples of multiclass-classification, whose highest number of workload levels classified has been 7 and, almost all, have been obtained by means of n-back and arithmetic tasks in a laboratory context." In their study, they conducted 45-minute scenarios, each consisting of 15-minute segments designed by operational experts to represent low, medium, and high workload levels, with 35 ATCOs participating. They recorded data from 16 EEG channels and applied multiple feature extraction techniques. Among the tested machine learning methods, kNN achieved the highest accuracy (84%) when using 28 features, with accuracy decreasing as the number of features was reduced. Sciaraffa et al. concluded that machine learning can effectively distinguish three workload levels based solely on EEG data with high accuracy.

Gianazza [31] endeavored to estimate ATCO workload based on past sector operations rather than physiological measures. He proposed that sector collapses indicate low workload, sector maintenance corresponds to medium workload, and sector splits reflect high workload. Using this approach, he achieved a correct prediction rate of approximately 82%.

3. Data Collection

In this section, we describe the data collection process used to investigate ATCOs' workload, focusing on the simulator-based procedures and subjective workload ratings. We provide detailed information on the study participants and outline the simulation-study setup to contextualize the experimental environment.

3.1 Simulator-Based Data Collection

We conducted a controlled laboratory study in the TopSky radar simulation environment at an Area Control Center, testing three scenarios with varying task loads: light, moderate, and heavy. These scenarios were adapted from pre-existing simulator exercises originally developed for ATCO training at Air Navigation Services of Sweden (LFV). They were specifically designed to manipulate ATCO workload by varying factors such as the number of aircraft, traffic density, and complexity of events. Each scenario aimed to represent a different level of operational complexity and task demand. All simulations took place within the same en-route sector in Swedish airspace.

During the experiments, we collected a variety of behavior-based and physiological data using the Smart Eye XO remote eye-tracking system [32]. This system, equipped with two infrared cameras operating at a 250 Hz sampling rate, captures detailed information on head position and movement, eyelid dynamics, and eye-gaze behavior. The recorded measures include blink data, pupil diameter, saccadic movements, fixations, and head rotation. The system offers high precision, with an accuracy of up to 1.5 degrees for head rotation and 0.5 degrees for gaze tracking. Importantly, it is also compatible with participants wearing eyeglasses.

3.2 Definition of Workload

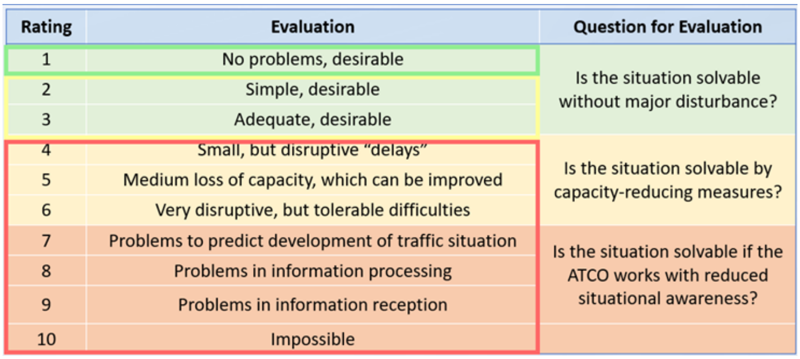
In this work, we follow Hart and Staveland's view of workload as the cost incurred by an operator to maintain performance [1]. While workload is inherently multidimensional and subjective, for the purpose of training machine learning models, we operationalize the construct not through task difficulty or traffic complexity, but through subjective ratings—thereby connecting the theoretical understanding of workload with a practical operationalization that enables predictive modeling based on eye-tracking features.

3.3 Subjective Workload Rating

Various rating scales exist for workload assessment. In this study, we employ an adapted **C**HS for workload evaluation, utilizing a ten-point numerical scale (see [33, 8]). The scale values are determined based on responses to three criteria: if the situation is solvable without major disturbances, the workload score falls within the range of 1–3; if the situation is solvable only through capacity-reducing measures, the score ranges from 4–6; if the situation is solvable only if the ATCO operates with reduced situational awareness, the score ranges from 7–9; a negative response to the final question results in a score of 10. A detailed description of each score is provided in Figure 1. We aim to distinguish three different workload levels: low, medium, and high. Initially, we planned to classify workload levels based on the structure of the CHS, assigning each question category to a corresponding workload level. However, the actual ATCO responses revealed that very few scores exceeded 4. Considering this and incorporating feedback from post-run discussions with ATCOs, we decided to reclassify the workload levels as follows: low as CHS score = 1, medium as CHS score $\in \{2, 3\}$, and high as CHS score ≥ 4. In Figure 1, the background color differentiates the question groups, while the frame color indicates the workload levels.

We selected the adapted CHS over alternatives like the five-point Instantaneous Self Assessment scale [34] due to several issues identified in previous research conducted by some of the authors [8]. Both numerical scales presented challenges. First, they lacked sufficient granularity to capture fluctuations in workload accurately. Second, the simplicity of numerical scales could lead ATCOs to provide quick, reasonable-sounding responses without mentally cross-referencing the corresponding verbal descriptions. Third, both scales were influenced by social desirability, defined as "the bias or tendency of individuals to present themselves in a manner that will be viewed favorably by others" [35], which may cause ATCOs to underreport their workload. Although these limitations affect both scales—and an instrument capable of detecting subtle variations at low workload levels would be ideal—we consider the more detailed CHS, with its 10-point scale, less susceptible to these issues compared to the Instantaneous Self Assessment 5-point scale. Specifically, as our study aims to differentiate between three distinct scenarios, we believe the CHS offers ATCOs a better opportunity to report varying workload levels, even in the presence of social-desirability biases.

The reclassification of CHS ratings was motivated by an observed phenomenon in the range of self-estimated workload values: participants did not utilize the full scale. This likely reflects the extensive training and expertise of ATCOs, who routinely operate under demanding conditions yet report only reasonable workload levels. Another contributing factor may be social desirability effects, which can inflate self-assessments of performance or ability. As a result, familiarity with high-demand situations reduces variability in subjective ratings and makes it difficult to induce high workload levels in simulations. This compression effect can lead to ceiling effects, where ratings fail to fully capture increases in task demands [36, 37]. Moreover, as shown in our earlier work, subjective workload measures such as CHS, NASA-TLX, or the Instantaneous Self Assessment scale also face challenges related to scale interpretation and the lack of calibration, which complicate their direct use in quantitative modeling [9]. Taken together, these considerations justify the reclassification of CHS scores to obtain meaningful workload categories suitable for machine learning model training and analysis.



Rating	Evaluation	Question for Evaluation
1	No problems, desirable	Is the situation solvable without major disturbance?
2	Simple, desirable	
3	Adequate, desirable	
4	Small, but disruptive "delays"	Is the situation solvable by capacity-reducing measures?
5	Medium loss of capacity, which can be improved	
6	Very disruptive, but tolerable difficulties	
7	Problems to predict development of traffic situation	Is the situation solvable if the ATCO works with reduced situational awareness?
8	Problems in information processing	
9	Problems in information reception	
10	Impossible	

Figure 1. Adapted Cooper-Harper Scale

3.4 Participants

A total of 18 licensed ATCOs participated in the study, comprising 9 females and 9 males, all holding valid unit endorsements for the sector used in simulations. Each participant completed three simulation runs, corresponding to three different task-load scenarios (see 3.6), resulting in a total of 54 runs. Eye-tracking measurements were successfully recorded in 53 of these runs (98.1% success rate).

The participants were experienced ATCOs, with a mean age of $M = 46.1$ years ($SD = 8.3$) and an average of $M = 19.7$ years of work experience ($SD = 8.4$). This demographic profile indicates a high level of professional expertise.

3.5 Simulation-Study Setup

Each of the 54 simulation runs lasted 35–45 minutes. At three-minute intervals, an audio signal prompted ATCOs to briefly state their CHS rating aloud. Before the simulations, all participants were

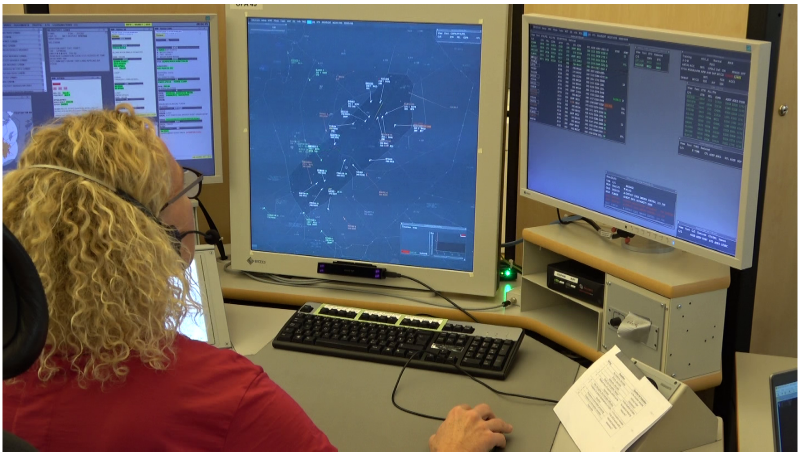


Figure 2. An ATCO performing a simulated scenario with an eye-tracking device on the bottom of the primary screen

briefed on how to provide their responses. They were not required to answer if they preferred to skip a prompt. This procedure minimized disruption to the simulation and ensured that all participants followed the same protocol. The minimal duration of the verbal response, together with the three-minute aggregation window used for feature extraction, further reduces the potential influence of these brief interruptions on the physiological measurements. We used a camera and a frame grabbing system to record the ATCOs and their working environment including the screen content as video files. The screen setup consisted of a large main screen for radar and two smaller (secondary) screens on the right for flight plan information and on the left for procedural and weather information. The setup utilized the original hardware and software in a hardware-in-the-loop configuration, and therefore appeared to the ATCO as if a true operational environment was being used. The eye-tracking device was attached to the bottom of the primary screen (see Figure 2).

3.6 Scenario Design Verification

The light, moderate, and heavy task-load scenarios were derived from pre-existing simulator exercises originally developed for training purposes at LFV. These exercises were designed to manipulate ATCO task load by varying factors such as the number of aircraft, traffic density, and event complexity. Scenario selection was guided by expert judgment to ensure that each exercise represented a distinct and realistic level of operational complexity and task demand.

To ensure that the scenarios were appropriate for workload variation, a dedicated scoring system was developed in collaboration with an operational expert. This system was adapted from the scoring framework introduced by [12]. The scoring approach provides a quantitative assessment of the traffic complexity encountered by ATCOs during simulation, applying an event-based scheme to assign points according to scenario-specific events and associated ATCO tasks. Table 1 summarizes the simulator events, the corresponding ATCO tasks, and their respective point values.

An initial validation of the methodology was performed using data from a single ATCO's simulation runs. The resulting scores were 142, 221, and 538 for the light, moderate, and heavy task-load scenarios, respectively. These results indicate a consistent and logical progression in task complexity across the scenarios. Accordingly, we conclude that the scenarios are clearly differentiated in terms of task load.

Table 1. ATCO tasks and corresponding scores

Event	Task	Base Score
A/c entering the sector, a/c crossing the sector	Acceptance + monitoring	2
A/c leaving the sector	Transfer	1
Conflict type 1: cruise-cruise, conflict due to overtaking	Determining conflict type 1 + taking decision how to solve + monitoring	5
Conflict type 2: cruise-climb, cruise-descent, climb-descent	Determining conflict type 2 + taking decision how to solve + monitoring	7
A/c landing, a/c climbing, conflict	Instruction to pilot (clearance/vectoring/level change/speed control)	4

4. Machine Learning Approach

In this section, we describe the machine learning techniques applied to classify ATCOs' workload based on the collected data. We detail the feature extraction process, model development, and evaluation methods, followed by the presentation of the corresponding results.

4.1 Data Preprocessing

The recorded dataset consisted of time-series recordings of eye-tracking and head-movement variables. These measurements capture both discrete events (e.g., saccades, fixations, blinks) and continuous signals (e.g., pupil diameter, blink dynamics, head rotations), thereby providing a detailed representation of visual and motor activity over time. Table 2 summarizes the variables included in the dataset, their types, and descriptions.

Table 2. Overview of eye-tracking and head-movement variables.

Variable	Type	Description
Saccade	Integer	Non-zero integer recorded for the duration of each saccade, as detected by the saccade filter; increments with each detected saccade.
Fixation	Integer	Non-zero integer recorded for the duration of each fixation, as detected by the fixation filter; increments with each detected fixation.
Blink	Integer	Non-zero integer recorded for the duration of each blink, as detected by the blink filter; increments with each detected blink.
Pupil diameter	Continuous	Diameter of the pupil, measured separately for left and right eyes.
Blink closing amplitude	Continuous	Difference in eyelid opening between the state before closing and the fully closed state, measured separately for each eye.
Blink opening amplitude	Continuous	Difference in eyelid opening between the fully closed state and the state immediately after opening, measured separately for each eye.
Blink closing speed	Continuous	Maximum speed of the eyelid during closing movement, measured separately for each eye.
Blink opening speed	Continuous	Maximum speed of the eyelid during opening movement, measured separately for each eye.
Head heading	Continuous	Left/right rotation of the head ("no" rotation).
Head pitch	Continuous	Up/down rotation of the head ("yes" rotation).
Head roll	Continuous	Tilt rotation of the head ("maybe" rotation).

A comprehensive data preprocessing was undertaken to ensure the quality and reliability of the dataset, which is crucial for accurate analysis. Given the complexity and sensitivity of eye-tracking and head-movement data, special attention was paid to identifying and addressing missing or physiologically unrealistic values. For pupil diameter data, physiologically unrealistic values were identified and marked as missing. Following the recommendations of Spector [38], a valid range of 2 to 8 mm was applied to filter out invalid measurements. Missing values in pupil diameter were then imputed using linear interpolation, which estimates missing data points based on adjacent valid values. A similar linear interpolation method was applied to handle missing values in other continuous features, such as blink-amplitude, blink-speed, and head-movement variables. For event-based data like saccades, fixations, and blinks, missing values were initially filled using forward propagation of the last valid observation. Any remaining gaps were then addressed using backward propagation with the next valid observation. This two-step approach ensured the completeness of the data, providing a robust foundation for subsequent analysis.

4.2 Features

We segment all data into non-overlapping time slots of three minutes. Accounting for eye-tracking and CHS data, this results in 667 slots. The three-minute interval is chosen to align with the measurement interval of the CHS scores, ensuring consistency between workload assessment and feature extraction. To derive meaningful features, we transform saccade, fixation, and blink data into duration-based metrics by calculating three descriptive statistics—mean, standard deviation, and median—of their respective durations within three-minute time intervals. Additionally, we compute the total number of saccades and blinks for each interval. For all other continuous variables, we calculate the same three descriptive statistics—mean, standard deviation, and median—within the same time intervals, yielding 50 derived features.

To refine the feature set, we exclude features with zero standard deviation, as they exhibit no variation across data points. Specifically, the median values for right and left blink opening and closing speeds, as well as amplitudes, are removed to eliminate redundant information and reduce potential noise. This process results in 42 features, which constitute the initial feature set used for training and evaluating the machine learning models.

4.3 Labels and Classification Tasks

We perform both binary and multi-class classification tasks, using the CHS scores as the labels for both. In the three-class classification, the objective is to distinguish between low, medium, and high workload levels. As detailed in Section 3.3, these workload levels are redefined based on the analysis of ATCO responses, with low workload corresponding to CHS score = 1, medium workload corresponding to CHS scores $\in \{2, 3\}$, and high workload corresponding to CHS scores ≥ 4. In the binary classification task, the goal is to distinguish high workload from all other levels, with the classes defined as low (CHS score < 4) and high (CHS score ≥ 4).

It is important to note that the scenario design (light, moderate, heavy) reflects the intended overall task load, whereas the labeling of each data interval is based on the ATCOs' CHS ratings. Consequently, a simulation run in a high task-load scenario may still contain intervals labeled as medium or even low workload, depending on the subjective assessment at that time.

To validate the consistency of low, medium, and high workload labels with the scenario scheme and assess their reliability, we analyze the distribution of these labels across each scenario, as illustrated in Table 3. In the light-workload scenario, nearly all labels fall under the low-workload category, with 100% classified as low in the binary scheme and 91.67% as low in the three-class scheme. The moderate-workload scenario exhibits a more balanced distribution, with the binary scheme predominantly labeling it as low (99.55%), while the three-class scheme reveals a substantial presence of medium workload labels (48.19%), alongside a slightly greater proportion of low labels (51.36%) and a negligible proportion of high labels (0.45%). In the heavy-workload scenario, the distribution shifts significantly, with 19.92% classified as high in both schemes. The three-class scheme further highlights a majority of medium labels (63.13%) and a smaller proportion of low labels (16.95%).

These results confirm that the workload labels align well with the expected task-load variations for each scenario. However, the dataset is imbalanced, with low-workload labels dominating in the light and moderate scenarios and high-workload labels being relatively scarce overall. This corresponds to a ratio of about 4:1 in the heavy scenario for high against other labels. Despite this imbalance, the binary and three-class schemes both capture the intended workload progression, and the three-class scheme provides a more nuanced representation of medium workload levels in the moderate and heavy scenarios. This consistency demonstrates the reliability and robustness of the CHS-based labeling process, making it a strong foundation for subsequent machine learning analysis, where appropriate techniques should be applied to address potential class imbalances.

Table 3. Distribution of low/medium/high workload across the scenarios

Scenario	Workload, binary		Workload, three-class		
	Low	High	Low	Medium	High
Light	100.00	0.00	91.67	8.33	0.00
Moderate	99.55	0.45	51.36	48.19	0.45
Heavy	80.08	19.92	16.95	63.13	19.92

We conducted the classification on a combined dataset including all participants. The models were trained on shuffled time slots pooled from all participants and then evaluated on other time slots from the same individuals. This setup corresponds to a subject-dependent cross-validation strategy, where training and test sets may contain data from the same participant.

4.4 Machine Learning Models

We apply eight classical machine learning models to evaluate their effectiveness in estimating ATCO workload levels: decision tree, AdaBoost classifier, bagging classifier, random forest, histogram-based gradient boosting classification tree, extra trees classifier, support vector classifier (SVC), and k-nearest neighbors classifier (kNN). The classification algorithms are implemented using the Python programming language and the scikit-learn library.

These models were selected to provide a comprehensive evaluation of different machine learning approaches, balancing interpretability, computational efficiency, and the ability to capture complex patterns. Given the non-linear and high-dimensional nature of the eye-tracking features, we anticipated that tree-based models and ensemble methods (decision tree, random forest, gradient boosting, ExtraTrees, AdaBoost, and bagging) might offer advantages, as they are designed to capture complex interactions among variables and have frequently demonstrated effectiveness on tabular data [39, 40]. A support vector classifier was chosen for its ability to handle moderately high-dimensional feature spaces, such as our dataset with 42 features, and in light of reports of good performance in certain studies, including eye-tracking-based workload tasks [40, 41]. Additionally, the k-nearest neighbors model was included for its non-parametric flexibility, which allows it to capture complex, non-linear relationships in the data without assuming any specific distribution, with some studies reporting favorable performance [40, 42]. By testing this diverse set of models, we aimed to identify not only the best-performing model but also to understand how different model architectures handle the challenges of ATCO workload classification, such as inter-individual variability, feature complexity, and potential non-linear relationships in the data.

4.5 Training and Validation Techniques

We employ a stratified k-fold strategy ($k = 10$) for model evaluation ("outer" cross-validation). This approach divides the dataset into 10 equally sized subsets (folds), maintaining the original class distribution in each fold. In each iteration, one fold is used as the test set, and the remaining nine are used for training, ensuring all data is utilized for testing once. The stratified split ensures proportional representation of the smaller high-workload class in both training and test sets, mitigating class imbalance and enabling a fair model evaluation.

To optimize the hyperparameters of each model, we use randomized search combined with stratified k-fold cross-validation ($k = 10$) to identify the parameter combinations that maximize overall model performance ("inner" cross-validation). Specifically, we randomly sample $n_{\text{iter}} = 100$ hyperparameter combinations from the specified hyperparameter space. To enhance efficiency, we begin with broad parameter ranges to capture general trends and then iteratively refine the search space around the most promising configurations. Each sampled hyperparameter set is evaluated by splitting the training dataset into 10 non-overlapping folds, ensuring class proportions are preserved. One fold serves as the validation set, while the model is trained on the remaining nine. This process is repeated until each fold has been used for validation once. The average *F1 score* across all folds is computed to assess hyperparameter performance. Once the optimal hyperparameters are identified, the model is retrained on the entire training dataset using these parameters to achieve the best possible performance. Details about the models and their hyperparameters can be found, for instance, in [43, 44].

To address the class imbalance between the majority (low-medium workload) and minority (high workload) classes, we apply class weights (where supported by the models) that are inversely proportional to the class frequencies in the training data. This approach assigns greater weight to the underrepresented high-workload class, encouraging the model to prioritize learning from these critical, yet less frequent data points.

When features have different scales, their contributions to the model fitting process and the resulting model function can become unequal. Features with larger scales may dominate, overshadowing the influence of features with smaller scales, which can lead to biased learning. To address this issue, we normalize each feature independently before the model fitting stage, ensuring all features contribute equally. To prevent data leakage, normalization is performed separately for the training and test sets, using the minimum and maximum values derived exclusively from the training set.

4.6 Model Evaluation Metrics

Given the inherent class imbalance within our ATCO workload prediction dataset, *accuracy* might not be an informative metric. To address this potential limitation, we incorporate the *precision*, *recall*, and *F1 score* metrics.

Precision measures the proportion of true positive predictions among all positive predictions made by the model, indicating its ability to correctly identify relevant instances (What fraction of the positive predictions turned out to be correct?). *Recall*, on the other hand, measures the proportion of true positives identified out of all actual positive cases, reflecting the model's ability to avoid missing relevant instances (What fraction of the actual positive cases was correctly identified?). The *F1 score* combines both *precision* and *recall* into a single metric by calculating their harmonic mean, offering a balanced assessment when there is a trade-off between the two.

All three metrics are macro-averaged, meaning they are computed as the arithmetic mean (aka unweighted mean) of the per-class scores. This approach treats all classes equally, regardless of the number of samples, ensuring that under-represented classes contribute equally to the overall evaluation. By using macro-averaging for *precision*, *recall*, and *F1 score*, we penalize the model for poor performance on under-represented classes, making this approach better suited to handle class imbalance. A high *macro F1 score* indicates balanced performance across all workload categories, even when they are not equally represented in the data. All performance metrics are reported to three decimal places to ensure precision and consistency.

4.7 Evaluation and Selection of Models

Table 4 presents the performance of the eight selected machine learning models on the unseen test set, averaged across cross-validation folds, using the initial set of 42 features. The table reports the three key metrics—*precision*, *recall*, and *F1 score*—for both binary and three-class classification tasks. The highest *F1 score* for each classification task is highlighted in bold. Among the models, decision tree showed the lowest performance, while SVC and kNN achieved the highest *F1 scores*. These two top-performing models are selected for further analysis.

Table 4. Models' performance metrics in classifying three and two levels of ATCO workload using 42 features

Model	Binary			Three-class		
	Precision	*Recall*	*F1 score*	*Precision*	*Recall*	*F1 score*
Decision tree	0.646	0.656	0.635	0.619	0.590	0.593
AdaBoost classifier	0.805	0.607	0.645	0.821	0.688	0.718
Bagging classifier	0.694	0.628	0.647	0.757	0.648	0.671
Random forest	0.765	0.726	0.733	0.710	0.665	0.678
Histogram-based gradient boosting	0.814	0.688	0.724	0.793	0.703	0.729
Extra trees classifier	0.783	0.761	0.758	0.732	0.727	0.727
Support vector classifier	0.841	0.800	0.808	0.783	0.753	0.757
k-Nearest neighbors classifier	0.903	0.821	**0.849**	0.822	0.765	**0.778**

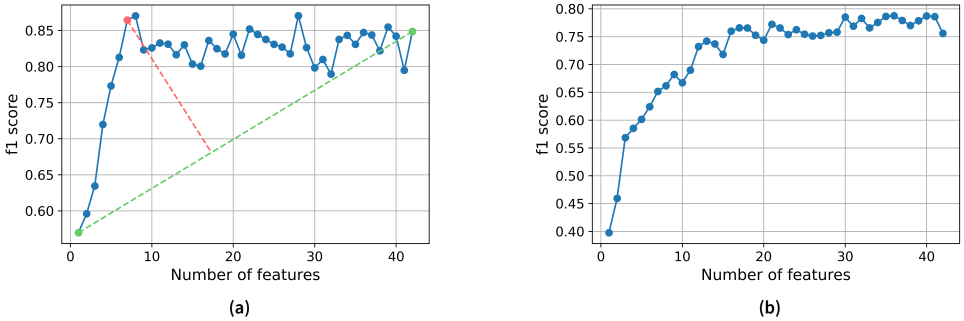


Figure 3. *F1 score* vs. number of features for the binary (a) and three-class (b) classification tasks (kNN model). (a) illustrates the "best economy" criterion: the red dashed line represents the distance from the "knee point" to the green dashed line, which connects the first and last points of the curve. Scaling factors were applied to accurately represent perpendicular distance, accounting for the differing x and y axis ranges.

5. Feature Selection

5.1 Recursive Feature Elimination

We apply **RFE** with permutation importance to iteratively remove the least informative features based on their impact on model performance. RFE begins with the full feature set, training a model and assessing feature importance using a method such as permutation importance. The feature contributing least to model performance (measured by *F1 score* in our case) is removed. This process repeats until a specified number of features remains or a predefined stopping criterion is met.

Permutation importance is a model-agnostic technique used to assess the significance of individual features in a predictive model. It works by randomly shuffling the values of a specific feature in the dataset while keeping all other features unchanged. The model's performance is then re-evaluated on this modified dataset, and the difference in performance compared to the original data is measured. A larger drop in performance suggests that the model relied heavily on that feature for making accurate predictions, indicating its higher importance. Conversely, if the model's performance remains largely unaffected, the feature is likely less informative or redundant.

In our RFE process, we apply stratified k-fold cross-validation ($k = 10$) to ensure a fair evaluation of feature importance across different training and validation splits. For each fold, we perform 20 shuffles per feature, measuring the corresponding impact on model performance. We then compute the average feature importance across all shuffles within each fold before taking the final average over all cross-validation splits. This approach provides a more stable and reliable estimate of each feature's contribution to the model.

We apply RFE separately to each of the two top-performing models, SVC and kNN (see 4.7), for both binary and three-class classification tasks. The process starts with 42 features and iteratively removes the least impactful ones until only a single feature remains. For each model and classification task, we then use the resulting ranked list of feature importances to reevaluate performance with the stratified k-fold cross-validation strategy, sequentially removing features one by one. Finally, we visualize the selection process for each combination of model and classification task by plotting the *F1 score* against the number of remaining features, illustrating the impact of feature reduction on performance. (See Figures 3a, 3b for kNN model examples).

We select the number of features based on two criteria: "best economy" and "best performance". The "best economy" criterion aims to achieve an optimal balance between model dimensionality (i.e., fewer features) and performance (i.e., higher *F1 score*). To identify this optimal feature subset, we locate the "knee point" on a performance curve. The knee point is defined as the point where "relative cost to increase some tunable parameter is no longer worth the corresponding performance benefit" ([45]). We determine this point by finding the maximum perpendicular distance from a straight line connecting the first and last points of the curve. This approach does not impose any assumptions about the curve's shape, making it more robust compared to other, more complex methods. The "best performance" criterion selects the number of features that yields the highest *F1 score*. Table 5 presents the results for both binary and three-class classification tasks. Notably, a modest reduction in feature dimensionality can be achieved while preserving model performance across both tasks. The complete list of the most important features for the best-performing model (kNN) in both classification tasks is presented in Table 7.

Table 5. Models' number of features and corresponding macro *F1 score* in RFE process for "best economy" and "best performance" criteria using ocular and head movements features. The following abbreviation is used: number of features (NoF).

Task	Model	Best economy		Best performance	
		NoF	*F1 score*	NoF	*F1 score*
Binary	SVC	11	0.769	35	0.817
Binary	kNN	7	0.864	28	**0.870**
Three-class	SVC	9	0.763	12	0.781
Three-class	kNN	13	0.742	36	**0.788**

5.2 Ocular Features

We investigate whether models using only ocular features can achieve performance comparable to those incorporating head-movement data, or if head-movement features play a crucial role in model performance. We achieve feature selection through a two-step process. First, we remove all head-movement features, leaving a set of 33 solely ocular features. Subsequently, we replicate the feature selection procedure outlined in Section 5.1 on this reduced feature set.

The results obtained are presented in Table 6, and the full set of key features for the best-performing model (kNN) in both classification tasks is provided in Table 7. The results show that the performance of models using only ocular features is slightly lower than that of models trained on the full feature set. However, the difference is not substantial, suggesting that ocular features alone capture a significant portion of workload-related information.

Table 6. Models' number of features and corresponding macro *F1 score* in RFE process for "best economy" and "best performance" criteria using ocular features. The following abbreviation is used: number of features (NoF).

Task	Model	Best economy		Best performance	
		NoF	*F1 score*	NoF	*F1 score*
Binary	SVC	5	0.709	28	0.784
Binary	kNN	8	0.818	12	**0.832**
Three-class	SVC	5	0.695	20	0.740
Three-class	kNN	6	0.660	33	**0.748**

Table 7. The most important features for the best-performing model (kNN) in binary and three-class classification tasks, presented for both the full dataset and the ocular-only dataset. The following abbreviations are used: left (L), right (R), mean (M), standard deviation (SD), median (Med), and closing (Cl).

Full dataset		Ocular dataset	
Binary	Three-class	Binary	Three-class
R Pupil Diameter SD	Saccades Duration Med	L Blink Cl Speed M	R Blink Cl Amplitude SD
L Pupil Diameter Med	R Blink Cl Amplitude M	R Pupil Diameter SD	L Pupil Diameter Med
Head Heading SD	Fixation Duration SD	R Pupil Diameter M	Saccades Duration Med
Blinks Duration M	Head Roll SD	R Blink Cl Amplitude M	R Pupil Diameter SD
Head Roll M	Head Pitch SD	Fixation Duration SD	R Pupil Diameter M
R Pupil Diameter Med	Fixation Duration Med	Blinks Duration Med	Blinks Duration SD
Head Pitch M	R Pupil Diameter SD	Blinks Duration SD	
	Head Roll M	L Pupil Diameter Med	
	Head Heading SD		
	L Pupil Diameter SD		
	Head Pitch Med		
	L Pupil Diameter Med		
	Blinks Duration SD		

5.3 Left-Right Eye Features Aggregation

In this section, we analyze a dataset that includes only features that are statistical summaries of eye-tracking variables measured independently for the left and right eyes. Namely, it contains the mean, standard deviation (SD), and median of pupil diameter for both eyes, along with left and right blink features, including the mean and SD of closing and opening amplitude and speed. This results in a total of 22 features ($SET1$). As in the previous analysis, we apply the RFE technique, but here, our objective is to determine the maximum achievable model performance given this specific feature set. By using this dataset, we investigate how different feature representations–maintaining separate left and right features, aggregating them, or applying additional transformations–affect model performance.

To investigate the influence of left- and right-specific features on model performance, we transform these features into aggregated metrics: their average, representing central tendency, and their absolute difference, capturing asymmetry ($SET2$). This transformation does not change the total number of features compared to using left and right features separately; instead, it reorganizes the information to distinctly separate symmetry (average) from asymmetry (absolute difference). Additionally, we evaluate the effect of augmenting the original left and right features with the absolute difference ($SET3$), which increases the total number of features to 33. The three configurations are summarized in Table 8.

Table 8. Left-Right Eye Features Aggregation Sets. The following abbreviation is used: number of features (NoF).

Name	NoF	Explanation
$SET1$	22	Original left/right features.
$SET2$	22	Original left/right features are transformed into their average and absolute difference.
$SET3$	33	Original left/right features are augmented with their absolute difference.

Scenario-based results (Table 9) show that using the original left and right eye features separately ($SET1$) generally provides the best performance, particularly for kNN in binary classification. Transforming these features into their average and absolute difference ($SET2$) leads to a decline in performance for most models, except for SVC in three-class classification, where it remains nearly unchanged. This suggests that preserving individual left and right eye features retains valuable asymmetry-related information that is lost when averaging. Augmenting the original features with their absolute difference ($SET3$) results in some improvements, particularly for SVC in both binary and three-class classification. However, these gains are not consistent across all cases. Overall, retaining the original left and right eye features separately ($SET1$) remains the most reliable approach, though $SET3$ might offer slight benefits in certain models.

Table 9. Maximum macro *F1 scores* ("best performance" criterion) obtained during the RFE process for $SET1$, $SET2$ and $SET3$

Features	kNN		SVC	
	Binary	Three-class	Binary	Three-class
$SET1$	0.819	0.713	0.721	0.668
$SET2$	0.787	0.669	0.715	0.669
$SET3$	0.799	0.724	0.768	0.719

6. Discussions

We aimed to explore whether non-intrusive physiological measures—specifically, eye-tracking data—can be used to predict ATCOs' workload using machine learning. We collected eye-tracking data from licensed controllers in realistic simulator scenarios and applied machine learning models to classify workload levels derived from an adapted CHS. Based on these analyses, we now discuss the main findings, their implications, and their relevance for workload assessment in air traffic control.

Among all models tested, the kNN classifier achieved the highest performance, suggesting that approaches based on local decision boundaries may be particularly effective for workload prediction from eye-tracking data. By classifying data points based on the majority class of their nearest neighbors, kNN captures patterns through proximity rather than relying on global decision surfaces. Its strong performance with a relatively small number of features indicates that preserving local relationships while limiting model complexity is beneficial in this setting. Increasing feature dimensionality may reduce kNN effectiveness due to the curse of dimensionality, where data points become equidistant, highlighting the importance of careful feature selection. In contrast, tree-based models rely on hierarchical partitioning of the feature space, which can offer greater flexibility but also increases susceptibility to overfitting when data are limited. This likely explains their lower performance in the present study, given the modest dataset size.

In the context of classifying ATCOs' workload using eye-tracking data, our *F1 scores* align with those reported in comparable cognitive workload studies based on physiological signals. Notably, similar performance has been observed in controlled laboratory settings, where participants' heads were stabilized with a chin and head rest and the focus was on general cognitive workload rather than complex, real-world ATCO tasks [15], with reported *F1 scores* typically ranging from 0.700 to 0.850.

In [16], cognitive workload classification during the Digit Symbol Substitution Test achieved a high *F1 score* (0.950). However, their study assessed "task load" based on objective task difficulty, rather than actual cognitive workload. Additionally, the controlled laboratory setting with standard fluorescent lighting and blocking outside light, standardized task, homogeneous participants, and balanced dataset contributed to the high performance. In contrast, our study examines real-world air traffic control, where workload is dynamic and influenced by individual differences. The use of subjective workload measures, such as CHS self-reports, further increases classification complexity. Despite these challenges, our model demonstrates strong performance, underscoring the potential of eye-tracking metrics in operational environments.

The relative importance of *precision* and *recall* depends on operational needs. Although we report macro-averaged metrics, class-specific trade-offs matter: high *precision* in the low-workload class reduces unnecessary interventions, whereas high *recall* in the high-workload class helps avoid missing critical overload situations. Thus, while macro metrics summarize overall performance, practical deployment would require prioritizing *precision* or *recall* per class.

The analysis of feature importance across different tasks (binary/three-class) and datasets (full/ocular) reveals several key insights into the physiological indicators most relevant to workload classification. Notably, pupil-diameter variability and blink metrics emerged as robust indicators, consistently appearing among the top features within the kNN model (Table 7). Specifically, the standard deviation of the right pupil diameter (R Pupil Diameter SD) and the median of the left pupil diameter (L Pupil Diameter Med) were identified as critical features in every dataset and task, together highlighting the strong correlation between bilateral pupil dynamics and cognitive workload. The fact that some features appear only on one side could be due to natural asymmetries in eye physiology or variations in measurement accuracy between the left and right eye. Blink-related features, including Blinks Duration and Blink Closing Amplitude/Speed (Mean, Median, SD), were also prominent, particularly in ocular-specific datasets, aligning with existing literature that associates blink characteristics with cognitive workload.

Head movement metrics, encompassing Head Heading, Roll, and Pitch (Mean/SD), played a substantial role in the full dataset, especially for the three-class classification tasks. This suggests that head movements provide valuable complementary information to ocular data, potentially reflecting physical signs of cognitive workload, like changes in posture or restlessness. The increased presence of both ocular and head movement features in the more complex three-class tasks, compared to the binary tasks, suggests that distinguishing between multiple workload levels benefits from incorporating a wider range of physiological data. While ocular features remain dominant, head dynamics capture additional subtle behavioral cues, such as postural adjustments, that enhance the model's ability to differentiate between nuanced workload levels.

The presence of head-movement features among the most important predictors in the full dataset indicates that they contribute valuable information. Their relevance suggests a potential link between physical movement and cognitive workload, likely reflecting how controllers adjust their visual scanning behavior in response to increasing task demands. The fact that removing head-movement features results in only a minor performance drop further implies some degree of redundancy between ocular and head-movement data. This overlap may indicate that aspects of head movement, such as scanning behavior, are already partially captured through eye-tracking features (e.g., saccade or fixation duration, see Table 7). The performance under the "best performance" criterion for the ocular dataset is approximately equivalent to the "best economy" criterion in the full dataset. That is, achieving a similar performance to the one we obtain for the full data set in the "best-economy" criterion (e.g., *F1 score* = 0.742 for 13 features for kNN in classification into three classes) necessitates significantly more features in the ocular data set (*F1 score* = 0.748 for 33 features for the same model and number of classes).

In the ocular dataset, the reliance on blink and pupil features was particularly pronounced, as expected due to the absence of head movement data. However, the inclusion of features like Fixation Duration SD and Saccades Duration Med indicates that a variety of eye movement behaviors contribute significantly to workload classification. As previously discussed, while the full dataset achieves a high *F1 score* using only the "best economy" feature set, the ocular dataset requires incorporating a larger set of features beyond this subset to attain comparable performance. This suggests that while ocular metrics alone are valuable, their predictive power is enhanced when a broader range of features is considered, compensating for the lack of head movement data.

An overarching trend across all datasets and tasks was the frequent identification of standard deviation metrics (e.g., Right Pupil Diameter SD, Fixation Duration SD, Head Heading SD) as top features. This suggests that variability in physiological responses is often more informative of cognitive workload than simple averages. Increased workload may lead to greater variability in eye movements and head movement behavior, potentially due to fluctuating attention or stress responses. This insight underscores the importance of considering dynamic aspects of physiological data in workload assessment models.

Limitations

First, the "reclassification" of the CHS scale—where ratings of ≥ 4 were classified as high workload—may influence how well the labels capture the underlying construct. Second, while the number of participating ATCOs is relatively large for this type of simulation study, the resulting dataset remains limited in size from a machine learning perspective, which may affect model generalizability. Third, the study was conducted within a single working environment, which may constrain the transferability of findings to other operational contexts. Fourth, subjective workload assessments are susceptible to social desirability bias, which can lead participants to under- or overstate their perceived workload. Fifth, the use of three-minute time slots for both feature extraction and CHS labeling assumes temporal stability of workload within these intervals. This temporal aggregation may smooth short-term workload fluctuations and reduce the model's sensitivity to rapid changes in cognitive state. Finally, workload itself is a soft, multidimensional concept, and while the operationalization used here enables predictive modeling, it cannot fully capture the richness of the construct.

7. Conclusion and Outlook

This study demonstrated the strong potential of machine learning techniques for predicting ATCOs workload using eye-tracking and head-movement data. The results support the feasibility of developing non-intrusive, data-driven workload monitoring systems to enhance operational safety and efficiency.

Compliance with Study Objectives

1. **Demonstration of machine learning effectiveness:** We confirmed that classical machine learning models, particularly the kNN classifier, can accurately predict ATCO workload from eye-tracking and head-movement features.
2. **Feature reduction analysis:** RFE showed that effective workload prediction can be achieved with a substantially smaller subset of features, enhancing model efficiency without compromising performance. The analysis further revealed that pupil diameter and blink metrics are strong general indicators of workload. Moreover, the prominence of variability-based metrics suggests that cognitive workload may manifest through fluctuations in physiological responses rather than through static changes.
3. **Evaluation of left–right eye metrics and head movements:** Analyses revealed that maintaining separate left and right eye features preserves critical asymmetry-related information, while head-movement metrics provide complementary predictive value.

Feature selection analysis indicated that the “curse of dimensionality” was not a limiting factor within the studied feature space. Increasing the number of features did not degrade model performance, although a “saturation point” was identified beyond which additional features contributed little or introduced noise.

By combining subjective assessments with objective, non-intrusive physiological measures such as eye-tracking data, this study mitigates the limitations of self-reported workload ratings, including bias and intrusiveness. The proposed machine learning approach thus offers a complementary and scalable method for real-time workload estimation, enhancing reliability beyond subjective measures alone.

In the present study, we employ a subject-dependent strategy, where models are trained and evaluated on shuffled data from all participants. Future work could explore a subject-independent approach, in which each participant’s data is reserved exclusively for either the training or the test set. This would provide a more rigorous evaluation and shed light on the models’ ability to generalize to new users. In addition, a subject-specific approach warrants investigation, where models are trained and evaluated using data from a single participant. Such an approach would allow us to examine whether individualized models outperform multi-subject models in predicting workload.

We also restricted our analysis to simple statistical descriptors of the collected variables (mean, median, standard deviation). While this approach provides a systematic baseline, more advanced feature engineering—including domain-specific metrics (e.g., saccade amplitude, gaze velocity), frequency-domain analysis or sequence models such as Long Short-Term Memory models—represent promising directions for future research.

Another promising direction is to explore alternative feature selection methods. In particular, applying Lasso regression for feature selection and comparing its performance with the RFE-based approach used in this study could provide further insights into the robustness of feature selection strategies for workload prediction.

Acknowledgement

We would like to acknowledge all involved persons of LFV en-route at air traffic control center Malmö, in particular ATCOs and pseudo-pilots involved in the experimental data collection. The study received ethical approval from the Swedish Ethical Review Authority (2023-06110-01).

Author contributions

- Anastasia Lemetti: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Writing - Original Draft, Writing - Review & Editing
- Lothar Meyer: Conceptualization, Funding Acquisition, Project Administration, Supervision, Writing - Review & Editing
- Maximilian Peukert: Conceptualization, Funding Acquisition, Supervision, Writing - Review & Editing
- Tatiana Polishchuk: Conceptualization, Funding Acquisition, Project Administration, Supervision, Writing - Review & Editing
- Christiane Schmidt: Conceptualization, Funding Acquisition, Supervision, Writing - Original Draft, Writing - Review & Editing
- Helene Alpfjord Wylde: Data Curation, Writing - Review & Editing

Funding statement

This study was supported in the scope of project On WorkLoad Measures (OWL), funded by the Swedish Transport Administration (TRV 2022/33636r).

Open data statement

The dataset utilized for the machine learning models in this study is publicly available on the Open Science Framework at https://doi.org/10.17605/OSF.IO/T93D2. It comprises two CSV files: one containing extracted features and one providing the corresponding CHS scores.

Reproducibility statement

The source code for this study is accessible in a public GitHub repository at https://github.com/anlemett/OWL_JOAS. Detailed instructions for reproducing the results are provided in the README file.

References

[1] Sandra G Hart and Lowell E Staveland. “Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research”. In: *Human mental workload*. Ed. by P. A. Hancock and N. Meshkati. Elsevier, 1988, pp. 139–183.

[2] European Aviation Safety Agency. *Commission Regulation (EU) 2015/340*. Accessed: 2025-02-17. 2015. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02015R0340-20240804.

[3] Robert M. Yerkes and John D. Dodson. “The relation of strength of stimulus to rapidity of habit-formation”. In: *Journal of Comparative Neurology and Psychology* 18.5 (1908), pp. 459–482. DOI: 10.1002/cne.920180503. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cne.920180503. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cne.920180503.

[4] José Juan Cañas, Pedro Ferreira, Patricia López de Frutos, Eva Puntero, Elena López, Fernando Gómez-Comendador, Francesca De Crescenzio, Francesca Lucchi, Fedja Netjasov, and Bojana Mirkovic. "Mental workload in the explanation of automation effects on ATC performance". In: *Human Mental Workload: Models and Applications: Second International Symposium, H-WORKLOAD 2018, Amsterdam, The Netherlands, September 20-21, 2018, Revised Selected Papers 2*. Springer. 2019, pp. 202–221.

[5] Kyle Kent Edward Ellis. "Eye tracking metrics for workload estimation in flight deck operations". Master's thesis. University of Iowa, 2009. DOI: 10.17077/etd.a773626l.

[6] Giovanni Pignoni and Sashidharan Komandur. "Development of a Quantitative Evaluation Tool of Cognitive Workload in Field Studies Through Eye Tracking". In: *Engineering Psychology and Cognitive Ergonomics - 16th International Conference, EPCE 2019*. Ed. by Don Harris. Vol. 11571. Lecture Notes in Computer Science. Springer, 2019, pp. 106–122. DOI: 10.1007/978-3-030-22507-0_9. URL: https://doi.org/10.1007/978-3-030-22507-0%5C_9.

[7] Maik Friedrich, Anneke Hamann, and Jörn Jakobi. "An Eye Catcher in the ATC Domain: Influence of Multiple Remote Tower Operations on Distribution of Eye Movements". In: *HCII*. Cham: Springer International Publishing, 2020, pp. 262–277. ISBN: 978-3-030-49183-3.

[8] Billy Josefsson, Lothar Meyer, Maximilian Peukert, Tatiana Polishchuk, and Christiane Schmidt. "Validation of controller workload predictors at conventional and remote towers". In: *9th International Conference on Research in Air Transportation*. 2020.

[9] Lothar Meyer, Maximilian Peukert, Tatiana Polishchuk, and Christiane Schmidt. "Investigating ocular and head-yaw measures as indicators for workload and fatigue under varying taskload conditions". In: *10th International Conference on Research in Air Transportation*. 2022.

[10] Anastasia Lemetti, Lothar Meyer, Maximilian Peukert, Tatiana Polishchuk, and Christiane Schmidt. "Discrete-Fourier-transform-based evaluation of physiological measures as workload indicators. A human-in-the-loop study linking workload and fatigue in a multi remote tower environment". In: *42nd Digital Avionics Systems Conference (DASC)*. 2023.

[11] Saeed Shadpour, Ambreen Shafqat, Serkan Toy, Zhe Jing, Kristopher Attwood, Zahra Moussavi, and Somayeh B. Shafiei. "Developing cognitive workload and performance evaluation models using functional brain network analysis". In: *npj Aging* 9.1 (Oct. 6, 2023), p. 22. DOI: 10.1038/s41514-023-00119-z. URL: https://doi.org/10.1038/s41514-023-00119-z.

[12] María Zamarreño Suárez, Rosa Maria Arnaldo Valdes, Francisco Pérez Moreno, Raquel Delgado-Aguilera Jurado, Patricia María López de Frutos, and Victor Fernando Gomez Comendador. "How much workload is workload? A human neurophysiological and affective-cognitive performance measurement methodology for ATCOs". In: *Aircraft Engineering and Aerospace Technology* 94.9 (2022), pp. 1525–1536.

[13] Patricia López de Frutos, Rubén Rodríguez Rodríguez, Danlin Zheng Zhang, Shutao Zheng, José Juan Cañas, and Enrique Muñoz-de-Escalona. "COMETA: An air traffic controller's mental workload model for calculating and predicting demand and capacity balancing". In: *Human Mental Workload: Models and Applications: Third International Symposium, H-WORKLOAD 2019, Rome, Italy, November 14–15, 2019, Proceedings 3*. Springer. 2019, pp. 85–104.

[14] Jorge Ibáñez-Gijón, David Travieso, José A. Navia, Aitor Montes, David M. Jacobs, and Patricia L. Frutos. "Experimental validation of COMETA model of mental workload in air traffic control". In: *Journal of Air Transport Management* 108 (2023), p. 102378. ISSN: 0969-6997. DOI: 10.1016/j.jairtraman.2023.102378. URL: https://www.sciencedirect.com/science/article/pii/S0969699723000212.

[15] Emmanouil Ktistakis, Vasileios Skaramagkas, Dimitris Manousos, Nikolaos S. Tachos, Evanthia Tripoliti, Dimitrios I. Fotiadis, and Manolis Tsiknakis. "COLET: A dataset for COgnitive workLoad estimation based on Eye-Tracking". In: *Computer Methods and Programs in Biomedicine* 224 (2022), p. 106989.

[16] Monika Kaczorowska, Małgorzata Plechawska-Wójcik, and Mikhail Tokovarov. "Interpretable machine learning models for three-way classification of cognitive workload levels for eye-tracking features". In: *Brain sciences* 11.2 (2021), p. 210.

[17] Ana Carolina Russo, M. M. Cardoso Junior, and Emilia Villani. "Eye-tracking analysis to assess the mental load of unmanned aerial system operators: systematic review and future directions". In: *The Aeronautical Journal* (2024), pp. 1–30.

[18] Jesus L. Lobo, Javier Del Ser, Flavia De Simone, Roberta Presta, Simona Collina, and Zdenek Moravek. "Cognitive workload classification using eye-tracking and EEG data". In: *Proceedings of the International Conference on Human-Computer Interaction in Aerospace*. HCI-Aero '16. Paris, France: Association for Computing Machinery, 2016. ISBN: 9781450344067. DOI: 10.1145/2950112.2964585. URL: https://doi.org/10.1145/2950112.2964585.

[19] Murillo Pagnotta, David M Jacobs, Patricia L de Frutos, Ruben Rodríguez, Jorge Ibáñez-Gijón, and David Travieso. "Task difficulty and physiological measures of mental workload in air traffic control: a scoping review". In: *Ergonomics* 65.8 (2022), pp. 1095–1118. DOI: 10.1080/00140139.2021.2016998.

[20] Nicolina Sciaraffa, Pietro Aricò, Gianluca Borghini, Gianluca Di Flumeri, Antonio Di Florio, and Fabio Babiloni. "On the use of machine learning for EEG-based workload assessment: Algorithms comparison in a realistic task". In: *Human Mental Workload: Models and Applications*. Ed. by Luca Longo and Maria Chiara Leva. Cham: Springer International Publishing, 2019, pp. 170–185.

[21] Anastasia Lemetti, Lothar Meyer, Maximilian Peukert, Tatiana Polishchuk, Christiane Schmidt, and Helene Alpfjord Wylde. "Predicting air traffic controller workload using machine learning with a reduced set of eye-tracking features". In: *35th European Association for Aviation Psychology Conference (EAAP)*. 2024.

[22] Anastasia Lemetti, Lothar Meyer, Maximilian Peukert, Tatiana Polishchuk, Christiane Schmidt, and Helene Alpfjord Wylde. "Eye in the sky: Predicting air traffic controller workload through eye tracking based machine learning". In: *43rd Digital Avionics Systems Conference (DASC)*. 2024.

[23] Gordon Hughes. "On the mean accuracy of statistical pattern recognizers". In: *IEEE Transactions on Information Theory* 14.1 (1968), pp. 55–63. DOI: 10.1109/TIT.1968.1054102.

[24] Rebecca L. Charles and Jim Nixon. "Measuring mental workload using physiological measures: A systematic review". In: *Applied Ergonomics* 74 (2019), pp. 221–232. ISSN: 0003-6870. DOI: https://doi.org/10.1016/j.apergo.2018.08.028. URL: https://www.sciencedirect.com/science/article/pii/S0003687018303430.

[25] Debashis Das Chakladar and Partha Pratim Roy. "Cognitive workload estimation using physiological measures: a review". In: *Cognitive Neurodynamics* 18.4 (2024), pp. 1445–1465. ISSN: 1871-4099. DOI: 10.1007/s11571-023-10051-3. URL: https://doi.org/10.1007/s11571-023-10051-3.

[26] Yuval Zak, Yisrael Parmet, and Tal Oron-Gilad. "Subjective Workload Assessment Technique (SWAT) in real time: Affordable methodology to continuously assess human operators' workload". In: *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 2020, pp. 2687–2694. DOI: 10.1109/SMC42975.2020.9283168.

[27] Robert Bixler and Sidney D'Mello. "Automatic gaze-based detection of mind wandering with metacognitive awareness". In: *User Modeling, Adaptation and Personalization: 23rd International Conference, UMAP 2015, Dublin, Ireland, June 29–July 3, 2015. Proceedings 23*. Springer. 2015, pp. 31–43.

[28] Lisa-Marie Vortmann, Jannes Knychalla, Sonja Annerer-Walcher, Mathias Benedek, and Felix Putze. "Imaging time series of eye tracking data to classify attentional states". In: *Frontiers in Neuroscience* 15 (2021), p. 664490.

[29] Shane D Sims and Cristina Conati. "A neural architecture for detecting user confusion in eye-tracking data". In: *Proceedings of the 2020 international conference on multimodal interaction.* 2020, pp. 15–23.

[30] Gano B Chatterji and Banavar Sridhar. "Neural network based air traffic controller workload prediction". In: *Proceedings of the 1999 American Control Conference (Cat. No. 99CH36251).* Vol. 4. 1999, 2620–2624 vol.4. DOI: 10.1109/ACC.1999.786543.

[31] David Gianazza. "Learning air traffic controller workload from past sector operations". In: *12th USA/Europe Air Traffic Management R&D Seminar (ATM Seminar).* 2017.

[32] Smart Eye. *Smart Eye XO Eye Tracking System.* Accessed: 2025-02-17. URL: https://www.smarteye.se/xo/.

[33] A. Papenfuss and M. Peters. "HMI laboratory report 8: Analysis of critical situations at remote tower operated airports". Master's thesis. DLR, Institut für Flugführung, Braunschweig, 2012.

[34] C. S. Jordan and S. D. Brennan. *An experimental report on rating scale descriptor sets for the instantaneous self-assessment (ISA) recorder.* Tech. rep. DRA, 1992.

[35] American Psychological Association. *Social Desirability.* https://dictionary.apa.org/social-desirability. [Accessed: 2025-02-17].

[36] Stephen M Casner and Brian F Gore. "Measuring and evaluating workload: A primer". In: *NASA Technical Memorandum* 216395 (2010), p. 2010.

[37] Joshua Harrison, Kurtuluş İzzetoğlu, Hasan Ayaz, Ben Willems, Sehchang Hah, Ulf Ahlstrom, Hyun Woo, Patricia A Shewokis, Scott C Bunce, and Banu Onaral. "Cognitive workload and learning assessment during the implementation of a next-generation air traffic control technology using functional near-infrared spectroscopy". In: *IEEE Transactions on Human-Machine Systems* 44.4 (2014), pp. 429–440.

[38] Robert H. Spector. "The pupils". In: *Clinical methods: The history, physical, and laboratory examinations. 3rd edition* (1990).

[39] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. "Why do tree-based models still outperform deep learning on typical tabular data?" In: *Advances in neural information processing systems* 35 (2022), pp. 507–520.

[40] Assaf Shmuel, Oren Glickman, and Teddy Lazebnik. "A comprehensive benchmark of machine and deep learning across diverse tabular datasets". In: *arXiv preprint arXiv:2408.14817* (2024).

[41] Vasileios Skaramagkas, Emmanouil Ktistakis, Dimitris Manousos, Nikolaos S Tachos, Eleni Kazantzaki, Evanthia E Tripoliti, Dimitrios I Fotiadis, and Manolis Tsiknakis. "Cognitive workload level estimation based on eye tracking: A machine learning approach". In: *2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE).* IEEE. 2021, pp. 1–5.

[42] Weiya Chen, Tetsuo Sawaragi, and Toshihiro Hiraoka. "Comparing eye-tracking metrics of mental workload caused by NDRTs in semi-autonomous driving". In: *Transportation research part F: traffic psychology and behaviour* 89 (2022), pp. 109–128.

[43] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning: with applications in R.* Springer, 2017. ISBN: 978-1-461-47137-0.

[44] Wei-Meng Lee. *Python machine learning.* John Wiley & Sons, Inc., 2019. ISBN: 978-1-119-54563-7.

[45] Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. "Finding a "kneedle" in a haystack: Detecting knee points in system behavior". In: *2011 31st international conference on distributed computing systems workshops.* IEEE. 2011, pp. 166–171.