

UNCERTAINTY ANALYSIS AND EYE-TRACKING-BASED CLUSTERING IN AIR TRAFFIC CONTROLLER WORKLOAD PREDICTION

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

[1]Linköping University (LiU), Norrköping, Sweden
[2]Air Navigation Services of Sweden (LFV), Research & Innovation, Norrköping, Sweden

Abstract

Accurate assessment of air traffic controller (ATCO) workload is essential for maintaining safety and efficiency in modern air traffic control (ATC) systems. Although machine-learning models based on eye-tracking and behavioral data have shown promise for objective workload prediction, their performance is often affected by inter-individual differences in eye-tracking data and contextual uncertainty. In this study, we investigate how socio-demographic factors, temporal context, and inter-individual differences influence machine-learning–based workload prediction in an operationally realistic ATC simulation.

Eye-tracking and head-movement data were collected from n = 18 licensed ATCOs performing simulated control tasks under varying workload conditions. We evaluate workload classification using k-nearest neighbors (kNN) and support vector classification (SVC) models and analyze the effects of socio-demographic attributes (gender, age, and professional experience), temporal factors (time on task and time of day), and feature-based representations derived from descriptive statistics of eye-tracking and head-movement signals computed across time intervals.

We show that time on task yields modest improvements in classification performance, while higher professional experience is associated with substantially higher classification performance and lower professional experience corresponds to reduced performance. Clustering ATCOs based on descriptive-statistics representations of eye-tracking and head-movement features yields more homogeneous subgroups and leads to marked performance gains compared to a single global model.

Keywords: ATCO; Workload; Eye Tracking; Machine Learning

1. Introduction

Monitoring and managing ATCOs' workload is essential for maintaining safety and efficiency in modern air traffic management (ATM). Workload fluctuates in response to traffic complexity, task demands, automation characteristics, and individual cognitive capacity. When workload exceeds manageable levels, ATCOs may experience reduced situational awareness, delayed decision-making, and an increased likelihood of operational errors, all of which can directly affect system safety [1]. Conversely, sustained periods of very low workload may also be detrimental, leading to under-arousal and reduced vigilance. Accurate assessment of workload is therefore a central concern for both operational safety and system design.

Subjective rating scales remain the most widely used approach for assessing ATCO workload in practice. However, such self-report measures are inherently limited by recall bias, coarse temporal resolution, and susceptibility to social desirability effects, particularly in safety-critical professional contexts. Consequently, researchers have recently increasingly explored objective physiological and behavioral indicators to support continuous workload assessment. These indicators span multiple sensing modalities, including eye-tracking measures, electroencephalography (EEG), electrocardiography (ECG), and electrodermal activity (EDA), and have been investigated in different domains [2–17].

Eye tracking is particularly attractive in air traffic control (ATC) environments, as it can be deployed remotely without encumbering air traffic controllers (ATCOs) and provides direct insight into visual attention allocation and information processing strategies. Several groups of authors have demonstrated robust associations between eye-tracking metrics and mental workload across a range of domains, including aviation, human–computer interaction, and remote operations [2, 3, 13–19].

Machine learning approaches trained on such measurements have shown promising potential for predicting workload in realistic ATCO simulation environments [17, 20]. However, reported model performance varies substantially across studies and between individuals, indicating that the relationship between physiological signals and perceived workload is not uniform. This variability represents a significant challenge for the development of generalizable and reliable workload prediction models suitable for operational deployment.

One important but underexplored source of variability arises from socio-demographic factors, such as age, gender, and professional experience. ATCOs with different levels of experience may differ not only in their operational strategies but also in how they perceive, regulate, and report workload. In addition, temporal factors—including time on task and time of day—may systematically influence cognitive state and physiological responses, introducing further uncertainty into workload prediction. Beyond these factors, substantial inter-individual differences exist in baseline eye-tracking and head-movement behavior, as well as in how these signals vary over time. Such differences are often treated as noise in machine-learning pipelines, yet they may reflect stable behavioral traits that can be exploited to improve model performance. Grouping ATCOs into more homogeneous subpopulations based on their physiological characteristics may therefore reduce heterogeneity and yield more accurate workload prediction models.

The objective of this study is to investigate how socio-demographic characteristics, temporal influences, and inter-individual differences in physiological behavior affect the performance of machine-learning models that predict ATCO workload from eye-tracking and head-movement data collected in a real-time simulation environment. More specifically, we examine socio-demographic and temporal variables as potential uncertainty factors and analyze their impact on model performance. Furthermore, we explore whether grouping ATCOs based on similarities in their oculometric and head-movement behavior can reveal more homogeneous subgroups that support improved workload prediction.

2. Related Work

Eye tracking has been explored as an objective method for quantifying cognitive workload across a variety of task contexts. Ktistakis et al. [21] 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. They indicated that multitasking and time pressure significantly increase cognitive workload.

Kaczorowska et al. [6] 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, they assessed workload in terms of task load, defined by objective task difficulty, in a controlled environment.

Dell'Acqua et al. [13] proposed a framework that leverages smooth pursuit eye movements as a continuous and non-intrusive indicator of cognitive load. They introduced metrics that do not require prior knowledge of the visual stimulus path, making the approach independent of predefined smooth-pursuit trajectories. Their method combines Kalman-filter–based modeling and B-spline approximation with machine-learning classifiers, and is evaluated on controlled tasks with varying workload levels.

Trigka et al. [14] investigated eye-based cognitive overload prediction in human–machine interaction using supervised machine-learning models trained on eye-tracking features. The authors evaluated fixation, saccade, and pupil-related metrics extracted from a visually demanding interaction task and

compared the performance of multiple classifiers. They confirmed that eye-tracking-only features can reliably distinguish between normal and overload states, with ensemble and neural models achieving the highest performance. They also highlighted a trade-off between interpretability and predictive performance across classifiers.

Atweh and Riggs [15] examined eye-tracking-based prediction of mental demand in team-based operations using eye-tracking metrics and supervised machine-learning models. They analyzed eye-tracking data collected from multiple unmanned aerial vehicle command-and-control experiments and evaluated fixation, saccade, and pupil-related features using a range of classifiers, with kNN achieving the highest performance. They demonstrated that eye-tracking features can support reliable workload classification not only at the individual level but also in collaborative environments.

Within aviation human-factors research, eye tracking has been used to study cognitive workload and visual attention in pilots and, more recently, in remote drone operators. Lobo et al. [3] 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 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.

Kucukosmanoglu et al. [22] proposed a multimodal framework for classifying cognitive workload using synchronized eye-tracking, head-movement, and ECG data collected during simulated military missions. Using subject-independent machine-learning models, they reported that multimodal fusion outperforms unimodal approaches, with eye-tracking features—particularly pupil size and fixation dispersion—emerging as the most informative predictors of workload, while head-movement and cardiovascular signals provide complementary contributions. Their findings reinforced the central role of eye-tracking measures while suggesting that inter-individual variability may limit the performance of subject-independent models in heterogeneous operator groups.

Russo et al. [19] 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.

The application of eye tracking to investigate cognitive workload and visual attention in ATCOs is a relatively recent development. Materne and Friedrich [23] investigated the emerging role of supervisors in multiple remote tower centers by evaluating a dedicated planning tool using combined subjective workload ratings and eye-tracking measures. In a simulation study with 15 professional ATCOs, fixation duration and fixation frequency were analyzed across five representative supervisory use cases and compared with subjective workload ratings. The results showed that, compared with scheduled events, unscheduled events were associated with higher fixation frequencies, indicating increased visual search demands, while fixation duration exhibited a statistically significant correlation with subjective workload, supporting its interpretation as an indicator of information processing effort.

Navia et al. [17] showed that eye-tracking metrics reliably capture mental workload in simulated en-route ATC, with pupil size and blink behavior closely mirroring subjective ratings. They further demonstrated that gaze-action patterns differentiate error mechanisms in conflict management, with simple conflicts failing mainly at detection and complex ones at the planning stage.

An et al. [24] presented a neuromorphic approach to cognitive load classification in ATC by deploying spiking neural networks (SNNs) on dedicated low-power hardware. The authors combined EEG and eye-tracking features collected during ATC tasks and evaluated both conventional machine-learning models and SNNs trained using biologically inspired learning rules. Their results showed that a minimal single-layer SNN achieved competitive performance and retained robust accuracy af-

ter quantization and deployment on neuromorphic hardware. This work highlighted the feasibility of ultra-low-power, event-driven workload monitoring.

Across the existing literature, two key limitations emerge. First, many studies rely on controlled tasks or externally defined workload levels, which may not fully capture the subjective and context-dependent nature of workload in operational environments. Second, inter-individual differences are often treated as noise, despite growing evidence that variability in physiological and behavioral responses reflects stable individual traits. In the present study, we address these gaps by explicitly analyzing socio-demographic and temporal sources of uncertainty, as well as inter-individual differences in physiological behavior. In addition, we examine whether grouping ATCOs based on socio-demographic characteristics and similarities in eye-tracking and head-movement behavior can improve machine-learning–based workload prediction in realistic ATC simulations, while temporal factors are incorporated as additional features to capture changes in cognitive state over time.

3. Data Collection

We conducted the study in a TopSky radar-simulation environment at ATCC Malmö, an area control center in Sweden. The setup included a display configuration comprising a large central screen used for radar visualization, complemented by two smaller auxiliary screens—one on the right presenting flight-plan information and one on the left displaying procedural and weather data. The system employed the original hardware and software in a hardware-in-the-loop setup, thereby providing the ATCO with the impression of a fully operational environment. The eye-tracking device was mounted at the lower edge of the primary screen (see Figure 1).

Task-load levels (light, moderate, heavy) were induced across three simulator runs per participant by varying traffic density, aircraft numbers, and event complexity, with the aim of generating different workload levels. Each simulation run lasted 35–45 minutes. All simulation runs were conducted during daytime hours, with each ATCO's three-run session starting either at 07:30 or at 12:00.

Eye-tracking data were recorded with a Smart Eye XO remote system operating at 250 Hz, capturing gaze behavior, pupil diameter, blink characteristics, and head-movement metrics. The use of a remote, contact-free eye-tracking system minimized interference with natural ATCO behavior and reduced the risk of introducing additional cognitive or physical load due to sensing equipment. Workload ratings were obtained using an adapted 10-point Cooper–Harper Scale (CHS). Scores were collected orally in a non-intrusive manner: ATCOs were prompted by an auditory cue, responded verbally, and the researcher recorded the responses using a dedicated data recording interface developed for this study. A detailed description of each score is provided in Figure 2, where the frame color denotes the corresponding workload level (low, medium, high) used in this study, while the background color reflects the original classification by design (see [20] for details on the adapted mapping).

Eighteen licensed ATCOs participated in the study (9 female, 9 male), all holding valid unit endorsements for the simulated airspace sector. The sample included ATCOs with a broad range of professional experience, with a mean age of M = 46.1 years (SD = 8.3) and an average of M = 19.7 years of professional experience (SD = 8.4). This heterogeneity in age, gender, and experience was intentional, as it enables the investigation of socio-demographic and experiential factors as potential sources of uncertainty in workload prediction. All participants completed all scenarios, producing 54 simulation runs, with valid eye-tracking recordings for 53 of them.

4. Research Objectives

We investigate whether partitioning ATCOs based on socio-demographic factors (gender, age, and experience) leads to subgroups with different workload classification performance. We further examine whether incorporating temporal context (time on task and time of day) affects the relationship between physiological signals and reported workload. Finally, we assess whether grouping ATCOs based on similarities in eye-tracking and head-movement behavior yields more homogeneous subgroups that improve workload classification performance compared to a model trained on the full dataset.



Figure 1 – An ATCO performing a simulated scenario with an eye-tracking device on the bottom of the primary screen

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 2 – Adapted Cooper-Harper Scale

5. Methodology

In our previous work [20], we applied classical supervised machine-learning techniques to classify ATCO workload from eye-tracking data. Eye-tracking recordings were segmented into three-minute intervals, matching the timing of the subjective workload reports. Each three-minute interval was converted into a fixed-length feature vector by computing statistical summaries of the ocular and head-movement signals; these summaries served as the model input features. The corresponding CHS scores for each interval were transformed into class labels. For the multi-class setting, CHS values were mapped to three workload levels (low: score = 1; medium: scores 2–3; high: scores $\geq$ 4). For the binary setting, low and medium workload were grouped into one class, and CHS scores ≥ 4 defined the high-workload class. Model development followed a nested cross-validation scheme, with stratified 10-fold cross-validation for performance estimation and an inner randomized search for hyperparameter optimization. Performance was evaluated using macro-averaged precision, recall, and F1 score. Macro-averaged metrics were selected to mitigate inflated performance estimates caused by majority-class dominance, which was present in our workload classification setting.

In this work, we focus on the kNN and SVC models, with kNN achieving the highest performance and SVC ranking second in our most recent analysis [20]. A previously validated eye-tracking-based workload prediction pipeline serves as the baseline model. To obtain more robust performance estimates, we refine the original stratified 10-fold cross-validation procedure by using a repeated stratified 10-fold scheme with 10 repetitions, each applying a new random partitioning of the data.

We extend the baseline in three ways. First, we evaluate performance across socio-demographic partitions (gender, age, experience) by comparing models trained on the full dataset with models trained within each subgroup. This analysis isolates the contribution of socio-demographic variability and allows us to examine whether more homogeneous subgroups yield more reliable workload predictions. For the socio-demographic analysis, we partition the dataset into subgroups based on gender, age, and experience, using median splits for age and experience to ensure balanced group sizes given the limited sample. For all subgroup analyses, both model training and evaluation are performed within the same subgroup. That is, cross-validation splits are constructed using only data from the corresponding subgroup, ensuring that performance estimates reflect within-group generalization.
Second, we add temporal attributes (time on task and time of day) to assess their contribution. We define time on task as ordinal data, incrementally increasing with consecutive time slots elapsed since the participant began their sequence of runs, providing a unified measure of how long the participant has been continuously engaged in the task. Time of day is defined as the number of time slots counted from the start of the day, capturing broader temporal effects such as potential fatigue later in the day. By explicitly modeling temporal context, we assess whether workload-related physiological patterns evolve systematically over time beyond what is captured by eye-tracking and head-movement features alone.
Third, we construct feature-based representations of each ATCO by computing descriptive statistics of eye-tracking and head-movement variables across all 3-minute intervals. For each variable, we extract the mean, median, standard deviation, minimum, and maximum, capturing both central tendency and variability. The resulting feature vector provides a compact representation of each ATCO's behavioral profile over time.
Using these feature vectors, we quantify similarity between ATCOs with Euclidean distance. This measure captures differences across all descriptive statistics, so that ATCOs with similar overall feature profiles are close in the feature space, while those with distinct profiles are farther apart. Consequently, similarity reflects both typical levels and variability of physiological signals over time. We then apply agglomerative hierarchical clustering with Ward linkage [25] to group ATCOs with similar overall feature profiles. Clustering is performed at the ATCO level, so the resulting groups represent subgroups of individuals with comparable oculometric and head-movement characteristics rather than similarities at individual time points.
To ensure that clusters correspond to meaningful subgroups rather than outliers or pairwise similarities, we impose a minimum cluster size of three ATCOs. For each cluster satisfying this constraint, a dedicated workload classifier is trained and evaluated exclusively using data from ATCOs within that cluster. This clustering-based approach enables the investigation of whether reducing inter-individual heterogeneity through subgrouping can improve machine-learning performance compared to a single global model.
To assess the impact of the proposed extensions, we evaluated performance differences between each condition and the baseline model across repeated runs of the evaluation procedure. In each run, model performance was estimated using cross-validation, and results were aggregated across folds to obtain a single performance value per condition. This procedure was repeated multiple times with different random splits, yielding independent performance estimates for each condition. Assessment of normality using the Shapiro–Wilk test [26] indicated that the performance distributions deviated from normality, motivating the use of nonparametric methods. As the datasets differ in composition and size, the resulting performance estimates are treated as independent samples. Given this independence across conditions, the Mann–Whitney U test [27] was selected to compare performance distributions between groups. Statistical significance was defined at $p \leq 0.05$.

6. Results

In this section, we first present the baseline model results for both kNN and CVS. We then examine how socio-demographic and temporal factors influence the relationship between physiological signals and reported workload, and assess the impact of clustering based on feature-level descriptive statistics on workload classification performance.

Table 1 – Distribution of workload labels (low, medium, high) across socio-demographic subgroups (gender, age, and experience). Abbreviation: F = female, M = male, yrs = years, exp. = experience.

Case	Binary		3 classes		
	low	high	low	medium	high
full dataset	619	48	501	118	48
gender: F	300	21	257	43	21
gender: M	319	27	244	75	27
age ≤ 46 yrs	316	12	251	65	12
age > 46 yrs	303	36	250	53	36
exp. ≤ 22 yrs	425	21	338	87	21
exp. > 22 yrs	194	27	163	31	27

Table 2 – Classification performance (F1 score) of the kNN model when the dataset is split by gender, age, and experience (median-based splits for age and experience). Reported differences (Δ) refer to mean differences relative to the original model. Values in bold indicate statistically significant changes ($p \leq 0.05$); blue shading denotes improvements relative to the baseline, while orange shading denotes decreases. Abbreviation: F = female, M = male, yrs = years, exp. = experience, p = p-value.

Case	Binary			3 classes		
	F1 $\pm$ SD	$\Delta \pm$ SD	p	F1 $\pm$ SD	$\Delta \pm$ SD	p
full dataset	0.815 $\pm$ 0.016			0.775 $\pm$ 0.011		
gender: F	0.827 $\pm$ 0.023	0.011 $\pm$ 0.021	0.212	0.781 $\pm$ 0.025	0.006 $\pm$ 0.026	0.521
gender: M	**0.754 $\pm$ 0.032**	**-0.061 $\pm$ 0.029**	**0.000**	**0.726 $\pm$ 0.025**	**-0.049 $\pm$ 0.026**	**0.000**
age ≤ 46 yrs	**0.677 $\pm$ 0.028**	**-0.138 $\pm$ 0.035**	**0.000**	**0.696 $\pm$ 0.024**	**-0.079 $\pm$ 0.027**	**0.000**
age > 46 yrs	0.800 $\pm$ 0.017	-0.015 $\pm$ 0.025	0.140	0.764 $\pm$ 0.021	-0.011 $\pm$ 0.022	0.162
exp. ≤ 22 yrs	**0.662 $\pm$ 0.046**	**-0.153 $\pm$ 0.048**	**0.000**	**0.692 $\pm$ 0.014**	**-0.083 $\pm$ 0.014**	**0.000**
exp. > 22 yrs	**0.862 $\pm$ 0.021**	**0.047 $\pm$ 0.025**	**0.001**	**0.808 $\pm$ 0.012**	**0.033 $\pm$ 0.019**	**0.000**

6.1 Baseline models

Under the updated evaluation procedure, the mean F1 scores ($\pm$ standard deviation) for the baseline models were as follows: for binary classification, kNN achieved 0.815 $\pm$ 0.016 and SVC 0.781 $\pm$ 0.010; for the three-class problem, kNN reached 0.775 $\pm$ 0.011 and SVC 0.747 $\pm$ 0.010. These results serve as a reference point for evaluating the impact of additional uncertainty-related factors introduced in the subsequent analyses.

6.2 Socio-Demographic Uncertainty Factors (Gender, Age, Experience)

The distribution of workload labels across the socio-demographic subgroups is presented in Table 1. The table reports the number of three-minute intervals assigned to each workload class within each subgroup, separately for the binary and three-class classification settings. This overview provides context for evaluating whether differences in classification performance may be influenced by variations in class distribution across groups.

The statistically significant effects showed the following pattern: for the kNN model, only the group with higher experience demonstrated better model performance than the baseline, whereas the male, younger, and less-experienced groups yielded worse model performance (see Table 2). For the SVC model, the same pattern was observed, with the additional finding that the female and older groups also exceeded the baseline (see Table 3). Taken together, these results suggest that age and gender alone do not define strongly homogeneous workload-response profiles, and that experience appears to be an important contributing factor in shaping these profiles.

Table 3 – Classification performance (F1 score) of the SVC model when the dataset is split by gender, age, and experience (median-based splits for age and experience). Reported differences (Δ) refer to mean differences relative to the original model. Values in bold indicate statistically significant improvements ($p \leq 0.05$); blue shading denotes improvements relative to the baseline, while orange shading denotes decreases. Abbreviation: F = female, M = male, yrs = years, exp. = experience, p = p-value Mann-Whitney U test.

Case	**Binary**			**3 classes**		
	F1 $\pm$ SD	**$\Delta \pm$ SD**	p	**F1 $\pm$ SD**	**$\Delta \pm$ SD**	p
full dataset	0.781 $\pm$ 0.010			0.747 $\pm$ 0.010		
gender, F	**0.843 $\pm$ 0.029**	**0.062 $\pm$ 0.031**	**0.001**	**0.799 $\pm$ 0.018**	**0.051 $\pm$ 0.022**	**0.000**
gender, M	**0.731 $\pm$ 0.021**	**-0.050 $\pm$ 0.020**	**0.000**	**0.718 $\pm$ 0.012**	**-0.029 $\pm$ 0.014**	**0.000**
age ≤ 46 yrs	**0.671 $\pm$ 0.017**	**-0.111 $\pm$ 0.017**	**0.000**	**0.701 $\pm$ 0.021**	**-0.046 $\pm$ 0.024**	**0.000**
age > 46 yrs	**0.832 $\pm$ 0.014**	**0.051 $\pm$ 0.018**	**0.000**	**0.806 $\pm$ 0.015**	**0.059 $\pm$ 0.017**	**0.000**
exp. ≤ 22 yrs	**0.658 $\pm$ 0.022**	**-0.123 $\pm$ 0.019**	**0.000**	**0.677 $\pm$ 0.017**	**-0.070 $\pm$ 0.019**	**0.000**
exp. > 22 yrs	**0.876 $\pm$ 0.018**	**0.094 $\pm$ 0.022**	**0.000**	**0.822 $\pm$ 0.023**	**0.074 $\pm$ 0.028**	**0.000**

Table 4 – Effect of adding temporal features on workload classification performance (F1 score) for the kNN model. Reported differences (Δ) refer to mean differences relative to the original model. Values in bold indicate statistically significant improvements ($p \leq 0.05$); blue shading denotes improvements relative to the baseline. Abbreviation: TonT = time on task, TofD = time of day, p = p-value.

Case	**Binary**			**3 classes**		
	F1 $\pm$ SD	**$\Delta \pm$ SD**	p	**F1 $\pm$ SD**	**$\Delta \pm$ SD**	p
full dataset	0.815 $\pm$ 0.016			0.775 $\pm$ 0.011		
+ TonT	**0.836 $\pm$ 0.009**	**0.021 $\pm$ 0.014**	**0.009**	**0.795 $\pm$ 0.011**	**0.020 $\pm$ 0.016**	**0.002**
+ TofD	0.823 $\pm$ 0.017	0.008 $\pm$ 0.016	0.307	0.777 $\pm$ 0.011	0.002 $\pm$ 0.013	0.623

Table 5 – Effect of adding temporal features on workload classification performance (F1 score) for the SVC model. Reported differences (Δ) refer to mean differences relative to the original model. Values in bold indicate statistically significant improvements ($p \leq 0.05$); blue shading denotes improvements relative to the baseline. Abbreviation: TonT = time on task, TofD = time of day, p = p-value.

Case	**Binary**			**3 classes**		
	F1 $\pm$ SD	**$\Delta \pm$ SD**	p	**F1 $\pm$ SD**	**$\Delta \pm$ SD**	p
full dataset	0.781 $\pm$ 0.010			0.747 $\pm$ 0.010		
+ TonT	**0.796 $\pm$ 0.015**	**0.014 $\pm$ 0.012**	**0.031**	**0.769 $\pm$ 0.009**	**0.021 $\pm$ 0.006**	**0.001**
+ TofD	0.778 $\pm$ 0.011	-0.004 $\pm$ 0.007	0.521	**0.773 $\pm$ 0.015**	**0.026 $\pm$ 0.007**	**0.005**

6.3 Temporal Uncertainty Factors (Time on Task, Time of Day)

Temporal factors showed small effects for both models. Time on task improved performance in both the kNN and SVC models, while time of day did not yield a statistically significant improvement overall; however, in the three-class setting, time of day produced a slight increase (see Tables 4, 5).

6.4 ATCO Grouping via Clustering Based on Descriptive Statistics of Eye-Tracking Features Across Time Intervals

The dendrogram illustrating the hierarchical structure of ATCO feature profiles, based on feature vectors derived from descriptive statistics, is shown in Figure 3. It represents pairwise similarity through the linkage height at which clusters are merged: ATCOs that merge at lower levels are more similar, while merges at higher levels indicate greater dissimilarity based on the euclidean distance.

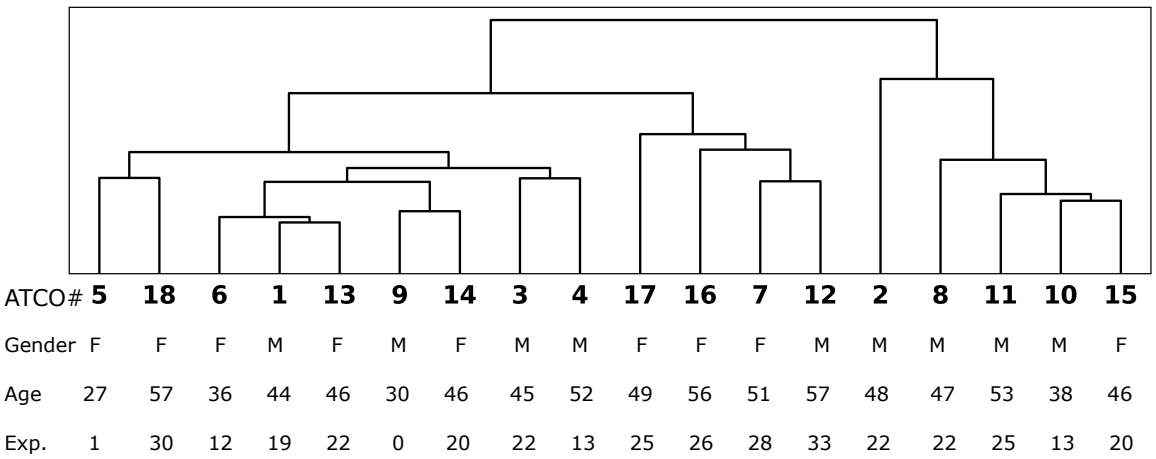


Figure 3 – Hierarchical clustering dendrogram of ATCO eye-tracking profiles. Abbreviation: F = female, M = male, exp. = experience.

Table 6 – ATCO composition of the selected clusters with workload labels shown for both binary and three-class schemes.

		Binary		3 classes		
Cluster ID	**ATCOs included**	**# low**	**# high**	**# low**	**# medium**	**# high**
C1.1	[6, 1, 13]	107	10	47	60	10
C1.2	[6, 1, 13, 9, 14]	183	10	91	92	10
C1.3	[6, 1, 13, 9, 14, 3, 4]	242	11	120	122	11
C2.1	[16, 7, 12]	112	4	76	36	4
C2.2	[17, 16, 7, 12]	145	11	86	59	11
C3.1	[11, 10, 15]	86	21	38	48	21
C3.2	[8, 11, 10, 15]	125	21	66	59	21
C3.3	[2, 8, 11, 10, 15]	162	22	89	73	22

We obtain a full hierarchy of possible ATCO groupings by applying hierarchical clustering. We examined clusters at different levels (altitude) of this hierarchy and evaluated their suitability for subsequent workload classification. Only clusters that contained at least three ATCOs and that produced improved or competitive classification performance relative to the baseline model were retained. Table 6 presents the clusters obtained from the hierarchical structure. Table 7 summarizes the classification performance for the identified clusters with sufficient class representation for binary classification, highlighting statistically significant differences relative to the baseline model. Cluster C2.1 is excluded because it contains fewer than ten high-workload labels, which is insufficient for the 10-fold evaluation procedure, as some folds would otherwise contain no positive samples. Table 8 presents the corresponding results for the three-class classification setting.

7. Discussion

In this study, we examine how different sources of uncertainty influence machine-learning–based workload prediction for ATCOs using eye-tracking and head-movement data. Explicit analysis of socio-demographic characteristics, temporal context, and inter-individual differences in physiological behavior provides insight into why workload prediction performance varies across ATCOs and modeling approaches.

Rather than viewing these differences as a limitation, we demonstrate that several sources of uncertainty can be systematically characterized and, in some cases, leveraged to improve model performance.

Table 7 – Binary classification performance (F1 score) of kNN and SVC models for the selected cluster derived from eye-tracking-based Euclidean-distance hierarchical clustering. Reported differences (Δ) represent mean differences relative to the baseline model trained on the full dataset. Values in bold indicate statistically significant changes ($p \leq 0.05$); blue shading denotes improvements relative to the baseline, while orange shading denotes decreases.

Dataset	kNN			SVC		
	F1 $\pm$ SD	**$\Delta \pm$ SD**	**p-value**	**F1 $\pm$ SD**	**$\Delta \pm$ SD**	**p-value**
full	0.815 $\pm$ 0.016			0.781 $\pm$ 0.010		
C1.1	0.787 $\pm$ 0.036	-0.028 $\pm$ 0.038	0.090	0.789 $\pm$ 0.028	0.008 $\pm$ 0.028	0.623
C1.2	**0.881 $\pm$ 0.028**	**0.066 $\pm$ 0.023**	**0.000**	**0.861 $\pm$ 0.011**	**0.080 $\pm$ 0.016**	**0.000**
C1.3	**0.875 $\pm$ 0.023**	**0.060 $\pm$ 0.032**	**0.000**	0.795 $\pm$ 0.025	0.014 $\pm$ 0.025	0.307
C2.2	**0.694 $\pm$ 0.025**	**-0.121 $\pm$ 0.015**	**0.000**	**0.603 $\pm$ 0.027**	**-0.179 $\pm$ 0.029**	**0.000**
C3.1	**0.878 $\pm$ 0.029**	**0.063 $\pm$ 0.034**	**0.000**	**0.885 $\pm$ 0.023**	**0.103 $\pm$ 0.028**	**0.000**
C3.2	**0.850 $\pm$ 0.032**	**0.034 $\pm$ 0.044**	**0.014**	**0.903 $\pm$ 0.007**	**0.121 $\pm$ 0.009**	**0.000**
C3.3	**0.838 $\pm$ 0.023**	**0.023 $\pm$ 0.022**	**0.038**	**0.873 $\pm$ 0.018**	**0.092 $\pm$ 0.019**	**0.000**

Table 8 – Three-classes classification performance (F1 score) of kNN and SVC models for the selected cluster derived from eye-tracking-based Euclidean-distance hierarchical clustering. Reported differences (Δ) represent mean differences relative to the baseline model trained on the full dataset. Values in bold indicate statistically significant changes ($p \leq 0.05$); blue shading denotes improvements relative to the baseline, while orange shading denotes decreases.

Dataset	kNN			SVC		
	F1 $\pm$ SD	**$\Delta \pm$ SD**	**p-value**	**F1 $\pm$ SD**	**$\Delta \pm$ SD**	**p-value**
full	0.775 $\pm$ 0.011			0.747 $\pm$ 0.010		
C1.1	0.758 $\pm$ 0.054	-0.017 $\pm$ 0.052	0.791	0.746 $\pm$ 0.030	-0.001 $\pm$ 0.030	0.273
C1.2	**0.825 $\pm$ 0.023**	**0.050 $\pm$ 0.025**	**0.001**	**0.789 $\pm$ 0.013**	**0.041 $\pm$ 0.015**	**0.000**
C1.3	**0.805 $\pm$ 0.025**	**0.030 $\pm$ 0.028**	**0.017**	0.742 $\pm$ 0.033	-0.006 $\pm$ 0.036	0.473
C2.2	**0.704 $\pm$ 0.019**	**-0.071 $\pm$ 0.024**	**0.000**	**0.684 $\pm$ 0.024**	**-0.064 $\pm$ 0.021**	**0.000**
C3.1	0.795 $\pm$ 0.026	0.020 $\pm$ 0.029	0.054	**0.866 $\pm$ 0.018**	**0.119 $\pm$ 0.016**	**0.000**
C3.2	0.783 $\pm$ 0.023	0.008 $\pm$ 0.030	0.385	**0.848 $\pm$ 0.019**	**0.101 $\pm$ 0.019**	**0.000**
C3.3	**0.798 $\pm$ 0.020**	**0.023 $\pm$ 0.022**	**0.017**	**0.824 $\pm$ 0.015**	**0.077 $\pm$ 0.018**	**0.000**

Role of Socio-Demographic Factors and Label Reliability. Gender and age do not define subgroups with consistent performance gains; however, male and younger ATCOs exhibit consistently lower performance across models. This indicates that these factors do influence workload prediction performance, but not in a way that produces clearly defined, high-performing subgroups. Instead, they introduce differences in performance across ATCOs, contributing to greater variability in the data. As a result, grouping ATCOs based on gender or age does not lead to more homogeneous subgroups that would support improved model performance.

Among the socio-demographic variables considered, professional experience emerged as the most influential factor. Models trained and evaluated within the more experienced subgroup achieved higher performance than the baseline, whereas performance decreased for the less-experienced subgroup (Table 2, Table 3). Specifically, for the more experienced group, performance increased by approximately 4.7% and 3.3% (absolute F1 score differences) in the binary and three-class settings, respectively, for kNN, and by 9.5% and 7.5% for SVC. This indicates that the more experienced ATCOs form a more homogeneous subset for workload classification.

A possible explanation for this pattern may lie in differences in reporting behavior. More experienced

ATCOs may provide better subjective workload ratings that are more stable and better aligned with the underlying physiological signals, allowing these labels to function more effectively as ground truth. They may also be less susceptible to social desirability pressures and therefore report high workload more consistently when it occurs. In contrast, less-experienced ATCOs may exhibit greater variability in both task execution strategies and workload reporting, leading to noisier and less reliable labels.
This suggests that a substantial portion of the observed performance variation may stem from label uncertainty rather than from limitations in the physiological features themselves. In this context, experience appears to be closely linked to the stability and reliability of subjective workload ratings used for supervised learning.

Interpretation of Temporal Effects. Temporal factors played a limited role in workload prediction. Of the temporal variables examined, only time on task yielded consistent, albeit modest, improvements in classification performance, with increases of approximately 2.1% and 2.0% for kNN and 1.4% and 2.1% for SVC in the binary and three-class settings, respectively. However, time of day did not produce a statistically significant effect overall (Table 4, Table 5).
These results suggest that workload-related physiological responses evolve gradually over sustained task engagement. One possible explanation is that time on task introduces cumulative cognitive demand and fatigue, which may alter how ATCOs perceive and report workload over time. In this sense, time on task can be interpreted as a moderating factor that influences the relationship between physiological signals and self-reported workload, rather than serving as a direct predictor. As fatigue increases, ATCOs may become more sensitive to variations in task demands, potentially leading to greater variability and stronger fluctuations in reported workload.
Additional factors may also contribute to this effect. In particular, the structure of the simulation scenarios may lead to increasing task complexity over time, making time on task indirectly associated with rising operational demand. Incorporating this variable may therefore provide complementary temporal context that helps align physiological patterns with reported workload levels.
The absence of a strong effect for time of day suggests that broader circadian influences were minimal within the experimental setting. This may partly be explained by the relatively short duration of the simulation runs, which limited the extent to which temporal effects, such as fatigue or sustained changes in cognitive state, could develop. Overall, temporal context appears to refine workload prediction only modestly rather than fundamentally altering model performance.

Inter-Individual Differences in Eye-Tracking Behavior and the Value of Subgrouping. The most substantial performance differences were observed when grouping ATCOs based on descriptive representations of their eye-tracking and head-movement behavior. The hierarchical clustering procedure produced multiple candidate subgroups at different levels of the dendrogram, several of which yielded improved workload classification performance relative to the global model, whereas only one showed reduced performance (Table 7). Cluster C1.2 demonstrated a marked improvement over the global model, with performance gains of 6.6% for kNN and 8.0% for SVC. In addition, several alternative clusters (C3.1–C3.3), comprising between three and five ATCOs, also yielded improved performance. These clusters achieved kNN gains ranging from 2.3% to 6.3% and SVC gains from 9.2% to 12.1%. This highlights a trade-off between cluster size and performance: larger clusters provide more data and potentially more stable models, while smaller clusters may offer higher performance gains but rely on fewer ATCOs. The selection of an appropriate cluster therefore depends on the desired balance between model robustness and performance improvement. ATCOs not included in the better-performing subgroups may be better handled by alternative modeling strategies; however, this may also reflect the limited sample size, and more stable subgroup structures could emerge in a larger dataset or real-world application. No consistent relationship was observed between cluster membership and socio-demographic or temporal factors such as age, experience, or time of day, suggesting that the identified groups primarily reflect differences in eye-tracking and head-movement behavior rather than external attributes.
A similar pattern is observed in the three-class setting (Table 8), where several clusters—most notably C1.2 and C3.1–C3.3—also demonstrate performance improvements over the baseline, particularly

for the SVC model. However, the magnitude and consistency of these gains are generally lower than in the binary case, and some clusters do not yield improvements, indicating that subgroup-specific modeling is more sensitive to class complexity and label distribution in the multi-class setting.
In addition, the results show that performance gains are not uniform across models, with different clusters benefiting more from either kNN or SVC. This suggests that each cluster may favor a different classification model.
Taken together, these findings suggest that inter-individual differences in physiological behavior are not merely noise but reflect stable characteristics that can be leveraged for modeling. At the same time, the results indicate that the benefits of subgroup-specific modeling are concentrated in a subset of ATCOs, whereas for others—either belonging to less well-performing clusters, clusters with insufficient label diversity, or not assigned to a stable cluster—alternative modeling approaches such as global or personalized models are likely to provide more reliable performance.
The effectiveness of clustering may also depend on the choice of feature representation and clustering strategy. The present approach relies on descriptive-statistics-based representations and hierarchical clustering; these capture differences in the central tendency and variability of the signals. Alternative approaches may capture complementary aspects of inter-individual differences and could further enhance subgroup identification.

Implications for Workload Modeling and ATM Applications. From an applied perspective, the results have important implications for the design of workload monitoring systems in ATC. First, they suggest that uncertainty-aware modeling—explicitly accounting for experience and inter-individual variability—can yield more robust and interpretable predictions than purely feature-driven approaches. Second, the strong effect of experience raises questions about the suitability of subjective workload ratings as universal ground truth. Future systems may benefit from combining subjective labels with objective performance measures or adaptive labeling strategies that account for individual differences in self-reporting. Finally, feature-based grouping offers a scalable path toward personalization without requiring fully individualized models. In addition, the results demonstrate the potential of workload monitoring as a quantitative tool for assessing and supporting personnel training.

Limitations. Several limitations should be considered when interpreting the results. Although the sample size is comparatively large for applied ATC studies, it limits the number and size of identifiable subgroups. In addition, the reliance on subjective workload ratings introduces potential biases, including self-reporting bias and social desirability effects. These affect the reliability of the ground-truth labels. Furthermore, the scarcity of high-workload instances in some clusters did not allow the testing for these clusters.

8. Conclusions and Future Work

We showed that inter-individual differences and professional experience provide meaningful gains in ATCO workload prediction based on eye-tracking and head-movement data, whereas temporal factors contribute only marginal improvements. Other socio-demographic factors, such as gender and age, do not yield consistent predictive benefits. The observed advantage of more experienced ATCOs may reflect greater stability and reliability in self-reported workload labels. Clustering of ATCOs based on descriptive feature representations highlights substantial inter-individual differences in oculometric and head-movement behavior. Overall, these findings underscore the importance of accounting for uncertainty and inter-individual differences and suggest that subgroup-specific or personalized models are a promising direction for future adaptive ATM applications.
Future work might explore alternative clustering strategies and feature representations to further characterize inter-individual differences in workload-related physiological behavior. In addition, complementary labeling approaches will be investigated to reduce uncertainty associated with subjective workload reports.
For real-world deployment, clustering-based workload prediction will require larger datasets with a greater number of ATCOs to adequately cover the diversity of operational behavior. In particular, additional data are needed to populate sparsely represented regions of the feature space, including

ATCOs that are currently isolated or weakly grouped in the clustering structure. Ensuring sufficient numbers of ATCOs within each cluster will enable the formation of stable and well-balanced subgroups, thereby supporting reliable cluster-specific models and robust performance estimates in operational settings. This provides a pathway toward adaptive, subgroup-specific modeling approaches that support more personalized workload prediction in operational ATC environments.
Additionally, supporting robust model performance across workload levels in operational contexts will require scenario designs or data acquisition strategies that systematically induce even higher workload levels, for example through increased traffic density, more complex conflict situations, or time-critical events. Expanding the representation of such conditions is essential for achieving a more balanced workload distribution and improving model robustness in demanding operational contexts.

9. Contact Author Email Address

Mailto: Anastasia.Lemetti@liu.se

10. Ethics Statement

The study received ethical approval from the Swedish Ethical Review Authority (2023-06110-01).

11. Funding Statement

This study was supported in the scope of project On WorkLoad Measures 2 (OWL2), funded by the Swedish Transport Administration (TRV 2024/99615).

12. 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] Brian Hilburn and PGAM Jorna. Workload and air traffic control. *Stress, workload, and fatigue*, pages 384–394, 2001.

[2] Kyle Kent Edward Ellis. Eye tracking metrics for workload estimation in flight deck operations. Master's thesis, University of Iowa, 2009.

[3] 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, New York, NY, USA, 2016. Association for Computing Machinery.

[4] 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*, pages 262–277, Cham, 2020. Springer International Publishing.

[5] 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.

[6] Monika Kaczorowska, Małgorzata Plechawska-Wójcik, and Mikhail Tokovarov. Interpretable machine learning models for three-way classification of cognitive workload levels for eye-tracking features. *Brain sciences*, 11(2):210, 2021.

[7] 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.

[8] William Romine, Noah Schroeder, Tanvi Banerjee, and Josephine Graft. Toward mental effort measurement using electrodermal activity features. *Sensors*, 22(19):7363, 2022.

[9] Chaeyoung Lee, MinJu Shin, David Eniyandunmo, Alvee Anwar, Eunsik Kim, Kyongwon Kim, Jae Keun Yoo, and Chris Lee. Predicting driver's mental workload using physiological signals: A functional data analysis approach. *Applied Ergonomics*, 118:104274, 2024.

[10] Aisy Al Fawwaz, Osmalina Nur Rahma, Khusnul Ain, Sayyidul Istighfar Ittaqillah, and Rifai Chai. Measurement of mental workload using heart rate variability and electrodermal activity. *Ieee Access*, 12:197589–197601, 2024.

[11] Merve Ekin, Krzysztof Krejtz, Carlos Duarte, Andrew T Duchowski, and Izabela Krejtz. Prediction of intrinsic and extraneous cognitive load with oculometric and biometric indicators. *Scientific Reports*, 15(1):5213, 2025.

[12] 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)*, pages 1–8, 2023.

[13] Pierluigi Dell'Acqua, Marco Garofalo, Francesco La Rosa, and Massimo Villari. Your eyes under pressure: Real-time estimation of cognitive load with smooth pursuit tracking. *Big Data and Cognitive Computing*, 9(11):288, 2025.

[14] Maria Trigka, Elias Dritsas, and Phivos Mylonas. Eye-based cognitive overload prediction in human-machine interaction via machine learning. In *21st International Conference on Web Information Systems and Technologies (WEBIST)*, 2025.

[15] Jad Atweh and Sara Riggs. Predicting mental demand of teammates using eye tracking metrics: A machine learning approach. In *Proceedings of the 2025 Symposium on Eye Tracking Research and Applications*, pages 1–3, 2025.

[16] Sachithra Karunathilake, Nurul Ahad Choudhury, Akash Deep, and Pratima Saravanan. Predicting cognitive workload in visuospatial tasks using pupillometry: A machine learning approach. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 69, pages 1680–1686. SAGE Publications Sage CA: Los Angeles, CA, 2025.

[17] José A Navia, Jorge Ibáñez-Gijón, David Travieso, Aitor Montes, and Patricia López de Frutos. Eye-tracking metrics for estimating workload and characterizing errors in conflict detection and resolution during simulated en route air-traffic control. *Frontiers in Psychology*, 16:1644721, 2025.

[18] Giovanni Pignoni and Sashidharan Komandur. Development of a quantitative evaluation tool of cognitive workload in field studies through eye tracking. In Don Harris, editor, *Engineering Psychology and Cognitive Ergonomics - 16th International Conference, EPCE 2019*, volume 11571 of *Lecture Notes in Computer Science*, pages 106–122. Springer, 2019.

[19] 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. *The Aeronautical Journal*, pages 1–30, 2024.

[20] Anastasia Lemetti, Lothar Meyer, Maximilian Peukert, Tatiana Polishchuk, Christiane Schmidt, and Helene Alpfjord Wylde. Predicting air traffic controller workload from eye-tracking data with machine learning. *Journal of Open Aviation Science*, 3(1), 2025.

[21] 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. *Computer Methods and Programs in Biomedicine*, 224:106989, 2022.

[22] Murat Kucukosmanoglu, Justin Brooks, Catherine Neubauer, and Andrea Krausman. Multimodal classification of cognitive workload using eye-tracking, ECG, and head motion data in simulated military missions, 2025. Preprint, Research Square, Version 1.

[23] Leo Julius Materne and Maik Friedrich. Supervision of multiple remote tower centers: Evaluating a new air traffic control interface based on mental workload and eye tracking. *IEEE Transactions on Human-Machine Systems*, 55(2):114–123, 2025.

[24] Jiahui An, Chonghao Cai, Olympia Gallou, Sara Irina Fabrikant, Giacomo Indiveri, and Elisa Donati. Neuromorphic deployment of spiking neural networks for cognitive load classification in air traffic control. In *2025 International Conference on Neuromorphic Systems (ICONS)*, pages 127–134. IEEE, 2025.

[25] Joe H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[26] Samuel S. Shapiro and Martin B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[27] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.