

European Association for Aviation Psychology Conference EAAP 35

Predicting Air Traffic Controller Workload using Machine Learning with a Reduced Set of Eye-Tracking Features

Anastasia Lemetti^{a,*}, Lothar Meyer^b, Maximilian Peukert^b, Tatiana Polishchuk^a,
Christiane Schmidt^a, Helene Alpfjord Wylde^b

^aCommunications and Transport Systems, ITN, Linköping University, Sweden

^bAir Navigation Services of Sweden (LFV), Research & Innovation, Sweden

Abstract

In this paper, we examine the feasibility of assessing air traffic controller (ATCO) workload (WL) using non-intrusive eye-tracking measures and machine learning (ML) algorithms. We concurrently acquire electroencephalography (EEG) data from a workload-optimized wearable device and subjective WL assessments through self-reported Cooper-Harper scale (CHS) workload-rating scores, employing both as label variables. A sample of $n = 18$ ATCOs participate in simulated work sessions encompassing tasks designed to induce three distinct task-load levels: light, moderate, and heavy. We evaluate the performance of five classical ML models. Focusing on the best-performing models, we apply feature selection techniques to identify reduced sets of eye-tracking features. Starting with 58 features, we use a recursive elimination method based on permutation importance, aiming to determine the minimal feature set while also striving for improved performance. The outcomes yielded promising results in the realm of workload-level estimation, achieving 96% accuracy (f1-score=0.87) with 34 features for high workload prediction and 88% accuracy (f1-score=0.82) using 57 features in predicting 3 different levels of workload. We further reduced the feature sets to 6-13 features for different tasks with minimal impact on performance. We identified a “knee point” as the optimal balance between model performance and dimensionality. Adding more features beyond this point did little to improve performance, but increased model complexity. These results indicate that even a small number (less than 10) of features can be sufficient for WL prediction.

© 2024 The Authors. Published by ELSEVIER B.V.

This is an open access article under the CC BY-NC-ND license (<https://creativecommons.org/licenses/by-nc-nd/4.0>)

Peer-review under responsibility of the scientific committee of the European Association for Aviation Psychology Conference EAAP 35

Keywords: ATCO; Workload; Eye Tracking; EEG; Machine Learning

1. Introduction

The mental WL of operators in socio-technical systems has long been identified as a crucial factor for safety and capacity. Hart and Staveland (1988) define WL as “a hypothetical construct that represents the cost incurred by a

* Corresponding author. Tel.: +46-76-285-0401

E-mail address: anastasia.lemetti@liu.se

human operator to achieve a particular level of performance”. To ensure optimal working conditions and performance commitment in ATC, ATCO WL should be kept at a moderate level - as indicated by the inverted-*U* theory of Yerkes and Dodson (1908). It is considered a non-linear multi-dimensional function with traffic load being only one variable, see Josefsson et al. (2020) and Meyer, Peukert, et al. (2022). This involves traffic complexity but also individual factors, such as experience, age and training. The level of WL may therefore vary even though the traffic situation appears very similar.

For this reason, the monitoring of WL should be considered a vital safety barrier. Consequently, ATCOs have a mandated duty to cease operations when their task-performance abilities are in doubt (European Aviation Safety Agency, 2015). However, self-assessment of WL is fraught with uncertainties, as concluded by Gopher and Donchin (1986, pp. 14–2) “... an operator is often an unreliable and invalid measuring instrument.” This is due to limitations in self-awareness and the challenges that individuals face in accurately measuring their WL due to their understanding of their own capacity (Cain, 2007; Hockey, 2003).

To address these limitations, recent studies focus on evidence-based support for WL assessment, e.g. Zak et al. (2020) and Zamarreño Suárez et al. (2022). This may involve physiological or behavioural indicators, which improve the accuracy of self-assessment and facilitate fact-based operational decisions, such as control sector opening/closing. A relevant concern with measurement of indicators is the possibility of disruption of ATCO’s work, which requires any monitoring system to be non-intrusive to ensure continued operational safety.

Following on from our findings (Lemetti et al., 2023; Meyer, Klang, et al., 2022), we consider non-intrusive eye tracking and head movement data to be promising candidates for predicting WL using machine learning (ML) methods. Eye-tracking generates a large number of features (NoF), but including all of them in ML models may not necessarily improve prediction accuracy. While it is generally believed that more features can improve the accuracy of statistical classifiers, too many features can lead to overfitting, increased computational complexity, and multicollinearity, which can reduce accuracy and reliability—a phenomenon known as the *curse of dimensionality* or “Hughes phenomenon” Hughes (1968). Therefore, it is crucial to identify which features are most relevant for predicting WL, as this enhances the interpretability of the results.

We hypothesize that accuracy may improve with a reduced, more significant set of features (best performance). We also hypothesize that there exists the optimal set of features that balances model performance and the NoF used (best economy). In our previous study Lemetti et al. (2024), we developed and evaluated five ML models to predict ATCO WL based on eye-tracking statistical features (summarized in Subsection 4.1). In this work, we extend our approach by incorporating a comprehensive feature selection process to identify a reduced set of features as hypothesized (Subsection 4.2). Additionally, we compare the models using only ocular features, to assess the importance of head movements.

2. Theoretical Background

Efforts to predict WL can succeed in the simplest case via a description of WL using explanatory models with defined causes and related effects. An example is Cañas et al. (2019), which proposes a psychological model that considers the dynamic allocation of cognitive resources influenced by factors such as engagement and effort. This comprehensive model helps to predict automation’s effects on ATCO performance and to develop strategies to mitigate negative outcomes. Another model-based approach, the Cognitive Complexity computerized model—the Cognitive Model for aTco workload Assessment (COMETA) model, has been developed by de Frutos et al. (2019). The model incorporates ATC events and the ATCO task model composed of ATC actions, and outputs the cognitive demand/mental WL of the tasks performed. CHS values were used for the calibration and validation of the model. Ibáñez-Gijón et al. (2023) validated an expanded COMETA model (they included the online effects of the controllers’ actions on the state of the airspace), recording Instantaneous Self Assessment (ISA) and NASA-TLX values, electrodermal activity, and heart rate.

The identification of metrics and related features that support valid indication of WL is an important element in such model developments. This was tackled by Zamarreño Suárez et al. (2022), who aimed to identify parameters for setting WL thresholds that support ATCOs working under safe and efficient conditions. They designed scenarios with overload conditions and used EEG and eye-tracking data to measure WL. They found the number of blinks to be the most significant eye-tracking parameter and confirmed that EEG metrics related to engagement, stress, and

focus correlated with scenario difficulty. They suggested using these variables to establish brain activity thresholds. A number of behaviour-based variables were identified by Josefsson et al. (2020), Meyer, Peukert, et al. (2022) and Lemetti et al. (2023) for ATC tower and en-route, such as blink rate and duration.

While the models presented above provide explanatory power for WL arousal, ML provides a new type of analysis shifting the focuses solely on predictive models. A pioneer of this idea, Zak et al. (2020), suggested a real-time WL assessment method using the Subjective WL Assessment Technique and ML models based on joystick interaction data from experienced military unmanned-aerial-vehicle operators. The study found a high correlation between joystick movement patterns and assessment scores, demonstrating the feasibility of this approach for real-time WL assessment.

The idea of predicting WL by combining electroencephalography (EEG) measurements and ML algorithms was addressed, among others, by Sciaraffa et al. (2019). They aimed to distinguish three levels of ATCO WL using EEG measurements and compared five machine-learning techniques. They concluded that ML can accurately distinguish WL levels based solely on EEG data. Zhou et al. (2022) used unsupervised domain adaptation to build an EEG-based cross-task WL recognition model using different machine-learning approaches (e.g., support vector machine (SVM)). Zhou et al. (2023) recently expanded this analysis to cross-subject WL recognition. Zhu et al. (2023) showed that machine-learning models can effectively identify the WL of pilots, based on EEG signals of six participants, and that the best approach could identify WL with an average accuracy of 90.5%. Safari et al. (2024) employed an open-access EEG data set of 48 users, applied feature selection on 150 features and then seven cross-validations on four machine-learning approaches and could classify WL levels with an SVM with an accuracy of 80.53%.

Kaczorowska et al. (2021) extended the idea of objective measurement and ML to eye-tracking as an elementary lab study. They analyzed eye-tracking data from 29 volunteers performing a computerized cognitive test with eight three-class classification ML models. In addition, they employed an interpretable model to gauge feature importance and the contribution to cognitive WL in a generic work environment. The authors were finally able to achieve high accuracy (97%) in classifying WL levels using only a reduced set of seven key features with this model.

3. Methods and Data Collection

We conducted a controlled lab study in the Area Control Center (ACC)/radar simulation environment Topsky with three en-route scenarios under varying task-load levels: light, moderate, and heavy. Each participant was exposed to the conditions of the scenario for 35 to 45 minutes in the simulation.

In total, we consider $n = 18$ licensed ATCOs (9 female, 9 male) with valid unit endorsements, performing three simulation runs each (one in each task-load condition). Hence, we have 54 runs in total. Eye-tracking data was successfully recorded in 53 runs (98.1%) and EEG data in 51 runs (94.4%). For all 54 runs, we recorded the CHS values. The participants had a mean age of $M = 46.1$ years ($SD = 8.3$), with a mean work experience of $M = 19.7$ years ($SD = 8.4$).

As dependent variables, we collected eye and head movement measurements using eye-tracking. We deployed the monitor-mounted eye-tracking system Smarteye XO with two infrared cameras and a 250 Hz sampling rate. We assume the system to capture a representative set of physiological and behavioural WL candidates features, covering the WL response to different levels of task-load as found in Meyer, Peukert, et al. (2022) and Lemetti et al. (2023). This involves head position and movement, eyelid movements, and eye-gaze data, providing measures such as head heading, pitch, and roll, pupil diameter, saccades, fixations, as well as eyelid closing/opening amplitude and speed.

Subjective WL ratings and EEG data are both captured as reference WL response data. For EEG data, we choose wireless gel-free passive EEG-based Brain-Computer Interfaces (BCIs), which hold promise for noninvasive assessment of human cognitive states. To explain, conventional passive BCIs often suffer from limitations in user comfort and data reliability, hindering their real-world implementation. To address this challenge, Sciaraffa et al. (2022) developed the Mindtooth Touch manufactured by Brainsigns, a wearable and portable semi-dry EEG device optimized for monitoring basic cognitive states. The device offers a 125 Hz sampling rate and outputs “workload”, “vigilance”, and “stress” neurometric variables. We consider Mindtooth to provide valid WL indications by using the neurometric “workload” as validated in Sciaraffa et al. (2022). Using EEG data as a label variable lies in contrast to Sciaraffa et al. (2019) who used EEG as a predictor variable instead. For the subjective WL rating, we used a ten-point adapted Cooper-Harper Scale (CHS) (see Josefsson et al., 2020; Papenfuss and Peters, 2012). We chose this scale over the five-point ISA due to concerns about granularity, anchoring bias, and social desirability associated with verbal scales (Jor-

dan & Brennan, 1992). Although both scales have limitations, the CHS offers finer resolution, potentially mitigating these issues, particularly for distinguishing WL levels across three scenarios. Notably, all values of 4 and above (indicating "solvable by capacity-reducing measures") were categorized as high WL based on post-trial discussions with participants. We queried the verbal WL self assessment on the CHS every three minutes (triggered by an audio signal).

4. Machine Learning Approach

4.1. Model Building

In this subsection, we briefly describe the methodology from Lemetti et al. (2024). First, we segmented the reference WL data into non-overlapping time windows of 3 minutes for CHS-based labels and 1 minute average for the EEG-derived WL variable. This created 667 and 1731 segments for CHS and EEG respectively as ML target variables, each representing a separate analysis case. As predictor variables, we calculated statistical summary features (see 4.2) from eye-tracking using the corresponding window length of each target variable.

We employed both binary and three-class WL classification tasks. We defined a three-class CHS-based WL classification scheme: low (CHS score = 1), medium (CHS score $\in \{2, 3\}$) and high (CHS score ≥ 4). In binary classification, the low and medium classes were combined. The distribution of CHS scores revealed 52% of data points in the "low" class and a cumulative 93% for "low" and "medium" classes combined. The percentiles derived from this distribution (52nd for the "low" class, and 93rd for the combined "low" and "medium" classes) were applied to the EEG WL variable for classification. For binary classification (high vs. medium/low), the 93rd percentile of the EEG data served as the threshold.

We applied five classical ML models to test their effectiveness in estimating the WL level of ATCOs: logistic regression (LR), decision tree (DT), random forest (RF), support vector classifier (SVC), and histogram-based gradient boosting classification tree (HGBC). We implemented the classification algorithms using Python programming language and the *scikit-learn* library.

To address class imbalance between the majority (low/medium WL) and minority (high WL) classes, we employed class weights inversely proportional to class frequencies during training. We used a hold-out validation strategy (90%/10% train/test split) to ensure an unbiased evaluation. To mitigate the influence of scale difference, we performed feature normalization (separately in training and test sets to avoid data leakage). Similarly, EEG WL variable values were transformed to WL classes based on the thresholds obtained from the training set. To optimize models hyperparameters, we implemented Randomized Search with 5-fold cross-validation.

We evaluated the performance of the ML model using a two-pronged approach: *accuracy* and *macro F1-score* metrics. Accuracy provides a baseline assessment of the overall prediction correctness. The macro F1-score is another performance measure, which considers both precision as well as recall, and addresses the potential bias of class imbalance in the ATCO WL prediction dataset. This approach ensures robust evaluation across all WL categories. For both performance metrics, higher values indicate a better performance. More detailed explanations of the applied performance metrics and techniques can be found, for instance, in Lee, 2019.

4.2. Feature Selection

The initial data set comprised 15 eye-tracking and head-movement variables encompassing saccade duration, fixation duration, pupil diameter (left and right), blink amplitude and speed (opening and closing for both eyes), head heading, pitch, and roll. For each variable, we calculate a set of five descriptive statistics across the designated time slot: mean, standard deviation, median, maximum, and minimum. Additionally, we include the total number of saccades. This yields a total of 76 features for subsequent analysis (as opposed to the initial set of 79 features in Lemetti et al. (2024)).

Statistical analysis identified features with zero standard deviation, indicative of no variation across data points. These features were excluded to avoid introducing noise: right/left blink opening/closing speed and amplitude minimum, right/left blink opening/closing speed and amplitude median, and saccade/fixation duration minimum. This initial culling resulted in a set of 58 features.

4.2.1. Evaluation and Selection of Models

Table 1: Algorithms' accuracies in classifying three and two levels of ATCO WL using 58 features

Labels	Classes	LR	SVC	DT	RF	HGBC
CHS	1 / 2–3 / ≥ 4	0.75	0.84	0.63	0.67	0.85
EEG	Percentiles: 0.52, 0.93	0.73	0.82	0.68	0.82	0.84
CHS	1–3 / ≥ 4	0.81	0.96	0.87	0.9	0.93
EEG	Percentile: 0.93	0.78	0.88	0.9	0.96	0.95

Table 2: Algorithms' macro F1-scores in classifying three and two levels of ATCO WL using 58 features

Labels	Classes	LR	SVC	DT	RF	HGBC
CHS	1 / 2–3 / ≥ 4	0.6	0.66	0.59	0.63	0.73
EEG	Percentiles: 0.52, 0.93	0.61	0.65	0.55	0.68	0.77
CHS	1–3 / ≥ 4	0.63	0.87	0.62	0.7	0.75
EEG	Percentile: 0.93	0.6	0.68	0.7	0.84	0.83

Tables 1 and 2 present the performance of the five chosen ML models on the unseen test set, using the initial set of 58 features. These tables report two key metrics: accuracy and macro F1-score. Additionally, each table details the classes and their corresponding labels used in the specific classification task. The model achieving the highest performance (either in terms of accuracy or F1-score) for each classification attempt is highlighted in bold.

The results revealed that HGBC and SVC achieved superior performance in predicting WL levels based on CHS. For WL level prediction using EEG data, HGBC and RF models demonstrated significantly better results compared to other evaluated models. Notably, HGBC consistently emerged as a strong contender for both multiclass classification tasks, suggesting its potential as a robust modeling approach. Consequently, HGBC and either SVC or RF (depending on the label) were selected for further analysis with feature selection techniques.

4.2.2. Wrapper Methods

Wrapper methods employ a greedy search strategy for feature selection. They iteratively evaluate subsets of features by training a ML model and assess its performance on a held-out validation set. Based on pre-defined stopping criteria, such as a decrease in model performance or reaching a desired NoF, features are added or removed from the subset.

We use the technique of **Recursive Feature Elimination** (RFE) with permutation importance. This approach iteratively removes the least informative feature based on its impact on model performance. RFE starts with the full set of features and trains a model. Then, it calculates the importance of each feature using a method like permutation importance. The feature with the least impact on the model's performance (e.g., F1-score) is removed. This process repeats until a desired NoF remains or a pre-defined stopping criterion is met.

Permutation importance is a model-agnostic technique that assesses a feature's importance by randomly shuffling its values in the dataset. The model's performance is then measured on this shuffled data. The decrease in performance compared to the original data indicates the feature's importance. A larger decrease suggests that the model relied heavily on that feature for accurate predictions.

After identifying the optimal models for each task (see 4.2.1), we performed RFE on these models, starting with 58 features and iteratively removing those with the least impact on performance until only one feature remained. The selection process was visualized by plotting the model's metrics (accuracy and F1-score) against the number of remaining features, revealing the impact of feature reduction on model performance (see Figure 1 for the EEG-labeled three-classes case example).

We chose the NoF according to two criteria: "best economy" and "best performance". The criterion "best economy" is defined as the best balance between model dimensionality (i.e., fewer features) and performance (i.e., higher accuracy and F1-score). The identification of this optimal subset of features involves locating the "knee point" on a performance curve. The knee point is defined as the "relative cost to increase some tunable parameter is no longer worth the corresponding performance benefit" (Satopaa et al., 2011). We determine the knee point by finding the point with the maximum distance from a line passing between the first and last points. This approach does not impose any assumptions about the curve's shape, making it more robust compared to other, more complex methods. The criterion "best performance" assumes choosing the NoF that corresponds to the maximum value of the model performance met-

ric (accuracy and F1-score). Table 3 presents the results for both criteria. Notably, it is possible to achieve a modest reduction in feature dimensionality while nearly preserving performance across all four tasks.

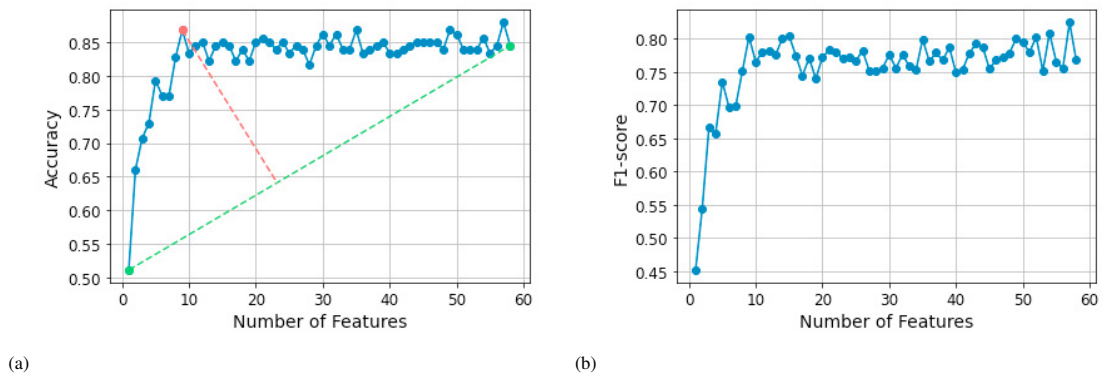


Fig. 1: Accuracy (a) and F1-score (b) vs. NoF for the 3-classes classification task labeled with EEG (HGBC model). (a) illustrates the “best economy” criterion: the red dashed line represents the distance from the “knee point” to the green dashed line, which connects the first and last points of the curve. Scaling factors were applied to accurately represent perpendicular distance, accounting for the differing x and y axis ranges.

Table 3: Models’ NoF and corresponding accuracies and macro F1-scores in RFE process for “best economy” and “best performane” criteria using ocular and head movements features

Labels	Classes	Model	NoF	Accuracy best economy	F1-score	NoF	Accuracy best performance	F1-score
CHS	3-classes	HGBC	9	0.82	0.7	48/54	0.87/0.85	0.74/0.78
EEG	3-classes	HGBC	9	0.87	0.8	57	0.88	0.82
CHS	binary	SVC	13	0.94	0.82	34	0.96	0.87
EEG	binary	RF	6	0.94	0.78	51	0.96	0.84

4.2.3. Ocular Features

We hypothesize that models utilizing solely ocular features can achieve performance comparable to those incorporating head-movement data. We achieve feature selection through a two-step process. First, we eliminate all head-movement features, resulting in a set of 43 solely ocular features. Subsequently, we replicate the feature selection procedure outlined in Section 4.2.2 on this reduced feature set. The results obtained are presented in Table 4. The results show that using the reduced set of ocular features, we achieve performance comparable to the original set.

Table 4: Models’ NoF and corresponding accuracies and macro F1-scores in RFE process for “best economy” and “best performane” criteria using ocular set of features

Labels	Classes	Model	NoF	Accuracy best economy	F1-score	NoF	Accuracy best performance	F1-score
CHS	3-classes	HGBC	6	0.82	0.74	39	0.85	0.78
EEG	3-classes	HGBC	13	0.8	0.77	31/13	0.83/0.8	0.76/0.77
CHS	binary	SVC	6	0.94	0.86	6	0.94	0.86
EEG	binary	RF	10	0.94	0.78	26	0.94	0.79

5. Discussion and Conclusion

We demonstrate the significant potential of ML techniques for predicting ATCO WL. Our findings highlight the efficacy of using eye-tracking data, either in conjunction with head-movement data or independently, for WL prediction. This paves the way for the development of a non-intrusive WL monitoring system. Such a system could be instrumental in identifying not only overload conditions, but also underload situations.

We effectively addressed the issue of feature dimensionality by using the Recursive Feature Elimination (RFE) method, which allowed us to reduce the NoF based on their impact on model performance. The relationship between the NoF and prediction performance showed that the curse of dimensionality was not a concern for our model when a large NoF were included. However, at a lower NoF, increasing the NoF improved model performance almost linearly. We observed a critical threshold, or “saturation point”, beyond which adding more features did not systematically enhance model performance, rather adding noise. By applying the “best performance” and the “best economy” criteria, we identified reduced feature sets that maintained model performance in both cases (EEG and CHS labels). We identified the knee point at “best economy” to be a good indicator for the “saturation point”.

In 14 out of 16 reduction processes, the performance metrics—accuracy and F1-score—identified the same set of features, indicating consistency between these metrics. It was observed that point of the “best performance” was identified in an interval of high variance beyond the saturation point. This might indicate an unstable result. However, NoF related to the “best economy” shows larger robustness against variations in-between tasks. The scenario with head-movements revealed 9 and 13 features for CHS labeled data at “best economy”. In turn, the corresponding case using ocular features only reveal 6 features. Against all expectation, this was not accompanied by a decrease in performance rather model performance remains the same or was even higher.

Models using EEG label data showed similar performance to those using CHS labels, but required more features, making EEG-based WL prediction less efficient than CHS-based WL. To explain, using EEG-based WL as label data is probably the more complex choice, as EEG-based WL itself is based on ML model prediction, as described in Sciaraffa et al. (2022), making the prediction more indirect and thus more complex.

Overall, we conclude that reducing the NoF generally results in a slight decrease in model performance, but significantly reduces measurement and computational efforts. The most favorable scenario was the CHS case with three classes, where only 6 features were needed instead of 58, with an accuracy of 0.82 compared to 0.85 and an F1-score of 0.74 compared to 0.73.

6. Future Work

While the current study employed 1-minute time slots for EEG labeling, adjusting these intervals could potentially enhance model performance in classifying ATCO WL. In future studies, it would be prudent to explore the impact of varying time slots for segmenting EEG data.

So far, we employ a subject-dependent strategy, which means that models are trained and evaluated on shuffled data from all participants. Future investigations could explore a subject-independent approach, where each participant’s data is segregated exclusively for either the training or test set, which will make the evaluation more rigorous and elucidate the model’s generalizability to new users. Additionally, a subject-specific approach merits exploration, where models are trained and evaluated solely on data from a single participant. This allows us to assess whether participant-specific models yield superior performance in WL prediction compared to models trained on multi-subject data.

Acknowledgments, Ethics and Funding

We would like to acknowledge all involved persons of LFV en-route at ATCC Malmö, in particular ATCOs and pseudo-pilots involved in the experimental data collection. The study was considered ethically safe by the Swedish Ethical Review Authority (2023-06110-01). This study was supported in the scope of project On WorkLoad Measures (OWL), funded by Swedish Transport Administration (TRV 2022/33636r).

References

- Cain, B. (2007). *A review of the mental workload literature* (tech. rep.). Defence Research; Development Canada, Toronto.
- Cañas, J. J., Ferreira, P., de Frutos, P. L., Puntero, E., López, E., Gómez-Comendador, F., De Crescenzo, F., Lucchi, F., Netjasov, F., & Mirkovic, B. (2019). Mental workload in the explanation of automation effects on ATC performance. *Human Mental Workload: Models and Applications: Second International Symposium, H-WORKLOAD 2018, Amsterdam, The Netherlands, September 20-21, 2018, Revised Selected Papers 2*, 202–221.

- de Frutos, P. L., Rodríguez, R. R., Zhang, D. Z., Zheng, S., Cañas, J. J., & Muñoz-de-Escalona, E. (2019). Cometa: An air traffic controller's mental workload model for calculating and predicting demand and capacity balancing. *Human Mental Workload: Models and Applications: Third International Symposium, H-WORKLOAD 2019, Rome, Italy, November 14–15, 2019, Proceedings 3*, 85–104.
- European Aviation Safety Agency. (2015). Commission regulation (eu) 2015/340 [Accessed: 2024-07-25]. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32015R0340>
- Gopher, D., & Donchin, E. (1986). Workload—an examination of the concept. In K. Boff, L. Kaufman, & J. Thomas (Eds.), *Cognitive processes and performance*. (Vol. 2). John Wiley & Sons, Inc.
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139–183). Elsevier.
- Hockey, G. (2003). Operator functional state as a framework for the assessment of performance degradation. *Operator Functional State*, 8–23.
- Hughes, G. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, 14(1), 55–63. <https://doi.org/10.1109/TIT.1968.1054102>
- Ibáñez-Gijón, J., Travieso, D., Navia, J. A., Montes, A., Jacobs, D. M., & Frutos, P. L. (2023). Experimental validation of COMETA model of mental workload in air traffic control. *Journal of Air Transport Management*, 108, 102378. <https://doi.org/10.1016/j.jairtraman.2023.102378>
- Jordan, C. S., & Brennan, S. D. (1992). *An experimental report on rating scale descriptor sets for the instantaneous self-assessment (ISA) recorder* (tech. rep.). DRA.
- Josefsson, B., Meyer, L., Peukert, M., Polishchuk, T., & Schmidt, C. (2020). Validation of controller workload predictors at conventional and remote towers. *9th International Conference on Research in Air Transportation*.
- Kaczorowska, M., Plechawska-Wójcik, M., & Tokovarov, M. (2021). Interpretable machine learning models for three-way classification of cognitive workload levels for eye-tracking features. *Brain sciences*, 11(2), 210.
- Lee, W.-M. (2019). *Python machine learning*. John Wiley & Sons, Inc.
- Lemetti, A., Meyer, L., Peukert, M., Polishchuk, T., & Schmidt, C. (2023). 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. *DASC*.
- Lemetti, A., Meyer, L., Peukert, M., Polishchuk, T., Schmidt, C., & Alpfjord Wylde, H. (2024). Eye in the sky: Predicting air traffic controller workload through eye tracking based machine learning [To appear in DASC].
- Meyer, L., Klang, K. J., Boonsong, S., Westin, C., Nordman, A., Lundberg, J., Josefsson, B., & Vrotsou, K. Mapping the decision-making process of conflict detection and resolution in en-route control: An eye-tracking based approach. In: *Proceedings of 12th sesar innovation days*. SESAR JU. Budapest, Hungary, 2022.
- Meyer, L., Peukert, M., Polishchuk, T., & Schmidt, C. (2022). Investigating ocular and head-yaw measures as indicators for workload and fatigue under varying taskload conditions. *10th International Conference on Research in Air Transportation*.
- Papenfuss, A., & Peters, M. (2012). *HMI laboratory report 8: Analysis of critical situations at remote tower operated airports*.
- Safari, M., Shalhaf, R., Bagherzadeh, S., & Shalhaf, A. (2024). Classification of mental workload using brain connectivity and machine learning on electroencephalogram data. *Scientific Reports*, 14(1), 9153. <https://doi.org/10.1038/s41598-024-59652-w>
- Satopaa, V., Albrecht, J., Irwin, D., & Raghavan, B. (2011). Finding a “kneedle” in a haystack: Detecting knee points in system behavior. *2011 31st international conference on distributed computing systems workshops*, 166–171.
- Sciaraffa, N., Aricò, P., Borghini, G., Flumeri, G. D., Florio, A. D., & Babiloni, F. (2019). On the use of machine learning for EEG-based workload assessment: Algorithms comparison in a realistic task. In L. Longo & M. C. Leva (Eds.), *Human mental workload: Models and applications* (pp. 170–185). Springer International Publishing.
- Sciaraffa, N., Di Flumeri, G., Germano, D., Giorgi, A., Di Florio, A., Borghini, G., Vozzi, A., Ronca, V., Babiloni, F., & Aricò, P. (2022). Evaluation of a new lightweight EEG technology for translational applications of passive brain-computer interfaces. *Frontiers in Human Neuroscience*, 16, 901387. <https://doi.org/10.3389/fnhum.2022.901387>
- Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, 18(5), 459–482. <https://doi.org/10.1002/cne.920180503>
- Zak, Y., Parmet, Y., & Oron-Gilad, T. (2020). Subjective workload assessment technique (SWAT) in real time: Affordable methodology to continuously assess human operators' workload. *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2687–2694. <https://doi.org/10.1109/SMC42975.2020.9283168>
- Zamarreño Suárez, M., Arnaldo Valdes, R. M., Pérez Moreno, F., Delgado-Aguilera Jurado, R., López de Frutos, P. M., & Gomez Comendador, V. F. (2022). How much workload is workload? A human neurophysiological and affective-cognitive performance measurement methodology for ATCOs. *Aircraft Engineering and Aerospace Technology*, 94(9), 1525–1536.
- Zhou, Y., Wang, P., Gong, P., Wei, F., Wen, X., Wu, X., & Zhang, D. (2023). Cross-subject cognitive workload recognition based on EEG and deep domain adaptation. *IEEE Transactions on Instrumentation and Measurement*, 72, 1–12. <https://doi.org/10.1109/TIM.2023.3276515>
- Zhou, Y., Xu, Z., Niu, Y., Wang, P., Wen, X., Wu, X., & Zhang, D. (2022). Cross-task cognitive workload recognition based on EEG and domain adaptation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30, 50–60. <https://doi.org/10.1109/TNSRE.2022.3140456>
- Zhu, W., Zhang, C., Liu, C., Yuan, J., Li, X., Wang, Y., & Jiang, C. (2023). Recognition of pilot mental workload in the simulation operation of carrier-based aircraft using the portable EEG. *Proceedings of the 2023 3rd International Conference on Human Machine Interaction*, 43–49. <https://doi.org/10.1145/3604383.3604391>