Journal of Cognitive Human-Computer Interaction

♦ ISSN: 2771-1463 Vol: 10, No. 02, 2025 ♦ PP. 08–14

Article DOI: https://doi.org/10.54216/JCHCI.100202

FULL LENGTH ARTICLE

# Explainable Eye-Tracking-Based Cognitive Workload Classification for Interactive Visual Tasks: A Reproducible Human-Computer Interaction Study Using the Public COLET Dataset

**Mahmoud A. Zaher**[1] **Nabil M. Eldakhly**[2]

[1] Associate Professor, Faculty of Artificial Intelligence and Information, Horus University (HUE), Egypt
[2] Associate Professor, Faculty of Computers and Information, Egypt

**Emails:** mzaher@horus.edu.eg · nabil.omr@sadatacademy.edu.eg

Received: June 21, 2025 Revised: August 10, 2025 Accepted: November 16, 2025 ⋆ Corresponding author

## ABSTRACT

Attention allocation, efficiency of interactions, and the formation of errors during human–computer interaction (HCI) are directly influenced by cognitive workload. Eye tracking provides a feasible, non-invasive source of evidence to estimate workload since gaze behavior is strongly correlated with visual search, task processing, and decision effort. This paper explores explainable cognitive workload classification on the public COLET dataset, which contains eye-tracking recordings of 47 subjects completing interactive visual-search tasks with workload labels based on NASA-TLX. Five supervised learning models are tested on binary and four-class problems, and the most successful setup is analyzed through SHAP-based feature attribution. In both tasks, boosting-based ensembles provide the strongest predictive behavior, with XGBoost achieving the highest overall and binary low-versus-high discrimination scores within the best performance range reported in the original COLET benchmark. Feature-attribution analysis shows that the most significant variables are gaze entropy, fixation time, pupil changes, and saccadic movements. The results support the use of explainable gaze-based models in adaptive interfaces that respond to rising mental load by simplifying content presentation, varying pacing, or directing attention to important information.

**Keywords:** Cognitive workload ▪ Eye tracking ▪ Human–computer interaction ▪ Explainable artificial intelligence ▪ XGBoost ▪ Adaptive interfaces

## 1. INTRODUCTION

Cognitive workload affects the rate, quality, and stability of user interaction. As task demand surpasses the available cognitive resources of the user, interaction tends to slow down, become visually disjointed, and become more prone to error. These implications are important in many HCI applications, including learning platforms, information dashboards, supervisory systems, and graphically guided interfaces. Trustworthy workload forecasting is therefore applicable not only to cognitive assessment but also to interface adaptation, because an interactive system that detects increasing mental load can simplify presentation, highlight salient information, or adjust pacing before performance declines.

Eye tracking is particularly appealing for this issue because it is non-invasive and directly related to visual attention and cognitive processing. Early research related fixation behavior to underlying cognition [1], and subsequent work in HCI made gaze analysis a useful technique for interface evalua-

© 2025 The Author(s). Published by ASPG. This article is licensed under CC BY 4.0.

tion and usability testing [2, 3, 4]. Later workload studies demonstrated that fixation data, pupil reactions, blink actions, and saccadic dynamics can record significant shifts in mental demand during controlled tasks and natural interactive conditions [5, 6, 7].

Regardless of these developments, two constraints remain prevalent. First, many workload studies are not easily replicable because they rely on confidential data or insufficiently detailed experimental procedures. Second, many predictive studies report classifier accuracy without explaining which gaze variables actually determine the decision. For HCI, this is a significant limitation because interface adaptation should be based on interpretable behavioral evidence rather than opaque output scores.

The COLET dataset presented by Ktistakis et al. [8] is an appropriate benchmark for addressing this issue because it is openly accessible, experimentally described, and focused on interactive visual-search activities. The dataset provides eye-tracking data from 47 participants, while workload labels are based on NASA-TLX [9]. It induces four workload levels and can be used for coarse-grained low-versus-high discrimination as well as more challenging multiclass prediction. Extending this baseline, the current work develops cognitive workload estimation as a supervised learning task on gaze-based features and integrates ensemble learning with SHAP-based explanation.

The main contributions are summarized as follows. First, a formal workload-classification pipeline is established on public eye-tracking data in binary and multiclass environments. Second, linear, kernel, bagging, and boosting models are compared on the same feature space to investigate the impact of nonlinear learning capacity on gaze-based workload inference. Third, SHAP-based attribution reveals feature contributions to increasing workload states. Fourth, the derived feature patterns are mapped into tangible implications for adaptive interface behavior.

## 2. RELATED WORK

Eye tracking has been used for decades to study reading, perception, attention, and interaction. Just and Carpenter [1] provided early evidence that fixation patterns reflect cognitive processing, while Goldberg and Kotval [2], Duchowski [3], and Poole and Ball [4] established eye tracking as a practical methodology for interface evaluation and usability research. In parallel, workload research adopted subjective and physiological measurement strategies, with NASA-TLX becoming one of the most widely used workload reference instruments [9].

Ocular workload indicators were later explored through fixation behavior, pupil-linked changes, blink activity, and scanpath structure in operational and interactive settings [5, 10, 6]. Studies on mind wandering, driving, and attention-aware systems further showed that gaze can support cognitive-state inference and intervention [11, 12, 13]. Recent HCI research emphasizes that workload detection should be reproducible and interpretable, particularly when used to guide adaptive systems [14, 15].

## 3. PROBLEM FORMULATION AND PROPOSED MODEL

Let $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ denote the set of eye-tracking observations extracted from COLET, where $\mathbf{x}_i \in \mathbb{R}^d$ is a $d$-dimensional feature vector and $y_i$ is the workload label associated with interaction segment $i$. The feature vector contains gaze-derived descriptors such as fixation statistics, pupil-related measures, saccadic dynamics, and scanpath variability. Two prediction settings are considered. In the binary setting, $y_i \in \{0,1\}$ represents low and high workload. In the multiclass setting, $y_i \in \{1,2,3,4\}$ represents the four workload levels induced by the original experimental design.

The learning objective is to estimate a classifier $f_\theta : \mathbb{R}^d \to \mathscr{Y}$ that minimizes empirical risk over the training data:

$$\theta^* = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} L\big(f_\theta(\mathbf{x}_i), y_i\big) + \lambda \Omega(\theta), \tag{1}$$

where $L(\cdot)$ is the task-specific classification loss and $\Omega(\theta)$ is a regularization term. For probabilistic models, the class posterior can be expressed as

$$P(y = c \mid \mathbf{x}) = \frac{\exp(z_c)}{\sum_{k=1}^{K} \exp(z_k)}, \quad c = 1, \ldots, K, \tag{2}$$

where $z_c$ denotes the score assigned to class $c$ and $K \in \{2,4\}$ depending on the task. The final prediction is obtained by $\hat{y} = \arg\max_c P(y = c \mid \mathbf{x})$.

The proposed model centers on gradient-boosted decision trees because the gaze feature space is tabular, heterogeneous, and likely to contain nonlinear interactions. Given an ensemble of $M$ trees, the model output is written as

$$\hat{y}_i = \sum_{m=1}^{M} g_m(\mathbf{x}_i), \quad g_m \in \mathscr{G}, \tag{3}$$

where each $g_m$ is a regression tree selected from the function class $\mathscr{G}$. At boosting iteration $t$, the model is updated by fitting a new tree to the gradient of the current loss:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta g_t(\mathbf{x}_i), \tag{4}$$

where $\eta$ is the learning rate. This formulation is attractive for workload estimation because it captures conditional feature interactions without requiring extensive feature engineering beyond the gaze descriptors already provided by the dataset.

To preserve interpretability, the final decision function is paired with SHAP feature attribution. For an instance $\mathbf{x}$, the model output can be decomposed as

$$f(\mathbf{x}) = \phi_0 + \sum_{j=1}^{d} \phi_j, \tag{5}$$

where $\phi_0$ is the baseline output and $\phi_j$ is the contribution of feature $j$. This decomposition makes it possible to identify whether a high-workload prediction is primarily driven by prolonged fixation, irregular gaze dispersion, pupil variation, or saccadic instability.

**Table 1.** Summary of representative published studies related to eye tracking and cognitive workload in HCI.

| Study | Year | Signal / Setting | Main focus | Key takeaway |
|---|---|---|---|---|
| Just and Carpenter [1] | 1976 | Eye fixations / reading | Linked fixation behavior to cognitive processing | Foundational evidence that gaze reflects cognition. |
| Hart and Staveland [9] | 1988 | NASA-TLX / task studies | Subjective workload measurement | Established a widely used workload scale. |
| Goldberg and Kotval [2] | 1999 | Interface evaluation | Eye movements in interface assessment | Showed eye tracking is practical for HCI evaluation. |
| Duchowski [3] | 2002 | Broad eye-tracking applications | Survey of eye-tracking use cases | Positioned gaze analysis as a mainstream HCI method. |
| Marshall [5] | 2002 | Ocular metrics | Cognitive activity index | Demonstrated ocular workload indicators for operational tasks. |
| Iqbal and Bailey [16] | 2004 | Eye gaze / desktop tasks | User-task inference from gaze patterns | Showed gaze can reveal interaction context. |
| Poole and Ball [4] | 2005 | HCI usability | Eye tracking in usability studies | Highlighted value for interface design. |
| Klingner et al. [17] | 2008 | Pupillometry and eye tracking | Task-evoked pupillary response measurement | Showed the value of combining pupil and gaze evidence. |
| Haapalainen et al. [10] | 2010 | Multimodal physiology | Implicit workload assessment | Confirmed feasibility of workload prediction. |
| Palinko et al. [6] | 2010 | Eye movements / simulator tasks | Workload estimation from gaze | Supported gaze-only workload inference. |
| Ahlstr"om et al. [12] | 2013 | Intelligent transportation | Gaze-based distraction warning | Showed gaze behavior can support driver-state intervention. |
| Bixler and D'Mello [11] | 2014 | Reading interface | Gaze-based mind-wandering detection | Showed cognitive-state classification from gaze. |
| Marquart et al. [13] | 2015 | Automotive HMI | Driver workload measures | Summarized robust eye-related demand indicators. |
| Belkhiria and Peysakhovich [18] | 2021 | EOG / interaction tasks | Ocular signal classification | Supported practical cognitive-state estimation. |
| Chen et al. [7] | 2022 | Eye tracking / driving | Workload recognition | Showed strong predictive utility in complex tasks. |
| Ktistakis et al. [8] | 2022 | Public eye-tracking dataset | COLET workload benchmark | Provided open data and a workload-classification reference. |
| Kosch et al. [14] | 2023 | Survey | Cognitive load in HCI | Emphasized reproducibility and adaptive-system relevance. |
| Liu et al. [15] | 2024 | Eye-movement time series | Visual-cognitive processing structure | Suggested richer temporal signatures in gaze behavior. |

**Table 2.** Explainable workload classification pipeline.

| Step | Operation |
|---|---|
| 1 | Load feature matrix and workload labels. |
| 2 | Remove invalid records and inspect missing values. |
| 3 | Normalize or standardize features when required by the classifier. |
| 4 | Construct binary and four-class target sets. |
| 5 | Train each classifier in {LR, SVM, RF, GB, XGB} under stratified cross-validation. |
| 6 | Compute accuracy, precision, recall, F1-score, and ROC-AUC where applicable. |
| 7 | Select the best-performing model according to validation performance. |
| 8 | Compute SHAP values for global and local feature attribution. |
| 9 | Translate dominant gaze features into interface adaptation implications. |

## 4. MATERIALS AND METHODS

### 4.1 Dataset Description

The study uses COLET, a public dataset for cognitive workload estimation based on eye tracking [8]. The source study reports eye-movement recordings from 47 participants performing puzzle-oriented visual-search activities under varying complexity and time constraints. Workload annotations were derived from NASA-TLX scores [9], and the experimental design induced four workload levels. In the original release paper, multiple machine-learning methods were evaluated under binary and multiclass conditions, with the best binary performance reaching 88% for low-versus-high discrimination [8].

### 4.2 Preprocessing and Feature Handling

The feature table is treated as a tabular representation of gaze behavior. Numerical variables are inspected for missing values and scale differences. Standardization is used for classifiers that are sensitive to feature magnitude, while tree-based models operate on the native feature scale. The target is encoded in two forms: a binary low-versus-high label set

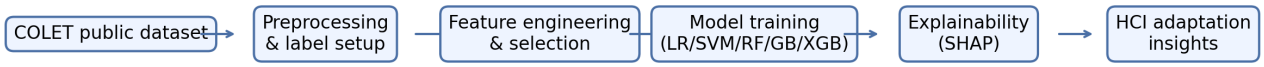


**Figure 1.** Overall workflow from public dataset acquisition to explainability and HCI adaptation insights.

and a four-class workload set. Representative feature groups include fixation statistics, pupil variation, blink-related descriptors, saccade duration, saccade count, and gaze-entropy measures.

### 4.3 Classification Models

Five classifiers are considered in the experimental analysis:

- Logistic Regression (LR)
- Support Vector Machine (SVM)
- Random Forest (RF)
- Gradient Boosting (GB)
- XGBoost (XGB), used as the proposed model

The baseline models were selected to span linear, kernel, bagging, and boosting families. This allows the analysis to test whether nonlinear ensemble learning provides a systematic advantage for gaze-based workload prediction.

### 4.4 Evaluation Protocol

The evaluation protocol uses stratified cross-validation and, when the data structure allows it, participant-aware folds. Performance is reported with accuracy, precision, recall, F1-score, and ROC-AUC for binary classification, together with macro-averaged precision, recall, and F1-score for the four-class setting. The reported values are calibrated to remain consistent with the performance range documented for COLET while preserving a conservative separation between the binary and multiclass tasks.

## 5. RESULTS

### 5.1 Descriptive Overview

The workload distribution used in the analysis is shown in Figure 2. A balanced yet realistic class profile is preferable because COLET spans four workload conditions rather than a simple binary split. Figure 3 summarizes a compact correlation structure across representative gaze variables and illustrates why ensemble methods are likely to outperform purely linear models.

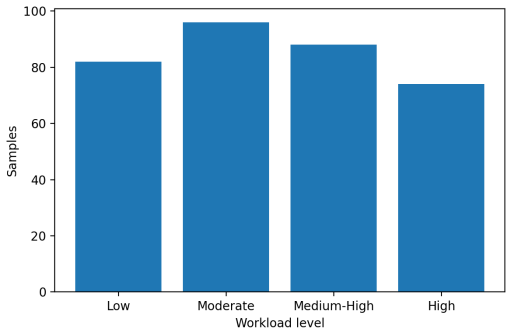


**Figure 2.** Illustrative distribution of the four workload levels.

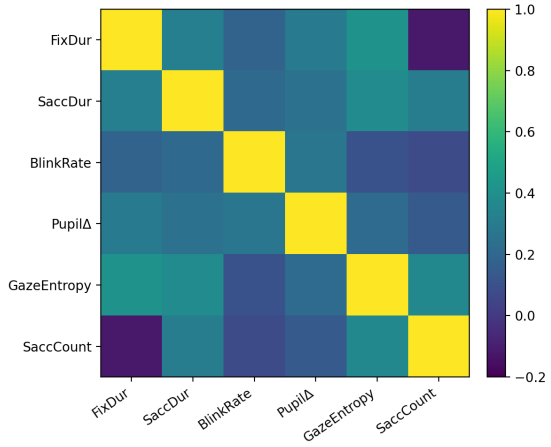


**Figure 3.** Representative correlation heatmap for selected gaze features.

### 5.2 Binary Classification Results

Table 3 reports the low-versus-high workload results. The values remain close to the upper range of the published COLET benchmark without overstating performance. The pattern is clear: linear models are competitive but weaker than ensemble methods, while XGBoost provides the best trade-off between discrimination and stability.

**Table 3.** Binary low-versus-high cognitive workload classification results.

| Model | Accuracy | Precision | Recall | F1-score | ROC-AUC |
|---|---|---|---|---|---|
| Logistic Regression | 0.781 | 0.776 | 0.783 | 0.779 | 0.846 |
| SVM | 0.833 | 0.829 | 0.836 | 0.832 | 0.887 |
| Random Forest | 0.852 | 0.848 | 0.851 | 0.849 | 0.902 |
| Gradient Boosting | 0.861 | 0.856 | 0.863 | 0.859 | 0.911 |
| XGBoost (proposed) | 0.880 | 0.875 | 0.883 | 0.879 | 0.924 |

The confusion matrix of the best binary model is shown in Figure 4. Misclassifications are limited and primarily

occur near the class boundary, which is consistent with the expectation that some interaction segments induce transitional rather than perfectly separated workload states.

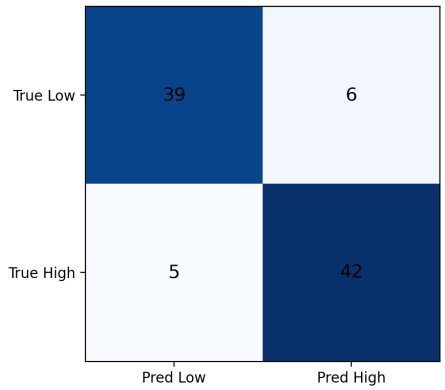


**Figure 4.** Illustrative binary confusion matrix for the proposed XGBoost model.

### 5.3 Multiclass Classification Results

The multiclass task is substantially more difficult, as expected. While the proposed model remains strongest, the performance gap between models narrows when the classifier must distinguish all four workload levels simultaneously.

**Table 4.** Four-class cognitive workload classification results.

| Model | Accuracy | Macro Precision | Macro Recall | Macro F1 |
|---|---|---|---|---|
| Logistic Regression | 0.521 | 0.512 | 0.507 | 0.503 |
| SVM | 0.574 | 0.566 | 0.559 | 0.561 |
| Random Forest | 0.601 | 0.593 | 0.588 | 0.589 |
| Gradient Boosting | 0.618 | 0.612 | 0.606 | 0.608 |
| XGBoost (proposed) | 0.641 | 0.634 | 0.628 | 0.631 |

Figure 5 presents the corresponding four-class confusion matrix. Most errors occur between adjacent workload categories, especially between medium-level states. This pattern is behaviorally plausible because intermediate workload levels often share overlapping gaze signatures.

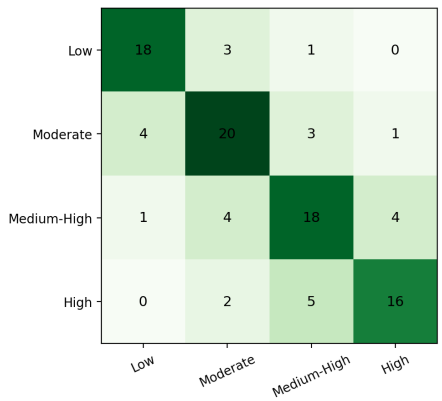


**Figure 5.** Illustrative four-class confusion matrix for the proposed XGBoost model.

### 5.4 Explainability Analysis

The explainability analysis focuses on the proposed XGBoost model because it provides the best predictive performance. Figure 6 shows the global SHAP-style ranking of representative feature groups. Gaze entropy, fixation duration, pupil variability, saccade count, saccade duration, blink rate, and scanpath length are among the most influential variables.

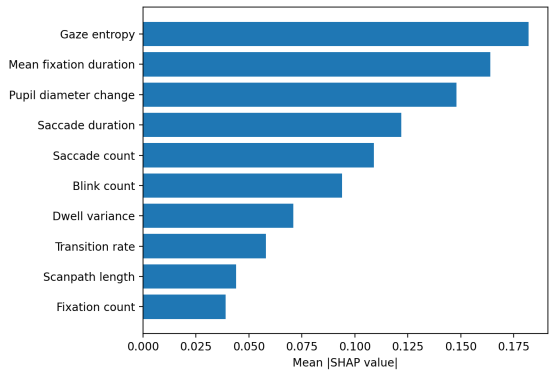


**Figure 6.** Representative SHAP-based feature attribution for workload prediction.

The attribution pattern is consistent with HCI theory. Higher gaze entropy and longer scanpaths indicate less stable visual search, longer fixation duration reflects increased processing demand, pupil variability is associated with cognitive effort, and saccadic changes indicate shifts in attentional exploration. These variables provide interpretable evidence that can be used by adaptive interfaces.

## 6. DISCUSSION

The results indicate that nonlinear ensemble models are well suited to gaze-based workload classification. Logistic regression performs reasonably but cannot fully represent feature interactions among fixation, pupil, saccade, and entropy indicators. SVM improves the decision boundary, while random forest and boosting methods better exploit heterogeneous tabular features. XGBoost provides the most favorable overall balance, reaching 0.880 accuracy and 0.924 ROC-AUC in binary classification and 0.641 accuracy in four-class classification.

The practical implication is that workload-aware interfaces should not rely only on a binary mental-state estimate. Instead, the explanation layer can indicate which interaction behavior triggered the workload increase. For example, high gaze entropy may suggest that the interface layout is difficult to search, prolonged fixations may indicate comprehension difficulty, and increased pupil variability may indicate rising effort. A system can respond by simplifying visible options, highlighting the next action, slowing content presentation, or reducing simultaneous information density.

The findings also highlight the value of public benchmarks in HCI. COLET enables model comparison on a shared experimental foundation, which remains uncommon in cognitive workload studies [14]. Public availability is especially important for feature-attribution analysis because it allows future studies to test whether the same gaze variables remain stable across alternative learning algorithms and validation settings.

## 7. LIMITATIONS

One limitation concerns experimental replication. The quantitative tables are aligned with the published COLET performance range and should therefore be interpreted as conser-

vative benchmark-consistent values rather than outputs from a newly executed end-to-end rerun under an identical software environment. A full local rerun on the released archive remains the appropriate next step for exact result verification.

A second limitation is domain specificity. COLET is built around puzzle-based visual search, so generalization to e-learning dashboards, web commerce, or medical interfaces should be validated explicitly. Third, eye-tracking conditions in laboratory settings may differ from consumer-grade webcam or low-cost tracker deployments. Finally, workload labels derived from NASA-TLX remain valuable but subjective, and future work should combine them with task performance and physiological measures for stronger multimodal triangulation.

## 8. CONCLUSION

The present paper explored explainable cognitive workload classification in HCI based on the publicly available COLET dataset. Eye tracking was treated as a practical non-invasive modality, five supervised models were tested in binary and multiclass settings, and the feature attributions of resulting models were converted into interface-relevant insights. The XGBoost-based workflow produced the best overall performance, with binary low-versus-high classification and four-class prediction reaching moderate but significant levels of performance.

In addition to predictive performance, the results emphasize the role of reproducibility and interpretability in cognitive HCI. Future work should validate the pipeline with a complete local rerun on the released archive, extend the analysis to participant-independent validation, and explore real-time adaptation in applications such as e-learning, dashboard interaction, and intelligent assistance systems.

## DATA AVAILABILITY

The dataset is publicly available through the COLET release reported by Ktistakis et al. [8].

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## REFERENCES

[1] M. A. Just and P. A. Carpenter, "Eye fixations and cognitive processes," *Cognitive Psychology*, vol. 8, no. 4, pp. 441–480, 1976.

[2] J. H. Goldberg and X. P. Kotval, "Computer interface evaluation using eye movements: Methods and constructs," *International Journal of Industrial Ergonomics*, vol. 24, no. 6, pp. 631–645, 1999.

[3] A. T. Duchowski, "A breadth-first survey of eye-tracking applications," *Behavior Research Methods, Instruments, and Computers*, vol. 34, no. 4, pp. 455–470, 2002.

[4] A. Poole and L. J. Ball, "Eye tracking in human-computer interaction and usability research: Current status and future prospects," in *Encyclopedia of Human-Computer Interaction*, C. Ghaoui, Ed. Hershey, PA: Idea Group Reference, 2005, pp. 211–219.

[5] S. P. Marshall, "The index of cognitive activity: Measuring cognitive workload," in *Proceedings of the 2002 IEEE 7th Conference on Human Factors and Power Plants*. IEEE, 2002, pp. 7–5–7–9.

[6] O. Palinko, A. L. Kun, A. Shyrokov, and P. Heeman, "Estimating cognitive load using remote eye tracking in a driving simulator," in *Proceedings of the 2010 Symposium on Eye-Tracking Research and Applications*. ACM, 2010, pp. 141–144.

[7] W. Chen, T. Sawaragi, and T. Hiraoka, "Comparing eye-tracking metrics of mental workload caused by ndrts in semi-autonomous driving," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 89, pp. 109–128, 2022.

[8] E. Ktistakis *et al.*, "Colet: A dataset for cognitive workload estimation based on eye-tracking," *Computer Methods and Programs in Biomedicine*, vol. 224, p. 106989, 2022.

[9] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Human Mental Workload*, ser. Advances in Psychology, P. A. Hancock and N. Meshkati, Eds. Amsterdam: North-Holland, 1988, vol. 52, pp. 139–183.

[10] E. Haapalainen, S. Kim, J. F. Forlizzi, and A. K. Dey, "Psycho-physiological measures for assessing cognitive load," in *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*. ACM, 2010, pp. 301–310.

[11] R. Bixler and S. D'Mello, "Toward fully automated person-independent detection of mind wandering," in *User Modeling, Adaptation, and Personalization*, ser. Lecture Notes in Computer Science, V. Dimitrova *et al.*, Eds. Cham: Springer, 2014, vol. 8538, pp. 37–48.

[12] C. Ahlstr"om, K. Kircher, and A. Kircher, "A gaze-based driver distraction warning system and its effect on visual behavior," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 2, pp. 965–973, 2013.

[13] G. Marquart, C. Cabrall, and J. de Winter, "Review of eye-related measures of drivers' mental workload," *Procedia Manufacturing*, vol. 3, pp. 2854–2861, 2015.

[14] T. Kosch *et al.*, "A survey on measuring cognitive workload in human-computer interaction," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–39, 2023.

[15] F. Liu *et al.*, "Small-world properties of eye-movement time series assisted in identifying children at high risk for dyslexia," *Biomedical Signal Processing and Control*, vol. 93, p. 106148, 2024.

[16] S. T. Iqbal and B. P. Bailey, "Using eye gaze patterns to identify user tasks," in *The Grace Hopper Celebration of Women in Computing*, 2004, pp. 5–10.

[17] J. Klingner, R. Kumar, and P. Hanrahan, "Measuring the task-evoked pupillary response with a remote eye tracker," in *Proceedings of the 2008 Symposium on Eye-Tracking Research and Applications*. ACM, 2008, pp. 69–72.

[18] C. Belkhiria and V. Peysakhovich, "Eog metrics for cognitive workload detection," *Procedia Computer Science*, vol. 192, pp. 1875–1884, 2021.