Journal of Open Aviation Science (2025), Vol.**3**
doi:10.59490/joas.2025.8579

JOAS

EDITORIAL

Reviews and Responses for Predicting Air Traffic Controller Workload from Eye-Tracking Data with Machine Learning

Authors: Anastasia Lemetti, Lothar Meyer, Maximilian Peukert, Tatiana Polishchuk, Christiane Schmidt, and Helene Alpfjord Wylde

Reviewers: Maik Friedrich and Timothé Krauth

Editor: Junzi Sun

1. Original paper

The DOI for the original paper is https://doi.org/10.59490/joas.2025.8034

2. Review - round 1

2.1 Reviewer 1

General Overview

The paper is well-organized, striking a solid balance between presenting state-of-the-art research and addressing the specific problem at hand. I found the work very interesting and valuable, with some comments and some minor changes for review.

Despite a relatively small sample size of 18 participants, this is a respectable number for a study involving Air Traffic Controllers (ATCOs). Additionally, the gender balance among participants is commendable. The study incorporates three distinct levels of workload and engages in feature selection to refine its analytical approach, that is a good approach in order to stepwise understand the possiblites of AI.

Literature Review

The focus on machine learning in workload assessment should primarily center on eye-tracking, as EEG−−while potentially valuable−−is not the main focus of the study and may feel misplaced. A major gap in the literature exists regarding the conceptualization of workload, particularly in the context of AI-based identification. The authors should clarify how they define workload in relation to AI training, as this is critical for understanding their methodology.

Data Collection

The reclassification of the Cooper-Harper Scale is justified by the high level of training ATCOs receive, as inducing workload in such highly trained professionals presents unique challenges. This consideration should also be reflected in the literature review. Also there are studies showing "Ceiling effects" with self assessment and workload that could suppport the authors with the reclassification.

The data collection process requires further elaboration. The paper would benefit from detailed descriptions of scenarios, outlining how they differ in terms of key factors and how they were designed to induce workload. Such clarifications would enhance transparency and robustness in the study's methodology.

Machine Learning Approach

The handling of missing values and outliers is thorough, and the feature selection and evaluation process are both well-documented and informative.

While the diversity of machine learning models used is impressive, questions arise regarding the choice of certain models. Given the task and the existing literature, logistic regression and linear discriminant analysis seem less likely to perform well compared to models like decision trees, random forests, gradient boosting, and bagging. If the latter models did not outperform, it might indicate issues with the dataset rather than the algorithms themselves. The inclusion of k-nearest neighbors and other models raises concerns about whether the authors are simply testing a wide variety of methods in hopes of finding significant results. While a combination of hypothesis-driven and exploratory research is valuable, excessive model testing can lead to unreliable conclusions.

Additionally, assumptions about the connection between head movement and redundancies are introduced prematurely in Chapter 4 and should be relocated to the discussion section for a more logical flow.

Discussion

The first section of the discussion reads more like a conclusion. If the authors are already confident in their findings at this stage, the discussion risks losing its purpose.

The discussion is primarily focused on side effects found during analysis rather than a broader contextualization of findings. Greater insights could be provided by comparing results with other studies in the field. The study achieves an 80% prediction accuracy, which is highly impressive—yet it would be valuable to investigate what specifically contributed to this superior performance compared to other models in similar research.

I would also encourage the authors to reflect on the limitations of their study.

Conclusion

The conclusion is highly similar to the discussion section, which reduces its impact. It would benefit from clearer differentiation, emphasizing key takeaways and broader implications rather than reiterating previous points.

Recommendation: Revisions Required

2.2 Reviewer 2

This paper presents a novel methodology that explores eye-tracking data and feature engineering coupled with machine learning to predict the workload of ATCOs. Overall, the paper is well strcutured and the reader understands well what are the contributions of the authors. Great care has been taken to justify the methodological choices in the machine learning approach. Here are comments that might improve the manuscript, essentially concerning the form:

Introduction:

- The introduction is well structured, and the main contributions of the paper are stated. However, it is hard to clearly identify the gap in the current state of the literature that is addressed. I suggest to add brief explanations about what are the shortcomings of the current methods, and how yours is tackling some of them.

Related work:

- This section is very well furnished with previous research that are relevant regarding the topic of the paper. However, the structure is difficult to follow, and the section is a stack of methodologies, making it rather difficult to understand how you used this literature to develop your methodology. I suggest reorganizing it to highlight the various improvements in the context of workload prediction, the important points you decided to retain, and, conversely, the areas for improvement you decided to work on.

- Line 77: CHS is used whereas it is only defined later.

Data Collection:

- I appreciated the justification of the choice of the CHS over other metrics. Though the CHS score is "human-based", it seems that the questions for assessment are sufficiently well-defined so that the choice of the rating is sharp enough and doesn't depend on the person conducting the evaluation. This noise reduction is further improved thanks to the 3 level aggregation that prevents having scenarios that can be classified in two different categories.

- I understood that the workload self-assessment is performed during the simulation runs. To you think that the fact ATCOs have to answer your question on top of working might introduce noise in the physiological data you collect ?

Machine Learning Approach:

- This section is very long, and thus quite difficult to read. I would suggest to split it into two sections: one describing the methodology, and one expliciting the results.

- Section 4.1 could be presented as a table (title, type, description). I would also explicitely state that they are time series data.

- The features that are computed are only simple statistical descriptors of the variables that are collected (mean, median, std). Have you ever tried computing other features like avg fixation duration, avg saccade amplitude, avg gaze velocity / acceleration, etc... Moreover, did you try more complex time series features such as time fromain or frequency domain features (cf. TSFresh library)? Finally, why did you only focus on scalar data models and did not explore sequence models (e.g. LSTM) ?

- The difference between the ATCO simulation runs corresponding to three different task load scenarios and the labeling of the data in section 4.3 is slightly misleading. For a given simulation run (e.g. high workload), there can be data points that correspond to low to medium workloads ? Table 1 confused me at first. Moreover, if the CHS score is assessed every 3 minutes during the simulations, don't you fear it introduce noise in the data you collect ?

- If I understand well, the cross-validation sets are made by shuffling the data points collected for all scenarios and all participants, while maintaining the class distribution. In my experience, it's often a good practice to perform a "participant-based" cross validation, so that you don't have data points belonging to one participant in both the training and validation sets. It avoids data leakage, in the sens that the machine learning model won't overfit the specificities of one particular participant. As a result, the model should be more robust to the different behaviors that can be observed between the paritcipants. Could you discuss this ?

- Why not testing the Lasso regression (instead of a regular LR) for feature selection, and comparing with RFE results ?

- Table 2: are those metrics representing the average over the cross-validation, or the best fold ?

- The description of SET1/2/3 could be improved with a table.

Discussion:

- The discussion of the results and the comparison with similar studies is very interesting. However, in my opinion, I think the description of the different ML models is too detailed, and seems too generic regarding the main goal of your study.

Recommendation: Revisions Required

3. Response - round 1

3.1 Response to reviewer 1

Literature Review

The focus on machine learning in workload assessment should primarily center on eye-tracking, as EEG—while potentially valuable—is not the main focus of the study and may feel misplaced. A major gap in the literature exists regarding the conceptualization of workload, particularly in the context of AI-based identification. The authors should clarify how they define workload in relation to AI training, as this is critical for understanding their methodology.

Response

We thank the reviewer for the comments and constructive feedback. First, we agree with the reviewer that the focus should remain on eye-tracking rather than EEG. Accordingly, we have removed related work primarily centered on EEG-based workload assessment to ensure that the section remains closely aligned with the scope of our study.

Second, regarding the conceptualization of workload in the context of AI training, we now make the link between the theoretical definition and its practical use more explicit. A new subsection, Definition of Workload, has been added to the Data Collection section, where we describe how Hart and Staveland's view of workload informs our use of CHS ratings in model training. Placing this explanation in the Data Collection section allows us to present the conceptual foundation alongside the description of how workload is measured and applied, making the connection clearer and more directly relevant to the methodological approach.

We were not entirely certain if we interpreted the reviewer's second part of the comment correctly. Should the reviewer have intended a different emphasis or clarification, we would be grateful for further guidance.

Data Collection

The reclassification of the Cooper-Harper Scale is justified by the high level of training ATCOs receive, as inducing workload in such highly trained professionals presents unique challenges. This consideration should also be reflected in the literature review. Also there are studies showing "Ceiling effects" with self assessment and workload that could support the authors with the reclassification.

Response

In the revised manuscript, we have strengthened the justification for the reclassification of CHS ratings in the Data Collection section. We now emphasize that the high level of training and expertise among licensed ATCOs makes it difficult to induce high workload levels in simulations, which reduces variability in subjective self-assessments and can lead to ceiling effects. To support this point, we added references discussing ceiling effects in workload assessment.

We also refer to our earlier work highlighting broader limitations of subjective workload measures such as CHS, NASA-TLX, and ISA, including intrusiveness, social desirability bias, and lack of calibration. Taken together, these considerations clarify the rationale for reclassifying CHS scores in order to obtain meaningful workload categories for ML model training and analysis.

The data collection process requires further elaboration. The paper would benefit from detailed descriptions of scenarios, outlining how they differ in terms of key factors and how they were designed to induce workload. Such clarifications would enhance transparency and robustness in the study's methodology.

Response

In the revised manuscript, we added a new subsection titled Scenario Design Verification, where we describe how the light, moderate, and heavy task-load scenarios were derived from simulator exercises originally developed for ATCO training at LFV. These scenarios were selected based on expert judgment to represent different levels of operational complexity. As the exercises were already established for training purposes, we did not modify their structure but instead verified their suitability for research by applying a dedicated scoring system. This system, adapted from Zamarreno et al. (2022), quantitatively assesses scenario complexity through an event-based scheme, summarized in Table 1. An initial validation using data from a single ATCO confirmed a clear progression in scenario scores (142, 221, and 538), demonstrating that the scenarios are appropriately differentiated in terms of task load. We believe these additions enhance both transparency and methodological rigor.

Machine Learning Approach

The handling of missing values and outliers is thorough, and the feature selection and evaluation process are both well-documented and informative. While the diversity of machine learning models used is impressive, questions arise regarding the choice of certain models. Given the task and the existing literature, logistic regression and linear discriminant analysis seem less likely to perform well compared to models like decision trees, random forests, gradient boosting, and bagging. If the latter models did not outperform, it might indicate issues with the dataset rather than the algorithms themselves. The inclusion of k-nearest neighbors and other models raises concerns about whether the authors are simply testing a wide variety of methods in hopes of finding significant results. While a combination of hypothesis-driven and exploratory research is valuable, excessive model testing can lead to unreliable conclusions.

Response

We agree that linear models are less suitable for eye-tracking-based workload classification; therefore, we removed logistic regression and LDA. We also revised Section 4.4 to justify the remaining models: tree-based methods, ensemble approaches, SVM, and kNN. This selection is supported by large-scale comparative benchmarking studies as well as by workload-focused eye-tracking studies (the corresponding references have been added to the text). In our analysis, tree-based methods did not yield the best performance, which may be attributable to the limited size of the dataset (667 records), as such methods are known to be prone to overfitting in small-sample settings.

Additionally, assumptions about the connection between head movement and redundancies are introduced prematurely in Chapter 4 and should be relocated to the discussion section for a more logical flow.

Response

In the revised manuscript, we have relocated the discussion of assumptions regarding the connection between head movement and redundancies from Chapter 4 (now Chapter 5 in the revised version) to the Discussion section. We believe this change improves the logical flow of the paper as suggested.

Discussion

The first section of the discussion reads more like a conclusion. If the authors are already confident in their findings at this stage, the discussion risks losing its purpose. The discussion is primarily focused on side effects found during analysis rather than a broader contextualization of findings. Greater insights could be provided by comparing results with other studies in the field. The study achieves an 80% prediction accuracy, which is highly impressive—yet it would be valuable to investigate what specifically contributed to this superior performance compared to other models in similar research.

Response

In the revised manuscript, we have removed statements from the Discussion section that were more conclusive in nature. The Discussion now focuses on interpreting the results in relation to previous research and providing broader contextual insights into the study's findings.

Regarding the request for broader contextualization, we would like to note that the manuscript already discusses our results with respect to two key aspects: (i) the factors that may explain why the best performing model performs better than others, and (ii) comparisons with the relatively few related studies available in this domain. We believe these discussions capture the intent of the reviewer's suggestion. However, if we have misunderstood the comment, we would be grateful for further clarification so that we can address it more precisely.

I would also encourage the authors to reflect on the limitations of their study.

Response

In the revised manuscript, we have added a dedicated subsection Limitations at the end of the Discussion section. This subsection reflects on key constraints of our study, including: (i) the reclassification of the CHS scale, (ii) the limited dataset size from a machine learning perspective despite the relatively large number of participating ATCOs, (iii) the fact that data were collected in a single working environment, (iv) the potential influence of social desirability bias in subjective ratings, and (v) the inherent softness and multidimensionality of the workload construct. We believe that this addition clarifies the study's boundaries and enhances the transparency of our findings.

Conclusion

The conclusion is highly similar to the discussion section, which reduces its impact. It would benefit from clearer differentiation, emphasizing key takeaways and broader implications rather than reiterating previous points.

Response

In the revised manuscript, we have restructured the Discussion and Conclusion sections to ensure clearer differentiation. The Discussion now focuses on interpreting the findings and situating them in relation to prior work, while the Conclusions and Outlook section emphasizes the key takeaways, broader implications of the study, and directions for future work. We believe this restructuring enhances the impact of both sections.

3.2 Response to reviewer 2

Introduction:

The introduction is well structured, and the main contributions of the paper are stated. However, it is hard to clearly identify the gap in the current state of the literature that is addressed. I suggest to add brief explanations about what are the shortcomings of the current methods, and how yours is tackling some of them.

Response

We thank the reviewer for the helpful comments and feedback. In the revised Introduction, we now explicitly highlight the shortcomings of existing work: most prior studies rely on subjective ratings with known limitations, and while some combine eye-tracking with machine learning, these have not been conducted in the ATC environment. We then clarify how our study addresses this gap.

Related work:

- This section is very well furnished with previous research that are relevant regarding the topic of the paper. However, the structure is difficult to follow, and the section is a stack of methodologies, making it rather difficult to understand how you used this literature to develop your methodology. I suggest reorganizing it to highlight the various improvements in the context of workload prediction, the important points you decided to retain, and, conversely, the areas for improvement you decided to work on.

Response

In the revised manuscript, we have streamlined the Related Work section by removing parts less relevant to our study (particularly EEG-focused work). In addition, we have added a new paragraph to the Introduction to provide a clearer overview of existing approaches to workload prediction, and explain how these informed our methodological choices. This addition makes the motivation for our study more explicit and clarifies how our work builds on and extends previous research. These approaches and their methodological implications are discussed in greater detail in the Related Work section.

- Line 77: CHS is used whereas it is only defined later.

Response

Fixed.

Data Collection:

- I appreciated the justification of the choice of the CHS over other metrics. Though the CHS score is "human-based", it seems that the questions for assessment are sufficiently well-defined so that the choice of the rating is sharp enough and doesn't depend on the person conducting the evaluation. This noise reduction is further improved thanks to the 3 level aggregation that prevents having scenarios that can be classified in two different categories.

Response

We agree that the CHS, together with the three-level aggregation, helps reduce subjectivity and noise in workload classification.

- I understood that the workload self-assessment is performed during the simulation runs. Do you think that the fact ATCOs have to answer your question on top of working might introduce noise in the physiological data you collect ?

Response

The workload self-assessments were indeed collected during the simulation runs; however, ATCOs only had to briefly state their score in response to an audio signal, which kept the interruption minimal. All participants were briefed beforehand on how to provide their responses, and they were not required to answer if they preferred to skip a prompt. We therefore consider the risk of introducing noise into the physiological data to be very limited. Moreover, all participants followed the same procedure, and the three-minute aggregation window further reduces sensitivity to any short-term disturbance. To make this clear to readers, we have expanded the description in the Subsection 3.5 Simulation-Study Setup to explain the assessment procedure and note its negligible impact on physiological measurements.

Machine Learning Approach:

- This section is very long, and thus quite difficult to read. I would suggest to split it into two sections: one describing the methodology, and one expliciting the results.

Response

We understand the intention behind splitting the Machine Learning Approach section into methodology and results; however, we believe that keeping them together provides greater clarity. Our current structure was designed to present the methodological choices alongside their corresponding outcomes, enabling readers to understand the reasoning behind each step and how it impacts the results. This approach maintains a clear narrative flow from problem formulation to evaluation.

Separating methodology from results would require readers to constantly switch between sections to connect specific decisions with their outcomes, which could make the paper harder to follow. Furthermore, the techniques we used are well-established and generally do not require a dedicated section on their own. Since our primary contribution lies in demonstrating the effectiveness of the chosen approach rather than introducing an entirely new algorithm, keeping these elements together helps maintain clarity and highlights the rationale behind our findings.

That said, to address the reviewer's concern that the section combined too many elements, we revised the structure by moving the last three subsections into a separate section entitled Feature Selection. We believe this strikes a balance between addressing the reviewer's suggestion and maintaining the clarity of the narrative.

- Section 4.1 could be presented as a table (title, type, description). I would also explicitly state that they are time series data.

Response

In the revised manuscript, Section 4.1 has been reformatted into a table that lists each variable together with its type and description (Table 2). We have also clarified in the accompanying text that all variables are recorded as time series, capturing both discrete events (saccades, fixations, blinks) and continuous signals (pupil diameter, blink dynamics, head rotations). We believe these changes improve both readability and clarity.

- The features that are computed are only simple statistical descriptors of the variables that are collected (mean, median, std). Have you ever tried computing other features like avg fixation duration, avg saccade amplitude, avg gaze velocity / acceleration, etc... Moreover, did you try more complex time series features such as time fromain or frequency domain features (cf. TSFresh library)? Finally, why did you only focus on scalar data models and did not explore sequence models (e.g. LSTM) ?

Response

In the present study, we restricted our analysis to simple statistical descriptors (mean, median, standard deviation) in order to provide a first systematic investigation of eye-tracking features for ML-based workload prediction in the ATC context. We agree that additional features such as saccade amplitude, or gaze velocity could provide further insights. Regarding sequence models, we conducted preliminary experiments with LSTM networks; however, these did not yield satisfactory performance compared to the scalar feature models reported here. Exploring such models with larger datasets remains an important avenue for future work. To acknowledge this point, we have expanded the end of the Conclusions and Outlook section to outline these directions in more detail.

Finally, with respect to frequency-domain analysis, we have explored this approach in a related study, Discrete-Fourier-Transform-Based Evaluation of Physiological Measures as Workload Indicators (Lemetti et al., 2023, DASC). While those results were promising in certain cases (e.g., fixation duration), the scope of the current paper is focused on statistical descriptors.

- The difference between the ATCO simulation runs corresponding to three different task load scenarios and the labeling of the data in section 4.3 is slightly misleading. For a given simulation run (e.g. high workload), there can be data points that correspond to low to medium workloads ? Table 1 confused me at first. Moreover, if the CHS score is assessed every 3 minutes during the simulations, don't you fear it introduces noise in the data you collect ?

Response

We agree that the distinction between scenario design and data labeling needs clarification. The task-load scenarios (light, moderate, heavy) were designed to differ in their overall complexity, but within each simulation run, the CHS-based labeling allows for variation: for example, a "high task-load" scenario may still include intervals labeled as medium or even low workload, depending on the ATCO's subjective rating at that time. We have clarified this point in Section 4.3 to avoid confusion.

Regarding the second point, the CHS scores were collected every three minutes during simulations, which introduces the possibility of noise. However, the impact is limited for two reasons: (i) ATCOs only had to briefly state their score in response to an audio signal, which minimized disruption, and (ii) the three-minute aggregation windows smooth short-term fluctuations. We have made this clearer in the revised manuscript.

- If I understand well, the cross-validation sets are made by shuffling the data points collected for all scenarios and all participants, while maintaining the class distribution. In my experience, it's often a good practice to perform a "participant-based" cross validation, so that you don't have data points belonging to one participant in both the training and validation sets. It avoids data leakage, in the sense that the machine learning model won't overfit the specificities of one particular participant. As a result, the model should be more robust to the different behaviors that can be observed between the participants. Could you discuss this ?

Response

The reviewer is correct that we employed a subject-dependent cross-validation strategy, which may overestimate performance compared to participant-based (subject-independent) cross-validation. We now clarify this in the manuscript and highlight in the Conclusions and Outlook section that subject-independent and subject-specific cross-validation approaches will be explored to better assess model generalizability. At the same time, we consider the subject-dependent setup a useful baseline for establishing the predictive potential of eye-tracking features in this context.

- Why not testing the Lasso regression (instead of a regular LR) for feature selection, and comparing with RFE results ?

Response

We agree that applying Lasso regression for feature selection and comparing it with the RFE results would provide useful insights. While this was not implemented in the current study, we now mention it explicitly in the Conclusions and Outlook section as a promising direction for further analysis.

- Table 2: are those metrics representing the average over the cross-validation, or the best fold ?

Response

The metrics in Table 4 (former Table 2) are averaged over the cross-validation folds. Clarification is added to Section 4.7.

- The description of SET1/2/3 could be improved with a table.

Response

The three sets are summarized in Table 8 now.

Discussion: - The discussion of the results and the comparison with similar studies is very interesting. However, in my opinion, I think the description of the different ML models is too detailed, and seems too generic regarding the main goal of your study.

Response

We have removed the detailed model descriptions and now include only a concise discussion focused on the kNN and tree-based models, ensuring stronger alignment with the study objectives.

4. Review - round 2

4.1 Reviewer 1

The authors have addressed all the comments I have made for the first review. In my opinion, the manuscript is much clearer and the findings better highlighted. I especially liked the clear identifications of the limitations of the study, and the future works. I think this paper provides valuable insights for future research that aim at working with eye-tracking data in an ATC environment.