Eye-Based Cognitive Overload Prediction in Human-Machine Interaction via Machine Learning

Maria Trigka [1] [a], Elias Dritsas [1] [b], and Phivos Mylonas [1] [c]
[1] *Department of Informatics and Computer Engineering, University of West Attica, Greece*
{mtrigka,idritsas,mylonasf}@uniwa.gr

Keywords: Human-centered computing, Supervised learning, User models, Eye tracking

Abstract: Cognitive overload significantly impacts human performance in complex interaction settings, making its early detection essential for the design of adaptive systems. This paper investigates whether gaze-derived features can reliably predict overload states using supervised machine learning (ML). The analysis is based on an eye-tracking dataset collected during cognitively demanding visual tasks, incorporating fixations, saccades, and pupil diameter measurements. Five classifiers, Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM), XGBoost (XGB), and Multilayer Perceptron (MLP), were evaluated using stratified training and testing splits, alongside 5-fold cross-validation, to identify the presence or absence of cognitive overload. Among them, XGB achieved the highest performance, with an accuracy of 0.902, a precision of 0.958, a recall of 0.821, an F1-score of 0.884, and an area under the ROC curve (AUC) of 0.956. The findings confirm that gaze-derived features alone can reliably distinguish cognitive overload states. The study also highlights trade-offs between model interpretability and predictive performance, with ensemble methods, such as XGB, offering superior results, which support their use in attention-aware systems. Future directions include personalization, temporal modeling, cross-task generalization, and the integration of adaptive feedback mechanisms.

1 INTRODUCTION

The ability to monitor users' cognitive states during task execution is increasingly essential in domains such as human-computer interaction (HCI), education, simulation training, and safety-critical operations. When cognitive demand surpasses an individual's capacity, performance degradation becomes likely, a phenomenon known as cognitive overload. Detecting this overload in real time enables systems to adapt their complexity, pacing, or feedback, thereby reducing user frustration and enhancing overall system usability (Kosch et al., 2023).

Recent advances in eye-tracking technology have made it feasible to non-invasively capture detailed gaze behavior, offering insights into attention, information processing, and cognitive effort. Compared to physiological measures such as electroencephalography (EEG) or functional near-infrared spectroscopy (fNIRS), gaze-based features are easier to integrate into practical environments and impose a minimal burden on users. Research has shown that saccade patterns, fixation durations, pupil dilation, and blink rates are modulated by cognitive load, making them useful input signals for classification models (Abbad-Andaloussi et al., 2022),(Gorin et al., 2024).

ML has become the predominant approach for modeling the relationship between gaze behavior and cognitive states. Classical models such as SVM and LR, as well as more recent deep learning and ensemble methods, have been applied to various cognitive estimation tasks. However, many studies rely on multimodal inputs or domain-specific datasets, which limits their generalizability (Aksu et al., 2024), (Skaramagkas et al., 2023).

Despite the growing body of literature on cognitive state monitoring, a gap remains in evaluating how well ML models can generalize cognitive overload detection using gaze features alone. Existing studies often involve complex sensor setups or focus on specific environments (e.g., virtual reality (VR) or driving), which limits their applicability in more general HCI scenarios (Ghosh et al., 2023).

This study is motivated by the need to support the design of cognitively ergonomic interfaces in profes-

[a] https://orcid.org/0000-0001-7793-0407
[b] https://orcid.org/0000-0001-5647-2929
[c] https://orcid.org/0000-0002-6916-3129

sional human-machine interactions. Predicting cognitive overload from eye-tracking data enables system designers to better align interface complexity with user capabilities, thereby minimizing cognitive strain while preserving interaction fluency. By identifying when users experience mental overload, designers can proactively adjust information flow and visual load, preventing frustration, reducing negative emotional responses, and maintaining effective decision-making. Such predictive insights are critical for ensuring that high-demand operational environments remain user-centered, without hindering human cognition or compromising task performance. A supervised learning framework is adopted to infer whether cognitive overload occurs or not from gaze-derived features, assuming a unified and efficient pipeline. To concretely position this work within the current research landscape and clarify its methodological scope, the key contributions are summarized as follows:

- Investigation of cognitive overload prediction using an existing dataset collected via eye-tracking device during simulated print configuration tasks, featuring gaze-derived metrics such as fixation, saccade, and pupil dynamics.
- A consistent preprocessing pipeline involving signal cleaning, event reconstruction, and feature standardization for downstream model training.
- Statistical and visual analysis of gaze features, confirming the relevance of fixation, pupil, and saccade metrics for distinguishing cognitive states.
- Comparative evaluation of five supervised learning models (LR, NB, SVM, XGB, MLP) using standardized gaze-derived features, stratified validation, and multiple performance metrics including accuracy, precision, recall, F1-score, and AUC.
- Demonstration of the discriminative power of gaze-only features, showing that interpretable models such as XGB can predict overload states with high accuracy and AUC without requiring multimodal inputs.

The remainder of this paper is structured as follows. Section 2 reviews relevant literature on cognitive workload estimation using gaze-based features and ML techniques. Section 3 details the proposed methodology, including dataset overview, preprocessing, feature analysis, model formulation, and evaluation strategy. Section 4 presents and analyzes the experimental results, providing a comparative assessment of model performance across key metrics. Finally, Section 5 summarizes the main findings of this work and outlines directions for future research.

2 Related Works

Recent research in cognitive workload estimation has increasingly focused on gaze-based indicators due to their unobtrusive nature and applicability in real-time systems. Several studies have utilized ML to model the relationship between eye behavior and cognitive demand across various domains, including VR, driving simulation, and tasks that require attention.

A foundational dataset in this area is COLET, which captures gaze behavior under multitasking and time pressure across multiple task conditions (Ktistakis et al., 2022). By training classical classifiers on fixation, saccade, and pupil-related features, the authors reported classification accuracies of nearly 88%, validating gaze signals as effective predictors of cognitive load. To advance generalization in unconstrained settings, the CLERA framework was introduced as a unified deep model for eye-region tracking and load estimation (Ding et al., 2023). It integrates keypoint localization with workload regression in a single trainable architecture, outperforming SVM-based approaches in naturalistic environments. In the context of immersive training, cognitive load was modeled during VR-based disassembly tasks using fixation duration and pupil dilation as inputs to MLP classifiers (Nasri et al., 2024). Results indicated high F1-scores, underscoring the discriminative value of gaze dynamics as task complexity increased.

Multimodal approaches have also been explored. One study combined gaze features with fNIRS signals and driving dynamics within a Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) pipeline, achieving near-perfect classification performance across n-back difficulty levels (Khan et al., 2024). This integration of physiological and behavioral data demonstrated the benefits of signal fusion for robust load inference.

Gaze and pupillary data alone have proven sufficient in low-latency contexts. A CNN-based model was developed to detect stimulus onset using short windows of pupil diameter and gaze vectors across multiple cognitive domains (Dang et al., 2024). Despite domain variation, the models maintained reliable performance, especially for attention-oriented tasks.

Workload prediction in gamified VR environments was also examined through a combination of ocular and biosignals such as heart rate and galvanic skin response (GSR) (Szczepaniak et al., 2024). Using SVM and Random Forest (RF) models, the study reported F1-scores above 0.9, with interpretability analysis highlighting pupil size and blink rate as dominant predictors. Finally, a systematic benchmark evaluated eleven ML algorithms on gaze-derived features extracted under dual-task and time pressure conditions (Skaramagkas et al., 2021). The study demonstrated that lightweight models, such as RF, can match more complex methods in both binary and multiclass cognitive load classification.

A comparative summary of the aforementioned studies is provided in Table 1, which outlines core elements, including domain, modality, feature types, and model classes. As shown in the table, most prior work emphasizes VR or driving contexts, often relying on additional biosignals (e.g., fNIRS, GSR). In contrast, this work focuses on fine-grained gaze-only signals in a visual-cognitive task, leveraging both classical and deep classifiers for robust binary prediction of overload states.

3 Methodology

A structured pipeline was followed to evaluate the predictive capacity of gaze-derived features in detecting cognitive overload. The key blocks of the process are shown in Figure 1 and encompass dataset collection, preprocessing, model training, and evaluation.

3.1 Dataset Overview & Preprocessing

This study utilized a dataset collected from an eye-tracking device during simulated print configuration tasks that involved complex visual interactions. Gaze data were recorded under ecologically valid conditions using the Gazepoint GP3 system (Mannaru et al., 2017; Clemson, 2021), capturing continuous streams of fixations, saccades, pupil diameters, and gaze coordinates. The features of the acquired dataset are summarized in Table 2. Participants underwent individual calibration procedures to ensure the spatial accuracy of gaze mapping. The dataset includes recordings from nine users, supplemented by demographic and interaction-related metadata (e.g., age, experience, task familiarity). It comprises 2,510 overload samples (43.9%) and 3,207 normal samples (56.1%).

The raw data were preprocessed to remove missing or invalid samples using system confidence scores and pupil validity flags. Blink episodes were excluded, fixation events were reconstructed, and saccade magnitudes were derived from gaze displacements. All time indices were aligned to session start to support temporal consistency. This data supported the supervised training and evaluation of cognitive state prediction models based exclusively on visual attention dynamics.

We examined gaze-derived features related to fixation duration, saccadic behavior, pupil size, and gaze distribution to explore behavioral signatures of cognitive overload. These features were selected based on prior literature and observed empirical variability. Our aim was not to reduce dimensionality, but to evaluate the extent to which feature distributions differed across cognitive states in a statistically and behaviorally meaningful way.

Figure 2 shows kernel density plots of the distributions for each gaze-derived feature across normal and overload conditions. Fixation duration and saccade magnitude demonstrate the most distinct separation, with overload samples characterized by longer fixations and reduced saccade amplitudes. Pupil diameter measures are generally elevated under overload, albeit with moderate overlap in distribution. Features such as blink-constricted pupil size, pupil motion magnitudes, and gaze coordinates exhibit less separability but still reflect subtle class-dependent shifts. These trends are consistent with findings linking prolonged fixation, reduced eye movement, and pupil dilation to cognitive load and sustained attention.

To assess the statistical separability of features across cognitive states, we applied the Mann–Whitney U test (Wall Emerson, 2023) to all relevant features in the dataset. This non-parametric test evaluates whether values in the two classes originate from distinct distributions without assuming normality. Features were then grouped based on whether they exhibited statistically significant differences at the $p < 0.01$ level:

- **Significant features** ($p < 0.01$): Most gaze-derived features showed strong evidence of distributional divergence between cognitive states. These include FPOGD (fixation duration), SAC_MAG, LPD, RPD (left/right pupil diameter), LPMM, RPMM, BKDUR, BKPMIN, gaze positions CX, CY. Other significant features are AGE, TOT_EXP, EXP_PLAT, TIME, CNT, FPOGS, FPOGID, BKID, and CS. These results are consistent with the relevant literature, which links these features to visual attention and cognitive load. Their statistical significance supports their inclusion in subsequent interpretation and model development.
- **Non-significant features** ($p \geq 0.01$): A small number of features did not show statistically significant differences. These include i) eye-specific gaze coordinates LPCX, LPCY, RPCX, RPCY, and ii) pupil validity flags LPV, RPV. The lim-

Table 1: Comparative overview of related works on cognitive workload estimation.

Study	**Domain**	**Input Modalities**	**Features Used**	**Model Type**	**Labels**
(Ktistakis et al., 2022)	Visual Search	Eye Tracking	Fixations, Saccades, Pupil, Blinks	RF, SVM, XGB	NASA Task Load Index
(Ding et al., 2023)	Driving (natural)	Eye region video	Keypoints, Pupil, Blinks	Deep multitask CNN	Binary
(Nasri et al., 2024)	VR Training	Eye Tracking	Pupil Dilation, Fixation Duration	MLP, RF	NASA Task Load Index (Binary)
(Khan et al., 2024)	Driving Sim	Eye + fNIRS + Vehicle data	Gaze, HbO2, Vehicle Signals	CNN-LSTM	N-back levels
(Dang et al., 2024)	Multi-domain	Eye Tracking	Pupil, Gaze Vectors	Task-specific CNNs	Stimulus Onset
(Szczepaniak et al., 2024)	VR Game	Eye + GSR + Heart Rate	Saccades, Pupil, Heart Rate, Electrodermal Activity	SVM, RF	Perceived Load
(Skaramagkas et al., 2021)	Visual + Dual Task	Eye Tracking	29 gaze metrics incl. Blink, Fixation	RF, Extra Trees	NASA Task Load Index (3-class)
This work	Visual Task	Eye Tracking only	Fixation, Saccade, Pupil	LR, NB, SVM, XGB, MLP	Cognitive Overload

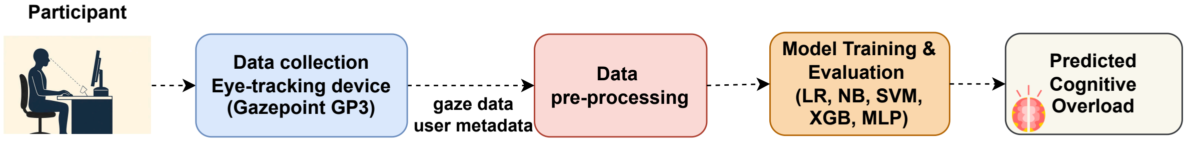

Figure 1: Overview of the experimental pipeline.

ited separability of these features is likely due to their dependence on external factors such as display layout or signal quality, rather than internal cognitive state. While retained for modeling purposes, these features were excluded from interpretative and visual analyses due to their minimal behavioral relevance.

In summary, the analysis confirms that fixation, saccade, and pupil-based features carry meaningful behavioral signals related to cognitive overload. These findings are supported by both statistical evidence and observable patterns in the feature distributions. The dataset involved nine participants with diverse demographic and experiential backgrounds. Table 3 summarizes each user's age, total and platform-specific professional experience, and the proportion of time they were classified as cognitively overloaded. The overload proportion is computed from the binary class labels derived from eye-tracking features using fixation duration and saccade magnitude thresholds.

Figure 3 presents exploratory correlations between overload proportion and selected participant-level variables. A weak negative association is observed between overload and both total and platform-specific experience, suggesting that greater familiarity with the task environment may reduce cognitive strain. In contrast, pupil diameter and fixation duration tend to increase with overload, consistent with established psychophysiological markers of elevated mental effort. Saccade magnitude shows an inverse trend, indicating more localized gaze behavior under higher cognitive load.

3.2 Machine Learning Models

To formulate cognitive overload detection as a binary classification problem, we consider five supervised learning models, each defined by distinct mathematical foundations. Let $\mathbf{x} \in \mathbb{R}^d$ denote a feature vector derived from gaze behavior and participant metadata, and $y \in \{0,1\}$ represent the binary label indicating cognitive state.

LR (Das, 2024) estimates the posterior probability of class membership via the sigmoid function: $P(y=1 \mid \mathbf{x}) = \frac{1}{1+\exp(-\mathbf{w}^\top \mathbf{x}-b)}$ where $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are the model parameters optimized by minimizing

Table 2: Structured summary of extracted features from eye-tracking data.

Feature(s)	Type	Description
Participant Metadata		
UID	Nominal	Participant identifier for grouping or stratified sampling, not predictive.
AGE	Numeric	Age of the participants in years.
TOT_EXP	Numeric	Total professional experience; reflects overall expertise.
EXP_PLAT	Numeric	Experience specific to the simulated platform.
Fixation and Saccade Features		
CNT	Numeric	Frame/sample index; useful for computing fixation order or timing.
TIME	Numeric	Time elapsed since session start; used for temporal analysis.
FPOGID	Nominal	Identifier for each fixation event.
FPOGS	Numeric	Fixation onset time, marking the start of a fixation.
FPOGD	Numeric	Fixation duration (ms); key indicator of cognitive effort.
SAC_MAG	Numeric	Saccade magnitude; amplitude of movement between fixations.
SAC_DIR	Nominal	Saccade direction; used in visual scanning analysis.
Pupil Metrics and Motion		
LPD, RPD	Numeric	Left and right pupil diameters.
LPV, RPV	Nominal	Validity flags for pupil diameter measurements.
LPMM, RPMM	Numeric	Eye motion magnitude; may reflect fatigue or stress.
LPMMV, RPMMV	Numeric	Pupil motion velocity; complementary to LPMM, RPMM.
Blink Features		
BKID	Numeric	Blink ID grouping samples during the same blink.
BKDUR	Numeric	Blink duration (ms).
BKPMIN	Numeric	Minimum pupil diameter recorded during a blink.
Gaze Coordinates and Confidence		
CX, CY	Numeric	Central gaze coordinates on screen.
CS	Numeric	System-provided confidence score for gaze sample validity.
LPCX, LPCY	Numeric	Left eye gaze X/Y screen coordinates.
RPCX, RPCY	Numeric	Right eye gaze X/Y screen coordinates.
BPOGX, BPOGY	Numeric	Raw base point of gaze coordinates; system-derived, not used directly.

Table 3: Participant demographics and overload proportion.

UID	Age	TotExp	PlatExp	Overload
1	45	21.0	15.00	0.44
2	46	20.0	14.00	0.37
3	24	0.0	0.67	0.45
4	32	10.0	0.00	0.39
5	45	22.0	15.00	0.41
6	21	0.0	0.67	0.42
7	27	0.2	1.00	0.46
8	59	30.0	16.00	0.09
9	42	18.0	12.00	0.26

the regularized negative log-likelihood. It offers interpretability and strong baseline performance in standardized feature spaces.

The **SVM** (Pisner and Schnyer, 2020) algorithm constructs a maximum-margin hyperplane in a (possibly nonlinear) transformed feature space. Given a kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$, the soft-margin SVM solves: $\min_{\mathbf{w},b,\xi} \ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i \quad \text{s.t. } y_i(\mathbf{w}^\top \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \ \xi_i \geq 0$ where $C > 0$ is a regularization parameter and $\phi(\cdot)$ denotes the implicit feature mapping.

The Gaussian **NB** (Chen et al., 2020b) classifier assumes conditional independence among features. Each feature x_j is modeled as: $P(x_j \mid y = k) = \mathcal{N}(x_j \mid \mu_{jk}, \sigma^2_{jk})$ The posterior is derived using Bayes' rule: $P(y = k \mid \mathbf{x}) \propto P(y = k) \prod_{j=1}^{d} P(x_j \mid y = k)$ This model is computationally efficient and effective under moderate violations of the independence assumption.

XGB (Chen et al., 2020a) constructs an additive ensemble of regression trees. At iteration t, the model prediction is: $\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(\mathbf{x}_i), \quad f_k \in \mathcal{F}$ where $\mathcal{F}$ denotes the space of CART models. The learning objective is: $\mathcal{L}^{(t)} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k), \quad \Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2$ with ℓ the logistic loss, T the number of leaves, and γ, λ regularization parameters.

An **MLP** (Cinar, 2020) defines a parametric function $f(\mathbf{x};\theta)$ as a composition of linear projections and nonlinear activations. For one hidden layer: $f(\mathbf{x}) = \sigma_2(\mathbf{W}_2 \cdot \sigma_1(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2)$ where σ_1 is typically ReLU and σ_2 is the sigmoid activation for binary classification. Parameters $\theta = \{\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1, \mathbf{b}_2\}$ are optimized by minimizing binary cross-entropy loss via backpropagation and stochastic gradient descent.

Each model offers a unique inductive bias, enabling a comparative evaluation under the same feature representation and data distribution.

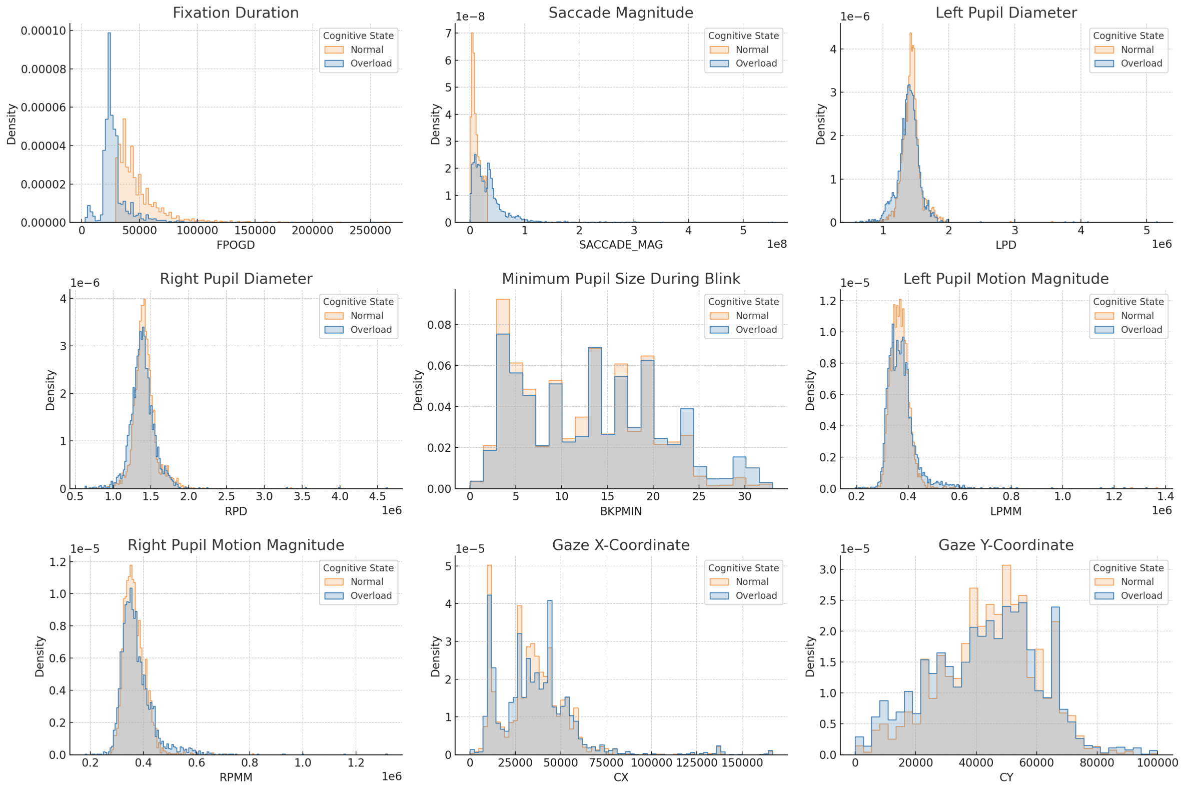

Figure 2: Distributions of 9 gaze-derived features across cognitive states (Normal vs. Overload), capturing fixation, saccade, pupil, and spatial attention dynamics.

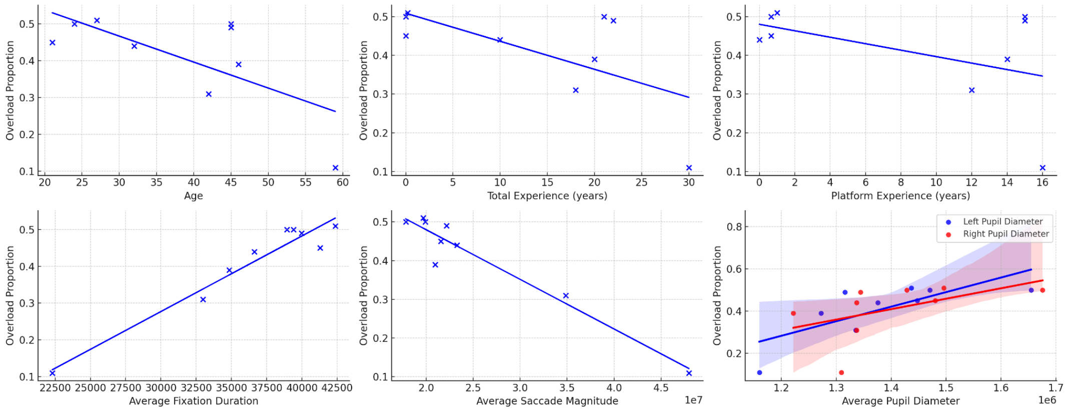

Figure 3: Participant-level correlations between overload proportion and gaze-derived features.

3.3 Model Training and Evaluation

Before model fitting, all numerical feature vectors were standardized using z-score normalization, transforming each feature x_j according to the formula: $x'_j = \frac{x_j - \mu_j}{\sigma_j}$, where μ_j and σ_j denote the empirical mean and standard deviation of the feature computed over the training set (Friedman and Komogortsev, 2019). From an ML perspective, this preprocessing step is essential for ensuring optimization stability in models such as LR, SVM, and MLP, where unscaled features can distort gradient magnitudes or impact kernel evaluations.

To ensure statistical robustness and minimize sampling bias, the dataset was first split into a stratified 80/20 train/test partition, preserving class distribution. Model selection and hyperparameter tuning were performed via stratified 5-fold cross-validation applied exclusively to the 80% training subset. After identifying the best-performing configuration, the selected model was retrained on the full training data and evaluated on the held-out 20% test set. This two-stage protocol ensures that final performance metrics reflect generalization to unseen data.

Each algorithm was trained using the following hyperparameter configurations. LR used the L-BFGS optimizer with ℓ_2-regularization and a maximum of 1000 iterations. The SVM employed an RBF kernel with $C = 1.0$ and $\gamma =$ scale, optimized via sequential minimal optimization. Gaussian NB estimated class-conditional statistics in closed form under a normality assumption. XGB utilized 100 boosted trees with a learning rate of 0.1, tree complexity regularization $\gamma = 0.1$, and L2 penalty $\lambda = 1.0$. Finally, the MLP was trained using the Adam optimizer with ReLU activations, a single hidden layer of 100 units, a learning rate of 0.001, mini-batches of 32, and 500 training epochs.

Model performance was evaluated using standard metrics, which are defined based on the elements of the confusion matrix (Naidu et al., 2023), with TP, TN, FP, and FN denoting true/false positives/negatives:

- Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$
- Precision $= \frac{TP}{TP+FP}$
- Recall $= \frac{TP}{TP+FN}$
- F1-score $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

AUC reflects the probability that a random overload instance is ranked above a normal one, offering threshold-independent discrimination. Models were implemented in Python 3.10 using `scikit-learn 1.3.0` and `XGB 1.7.6`, executed on Ubuntu 22.04 with an Intel i7 CPU and 32GB RAM. GPU acceleration was not required.

4 Results and Discussion

The five ML models were evaluated on a stratified 20% test set using accuracy, precision, recall, F1-score, and AUC. These metrics capture both overall performance and sensitivity to cognitive overload.

Table 4 summarizes the performance of all evaluated models. XGB consistently outperformed the others across all metrics, achieving the highest accuracy (0.902), F1-score (0.884), and AUC (0.956), highlighting its ability to model nonlinear dependencies effectively. MLP also yielded strong results, particularly in recall and F1-score, indicating its capacity to learn complex interactions between gaze-based features. LR and SVM demonstrated comparable but more conservative behavior, with high precision but lower recall, while NB trailed slightly due to its simplifying independence assumptions.

These findings highlight the value of using complementary evaluation metrics beyond accuracy, as F1-score and AUC provide a more nuanced view of model behavior under class imbalance and subtle cognitive effects. They also emphasize the importance of model selection strategies that account for deployment constraints, including interpretability, responsiveness, and tolerance to false negatives. In this context, ensemble-based models like XGB emerge as highly dependable, whereas MLP presents a compelling trade-off between adaptability and complexity.

Beyond aggregate performance, consistency and reliability are essential for cognitive state monitoring in operational settings. While the reported results reflect strong overall performance, future work may benefit from additional validation across multiple data splits to reinforce generalization claims. Furthermore, calibration (i.e., how well predicted probabilities reflect actual outcomes) is critical when system actions depend on confidence thresholds. Incorporating calibration analysis can further improve deployment readiness in practical settings.

Table 4: Experimental results on the test set.

Model	Accuracy	Precision	Recall	F1-score	AUC
XGB	**0.902**	**0.958**	**0.821**	**0.884**	**0.956**
MLP	0.870	0.925	0.765	0.837	0.918
LR	0.851	0.910	0.723	0.805	0.894
SVM	0.846	0.902	0.714	0.797	0.889
NB	0.828	0.872	0.690	0.770	0.871

5 Conclusions

This study examined cognitive overload detection using only gaze-derived features, applying five supervised ML models to data from visually demanding tasks. Among them, XGB delivered the best performance, achieving an accuracy of 0.902, precision of 0.958, recall of 0.821, F1-score of 0.884, and AUC of 0.956. These results demonstrate that eye-based metrics, including fixations, saccades, and pupil diameter, are sufficient for reliable binary classification, thereby eliminating the need for multimodal input.

Beyond predictive performance, the findings highlight the feasibility of deploying lightweight, gaze-based models in real-time HCI systems. Unlike multimodal approaches, this method offers a focused and interpretable solution based solely on ocular behavior.

Models were trained on pooled gaze data from multiple users, allowing intra-sample generalization but not evaluating performance on unseen individuals. Future work should investigate subject-independent validation and extend the framework to support personalized modeling, multi-class classification, temporal gaze dynamics, and cross-task generalization, thereby enhancing adaptive, user-aware cognitive monitoring.

REFERENCES

Abbad-Andaloussi, A., Sorg, T., and Weber, B. (2022). Estimating developers' cognitive load at a fine-grained level using eye-tracking measures. In *Proceedings of the 30th IEEE/ACM international conference on program comprehension*, pages 111–121.

Aksu, Ş. H., Çakıt, E., and Dağdeviren, M. (2024). Mental workload assessment using machine learning techniques based on eeg and eye tracking data. *Applied Sciences*, 14(6):2282.

Chen, J., Zhao, F., Sun, Y., and Yin, Y. (2020a). Improved xgboost model based on genetic algorithm. *International Journal of Computer Applications in Technology*, 62(3):240–245.

Chen, S., Webb, G. I., Liu, L., and Ma, X. (2020b). A novel selective naïve bayes algorithm. *Knowledge-Based Systems*, 192:105361.

Cinar, A. C. (2020). Training feed-forward multi-layer perceptron artificial neural networks with a tree-seed algorithm. *Arabian Journal for Science and Engineering*, 45(12):10915–10938.

Clemson (2021). *Gazepoint GP3 Eye Tracker User Manual*.

Dang, Q., Kucukosmanoglu, M., Anoruo, M., Kargosha, G., Conklin, S., and Brooks, J. (2024). Auto detecting cognitive events using machine learning on pupillary data. *arXiv preprint arXiv:2410.14174*.

Das, A. (2024). Logistic regression. In *Encyclopedia of quality of life and well-being research*, pages 3985–3986. Springer.

Ding, L., Terwilliger, J., Parab, A., Wang, M., Fridman, L., Mehler, B., and Reimer, B. (2023). Clera: a unified model for joint cognitive load and eye region analysis in the wild. *ACM Transactions on Computer-Human Interaction*, 30(6):1–23.

Friedman, L. and Komogortsev, O. V. (2019). Assessment of the effectiveness of seven biometric feature normalization techniques. *IEEE Transactions on Information Forensics and Security*, 14(10):2528–2536.

Ghosh, S., Dhall, A., Hayat, M., Knibbe, J., and Ji, Q. (2023). Automatic gaze analysis: A survey of deep learning based approaches. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):61–84.

Gorin, H., Patel, J., Qiu, Q., Merians, A., Adamovich, S., and Fluet, G. (2024). A review of the use of gaze and pupil metrics to assess mental workload in gamified and simulated sensorimotor tasks. *Sensors*, 24(6):1759.

Khan, M. A., Asadi, H., Qazani, M. R. C., Lim, C. P., and Nahavandi, S. (2024). Functional near-infrared spectroscopy (fnirs) and eye tracking for cognitive load classification in a driving simulator using deep learning. *arXiv preprint arXiv:2408.06349*.

Kosch, T., Karolus, J., Zagermann, J., Reiterer, H., Schmidt, A., and Woźniak, P. W. (2023). A survey on measuring cognitive workload in human-computer interaction. *ACM Computing Surveys*, 55(13s):1–39.

Ktistakis, E., Skaramagkas, V., Manousos, D., Tachos, N. S., Tripoliti, E., Fotiadis, D. I., and Tsiknakis, M. (2022). Colet: A dataset for cognitive workload estimation based on eye-tracking. *Computer Methods and Programs in Biomedicine*, 224:106989.

Mannaru, P., Balasingam, B., Pattipati, K., Sibley, C., and Coyne, J. T. (2017). Performance evaluation of the gazepoint gp3 eye tracking device based on pupil dilation. In *Augmented Cognition. Neurocognition and Machine Learning: 11th International Conference, AC 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9-14, 2017, Proceedings, Part I 11*, pages 166–175. Springer.

Naidu, G., Zuva, T., and Sibanda, E. M. (2023). A review of evaluation metrics in machine learning algorithms. In *Computer science on-line conference*, pages 15–25. Springer.

Nasri, M., Kosa, M., Chukoskie, L., Moghaddam, M., and Harteveld, C. (2024). Exploring eye tracking to detect cognitive load in complex virtual reality training. In *2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 51–54. IEEE.

Pisner, D. A. and Schnyer, D. M. (2020). Support vector machine. In *Machine learning*, pages 101–121. Elsevier.

Skaramagkas, V., Ktistakis, E., Manousos, D., Kazantzaki, E., Tachos, N. S., Tripoliti, E., Fotiadis, D. I., and Tsiknakis, M. (2023). esee-d: Emotional state estimation based on eye-tracking dataset. *Brain Sciences*, 13(4):589.

Skaramagkas, V., Ktistakis, E., Manousos, D., Tachos, N. S., Kazantzaki, E., Tripoliti, E. E., Fotiadis, D. I., and Tsiknakis, M. (2021). Cognitive workload level estimation based on eye tracking: A machine learning approach. In *2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 1–5. IEEE.

Szczepaniak, D., Harvey, M., and Deligianni, F. (2024). Predictive modelling of cognitive workload in vr: An eye-tracking approach. In *Proceedings of the 2024 Symposium on Eye Tracking Research and Applications*, pages 1–3.

Wall Emerson, R. (2023). Mann-whitney u test and t-test.