AI

MDPI

Article

Evaluating Feature-Based Machine-Learning Models with Post Hoc Explainability for Eye-Tracking-Based Task Type and Workload Inference

Tomi Božak [1,2,*,†], Shivalika Goyal [3,4,5,*,†], Marc Langheinrich [3], Martin Gjoreski [3] and Gašper Slapničar [1,6]

[1] Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[2] Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, 1000 Ljubljana, Slovenia
[3] Faculty of Informatics, Università della Svizzera italiana (USI), Via Buffi 13, 6900 Lugano, Switzerland
[4] Academy of Scientific and Innovative Research (AcSIR), Ghaziabad 201002, Uttar Pradesh, India
[5] School of Engineering, RMIT University, Melbourne, VIC 3000, Australia
[6] Jožef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia
* Correspondence: tb85088@student.uni-lj.si (T.B.); shivalika.goyal@usi.ch or email@shivalikagoyal.com (S.G.)
† These authors contributed equally to this work.

Abstract

Eye tracking is a valuable behavioral signal for human-centered AI, yet the reliability of feature-based machine-learning models for inferring task type and workload across users and tasks remains uncertain, because experimentally defined workload labels may reflect task type and visual structure as much as cognitive demand. The practical problem is that designers of gaze-adaptive systems need to know which inferences are dependable enough to act on, and reported accuracies alone do not answer this, because the choice of prediction target and validation split can determine the result. This study systematically evaluates feature-based machine-learning models with post hoc explainability across three prediction targets: task type, binary load-versus-rest, and three-level workload. Eye-movement features derived from fixations, saccades, pupils, and blinks were extracted from short temporal windows collected from 54 participants performing attention, visual-spatial, and memory tasks under rest, easy, and difficult conditions, and evaluated using leave-one-subject-out (LOSO) and leave-one-group-out (LOGO) validation. Task type was classified most reliably (85.9% LOSO, 83.4% LOGO), binary load-versus-rest showed moderate, validation-sensitive robustness (81.4% LOSO, 63.9% LOGO), and three-level workload classification was substantially more challenging (56.4% LOSO, 44.3% LOGO). SHAP and statistical analyses consistently identified fixation dispersion, pupil-related measures, and subject-normalized features as the strongest contributors across all three targets. These findings show that prediction target definition, validation strategy, and post hoc explainability jointly determine what can be reliably inferred from gaze-based machine-learning models. Eye tracking alone therefore appears promising for task-type recognition and may support coarse engagement-related inference when the deployment task family is represented during model development, whereas task-independent fine-grained workload estimation remains unsupported by the present evidence.

Keywords: eye tracking; task-type inference; workload inference; feature-based machine learning; post hoc explainability; human-centered AI; cross-task generalization

Academic Editor: Tihomir Orehovački

Received: 19 July 2026
Revised: 16 August 2026
Accepted: 18 August 2026
Published: 21 August 2026

1. Introduction

Machine-learning systems that infer a person's internal state from behavioral signals like gaze, posture, and physiological response are only as trustworthy as the label they are trained to predict. When that label is a proxy for the construct of interest rather than the construct itself, a model can achieve high accuracy while answering a different question than the one it claims to answer. This is not specific to any one modality; it can arise wherever an experimentally convenient label stands in for an internally experienced state. Eye tracking is a particularly instructive case, both because it is now a widely deployed behavioral signal for human-centered AI systems that adapt to a user's task type and interaction state [1–3], and because this literature offers a clear, well-documented example of the pattern.

Oculometric measures like pupil diameter, fixation dynamics, saccadic behavior, and blinks are attractive for state inference because they are captured non-intrusively through a modality already embedded in consumer devices, head-mounted displays, and automotive systems [4], and because, unlike many physiological signals, they are open to direct physiological interpretation. This matters because high-capacity models can achieve strong predictive accuracy while offering little insight into which signals drive a decision [5,6]. Feature-based models built on statistically characterized oculometric measures avoid this trade-off: combined with post hoc explainability methods such as SHapley Additive exPlanations (SHAP) [7], they trace a prediction back to specific, physiologically grounded behaviors like fixation dispersion, pupil response, saccadic amplitude, rather than an opaque internal representation. Prior work confirms this is achievable without a substantial accuracy cost [8].

Yet a substantial share of this literature describes its models as "cognitive workload" or "cognitive load" inference while training them on experimentally defined task-design labels—an "easy" versus "difficult" condition set by the study protocol—and using subjective instruments such as NASA-TLX [9] only as a post hoc, secondary validation measure, never as the prediction target itself. A model trained this way learns to recognize task type and experimentally defined condition structure, not internally experienced workload; the two are correlated, but not interchangeable, and retrospective instruments like NASA-TLX cannot resolve moment-to-moment state and carry their own task-structure and self-report confounds [10]. The consequence is concrete: models trained on a single task generalize poorly to new task layouts, because oculomotor demand is shaped by visual structure as much as by cognitive effort [11,12]. Multimodal approaches that add electroencephalography, electrodermal activity, or electrocardiography [13–15] improve resolution but do not resolve this underlying conflation; they add sensing modalities without clarifying what the target variable represents.

Three specific gaps follow, and Table 1 makes them concrete. First, task type, a coarse experimentally defined proxy for workload such as binary load-versus-rest, and a finer-grained workload-related target have not been treated together as a single explicit hierarchy of increasing inferential difficulty evaluated within one common analytical framework. Second, validation practice compounds the problem: models are typically validated only across held-out participants, not across held-out task groups, so it remains unknown how much reported accuracy reflects overfitting to task-specific visual layout rather than transferable signal. A recent systematic review of machine-learning workload classification in the adjacent EEG literature found that only about a third of studies used a cross-subject-robust evaluation, while roughly one in eight attempted cross-task evaluation at all [16]. Third, explainability analyses are rarely checked for consistency across targets, so it is not established whether the gaze features that explainability methods highlight are candidate robust markers or artifacts of a single task setup.

https://doi.org/10.3390/ai7080325

This work closes that gap directly. Using the Cognitive Load Understanding through Experimental Sensing (CLUES) dataset (54 analyzed participants performing attention, visual-spatial, and memory tasks under rest, easy, and difficult conditions) [17], we extracted curated oculometric feature sets from fixations, saccades, pupil dynamics, and blinks within short temporal windows, and evaluated feature-based classifiers with post hoc explainability under both Leave-One-Subject-Out (LOSO) and Leave-One-Group-Out (LOGO) validation.

Research Question: How do prediction targets and validation strategies affect feature-based eye-tracking models for task-type and workload inference?

We answer this along three lines: how predictive performance differs across a hierarchy of prediction targets of increasing inferential difficulty, including task type, binary load-versus-rest, and three-level workload; how that performance changes under subject-held-out versus task-group-held-out validation; and which gaze features remain consistently important across this hierarchy according to SHAP and statistical analysis, as opposed to features specific to a single target or validation setting. As shown in Section 4, reliability degrades systematically across this hierarchy rather than uniformly, a pattern with direct implications for what gaze-based AI systems can and cannot be trusted to infer.

The remainder of this paper is organized as follows: Section 2 reviews related work on eye tracking for human-centered AI, task-type and workload-related inference, feature-based versus deep learning approaches, explainable AI in eye tracking, and validation protocols. Section 3 describes the dataset, prediction targets, feature extraction, and modeling pipeline. Section 4 presents results organized by prediction target. Section 5 discusses what these results reveal about target definition, validation strategy, and explainability jointly, along with limitations. Section 6 discusses limitations and future work, and Section 7 concludes.

Contributions

The novelty claimed here is not eye-tracking-based workload classification, feature-based modeling, or the use of SHAP individually; each is established, as Section 2 documents. It is instead the controlled comparative design that places these elements in one framework, and the diagnostic finding that follows from it. This work makes three contributions. First, it introduces an explicit hierarchy of prediction targets including task type, binary load-versus-rest, and three-level workload. These are evaluated within a common feature-based modeling and explainability framework, isolating a conflation between task-design labels and workload state that prior work has largely left implicit. Second, it evaluates this hierarchy under both subject-held-out and task-group-held-out validation, examining the extent to which predictive performance remains stable under participant-held-out and task-group-held-out evaluation. Third, it applies SHAP-based and statistical explainability consistently across the full hierarchy, rather than to a single target in isolation, identifying which gaze features recur across targets and validation settings versus those that are more target- or context-specific. Together, these contributions show that prediction target definition, validation strategy, and post hoc explainability are not independent design choices but jointly determine what a gaze-based machine-learning model can be trusted to infer. It is a finding with implications for how transparent, robust human-centered AI systems built on behavioral signals should be defined and evaluated, in eye tracking and beyond.

2. Related Work

This section reviews prior work relevant to the three-target hierarchy and validation design used in this study. We first situate eye tracking within human-centered AI

(Section 2.1), then separate two literatures that are frequently, and, we argue, imprecisely, treated as interchangeable: inference of task type (Section 2.2) and inference of workload-related state (Section 2.3). We then review feature-based versus deep-learning approaches to modeling ocular data (Section 2.4), explainable AI in eye tracking (Section 2.5), and validation protocols for assessing generalization (Section 2.6), closing with the specific gap this work targets.

2.1. Eye Tracking for Human-Centered AI

Eye tracking has become a practical sensing channel for human-centered AI systems that adapt to a user's attentional and interaction state, because gaze behavior is captured non-intrusively and is increasingly available through hardware already embedded in consumer devices, head-mounted displays, and vehicles [1,4]. Reviews of oculometric measurement [18,19] consistently identify pupil diameter, fixation dynamics, saccadic behavior, and blink patterns as the core signal families for attentional allocation [19,20]. These reviews also show that the strength of any single ocular indicator depends on what is being inferred, task identity, coarse engagement, or fine-grained workload, motivating the two-part split below rather than one undifferentiated problem.

2.2. Task-Type Inference

Ocular behavior is strongly shaped by the structure of the task itself, independent of any variation in internal effort: internally focused tasks such as memory recall reduce exploratory scanning, whereas externally demanding visual-search tasks promote broader fixation dispersion and larger saccades [21]. Task-type classification from gaze has an established methodological lineage independent of workload research: eye movement analysis was first demonstrated as a standalone sensing modality for activity and task recognition using electrooculography [22], later extended to high-level contextual-cue recognition [23], and more recently approached with convolutional networks classifying task directly from raw gaze and pupil data [24]. Within the narrower workload literature, multivariate machine-learning approaches applied to oculometric features distinguish between task types with substantially higher separability than they achieve for internal-state classification [25]. This asymmetry motivates treating task type as a distinct, comparatively easier prediction target, rather than folding it into "workload inference" as a secondary signal, as much prior work implicitly does.

2.3. Workload-Related Inference

Cognitive Load Theory frames mental effort as the allocation of a limited pool of cognitive resources across intrinsic, extraneous, and germane demands [26]. Ocular indicators are attractive correlates of this allocation because several are mechanistically linked to attentional and executive control: the task-evoked pupillary response is one of the more robust markers of cognitive effort, modulated by the locus coeruleus-norepinephrine system [27], with increases observed in memory paradigms [28] and operational settings such as air-traffic control [29]. Fixation duration, count, and spatial dispersion are similarly sensitive to moment-to-moment attentional demand [30], and cognitive load unfolds dynamically, with pupil dilation rising with demand and stabilizing as strategies adapt [31].

This body of work links ocular metrics to workload-related constructs, but most operationalize "workload" through the same experimentally defined condition structure used to design the task, with instruments such as NASA-TLX [9] used only as convergent, post hoc validation rather than as the training target [10]. Reviews of pupillometric and blink-based indicators in VR learning [32] and multichannel activity in university students [33] both illustrate this: eye-tracking reliably differentiates experimentally manipulated condition levels, a narrower claim than "measuring cognitive workload" as such. Recent work apply-

ing interpretable, feature-based classifiers directly to a three-level workload classification problem [8] reinforces both the feasibility and the comparative difficulty of this specific target relative to coarser classification problems.

2.4. Feature-Based Versus Deep Learning Approaches

Machine-learning approaches to ocular data broadly split into high-capacity models (deep neural networks, gradient boosting ensembles) and feature-based models built on statistically characterized, hand-engineered oculometric measures. Multivariate feature-based models have repeatedly outperformed univariate statistical analysis for workload-related classification, while offering an interpretability advantage that high-capacity models generally lack: deep models can achieve strong predictive accuracy but provide limited transparency about which ocular features drive a given prediction [5,6], a limitation demonstrated concretely by convolutional-network approaches that classify task directly from raw gaze coordinates and pupil size, achieving competitive accuracy while flagging that opacity impedes generalization to applied contexts [24]. Three further considerations favor the feature-based design here. First, the modeling unit is a short window already reduced to aggregate event statistics rather than a raw gaze sample stream, so sequence models would change the object of study from interpretable oculometric features to learned representations, which is precisely the contrast this work examines. Second, at the data scale used here, on the order of tens of thousands of windows from tens of participants, benchmark evidence indicates that gradient-boosted tree ensembles remain competitive with or superior to deep architectures on tabular inputs, while requiring substantially less tuning [34–36]; gradient boosting has also performed strongly in recent eye-tracking and physiological workload studies, giving the best reported result on a multimodal driving dataset under standard cross-validation [37] and exceeding 75% cross-participant accuracy from eye-tracking features alone, although a deep model reached a higher figure in that second case [38]. Third, exhaustive leave-one-out evaluation over three targets, two protocols, and a feature-family ablation grid requires many hundreds of model fits, which a tree ensemble makes tractable. We do not claim that gradient boosting would outperform sequence models on these data; that comparison is outside the present scope and is noted in Section 6. This trade-off is the primary methodological justification for the feature-based design adopted in this work, consistent with evidence that feature-based, interpretable classifiers can approach the performance of less transparent alternatives on closely related workload-classification problems [8].

2.5. Explainable AI in Eye Tracking

Explainability is increasingly treated as a requirement for responsible state inference in human-centered AI, allowing predictions to be checked against a stated rationale rather than accepted on trust [39,40]. Post hoc attribution methods such as SHAP [7] provide a principled way to connect a model's decision back to specific oculometric features. Explainable modeling remains comparatively underused in eye-tracking-only workload research, and fewer studies validate their explainability outputs against independent statistical or behavioral evidence, a gap that leaves open whether a model's highlighted features reflect genuine psychophysiological mechanisms or dataset-specific artifacts [41–43]. Checking whether this convergence holds as the prediction target itself changes is treated in this work as a methodological requirement rather than a supplementary analysis.

2.6. Validation Protocols and Generalization

Even where feature-based, explainable models perform well, most evaluations rely on cross-participant (subject-held-out) validation alone, leaving unaddressed how much reported accuracy reflects gaze behavior that generalizes across task structures, as opposed

to behavior tied to a specific visual layout. This is a known limitation: models trained on a single task frequently fail to generalize to new task structures, because oculomotor patterns shift with layout and cognitive demand [12], yet it is rarely tested directly with a cross-task-group protocol: Skaramagkas et al. [25] and Pillai et al. [12] evaluate workload-related prediction with cross-participant validation only, while Martinez-Cedillo et al. [11] examine task-related gaze dynamics without assessing predictive cross-task transfer. Multimodal and domain-adaptation approaches combining eye tracking with electroencephalography or other physiological signals explore generalization from a different angle [14,15] but depend on intrusive sensors or participant-specific calibration, and do not resolve whether unimodal gaze features alone can generalize across users and task groups.

To situate the present study against this literature, Table 1 compares representative eye-tracking-based studies of task and workload-related inference along the dimensions most relevant to generalization and explainability. The table also records the prediction target of each study, the dimension on which the present contribution turns. Kosch et al. [2] is discussed here as a survey-level reference rather than a directly comparable empirical study, and is listed in the extended comparison in Supplementary Table S8, because it provides the field's most comprehensive synthesis of measurement and evaluation practice and serves as the reference point against which the empirical studies below, and the present work, can be judged for methodological completeness.

As Table 1 shows, the existing literature on eye-tracking-based task and workload-related inference is concentrated on single-task settings with cross-participant evaluation only; cross-task-group generalization and explainability validated against independent statistical evidence are rarely reported together. Cross-task transfer has been demonstrated for eye-tracking-based load classification, notably by training on a working-memory task and testing on an emergency simulation game, where easy-versus-difficult accuracy reached 63.78–67.25% [44]; recent deep-learning work on eye-movement indices has likewise addressed workload and task identity as separate targets, though under random rather than participant-held-out splits [45]. What has not, to our knowledge, been reported is the combination examined here: an explicit hierarchy of prediction targets ranging from task type to fine-grained workload state, evaluated within one common analytical framework using both participant-held-out and task-group-held-out validation, and with explainability checked for convergence with independent statistical evidence across every target in that hierarchy. This is precisely the gap the present study addresses: Section 3 describes a dataset and modeling pipeline designed to evaluate task type, binary load-versus-rest, and three-level workload as separate, explicitly compared targets, under both LOSO and LOGO validation, with SHAP-based and statistical explainability applied consistently across all three.

Table 1. Representative comparison of prior eye-tracking-based task and workload-related inference studies, contrasting sensing modality, task diversity, evaluation protocol, and explainability.

Study (Year)	Task Domain (s)	Prediction Target (s)	Sensing Modality	Cross-User Eval.	Cross-Task Eval.	Explainability
[46]	Single lab task	Task difficulty (single task)	Eye tracking	—	—	—
[14]	Driving (real-world)	Cognitive load (binary)	Vision/video	✓	—	—
[8]	Cognitive workload (3-level)	Workload (3-level)	Eye tracking	✓	—	-

Table 1. *Cont.*

Study (Year)	Task Domain (s)	Prediction Target (s)	Sensing Modality	Cross-User Eval.	Cross-Task Eval.	Explainability
[12]	Driving simulator	Cognitive load (binary)	Eye tracking	✓	—	—
[32]	VR learning (visual fatigue & cognitive load)	Cognitive load/visual fatigue (review)	Eye tracking	—	—	—
[20]	Visual search/puzzles	Workload (multi-level)	Eye tracking	✓	—	—
[47]	Semi-autonomous driving	Mental workload (task conditions)	Eye tracking	—	—	—
[48]	Resource-management game	Cognitive load (theory-derived levels)	Eye tracking	✓	—	✓ Theory-driven
[33]	Multichannel learning (university students)	Cognitive load (task conditions)	Eye tracking	—	—	—
[11]	Real-world tasks (gaze dynamics)	Gaze dynamics under load (no prediction)	Eye tracking	✓	—	—
This work	Attention, visual-spatial, memory	Task type + binary load + 3-level workload	Eye tracking	✓ (LOSO)	✓ (LOGO)	✓ (SHAP + stats)

✓ = reported in the study; — = not reported. LOSO = leave-one-subject-out; LOGO = leave-one-group-out; SHAP = SHapley Additive exPlanations.

3. Methodology

3.1. Dataset and Task Design

We used the CLUES dataset [17], collected under controlled laboratory conditions at two sites. CLUES contains multimodal recordings from tasks performed at systematically varied difficulty levels. In this study, we use only the eye-tracking modality to evaluate three related prediction targets: task type, binary load-versus-rest, and three-level experimentally defined workload conditions. Recordings were collected from 55 participants, of whom 54 entered the analyses reported here; one participant was excluded during data curation owing to insufficient valid eye-tracking data. The cohort averaged 27.7 years of age (SD 6.6, range 19–47), was 63.6% male, and comprised students and employed adults [17]; further participant characteristics, and an analysis relating them to per-participant model performance, are reported in Supplementary Section S1.1. Eye-tracking data were recorded using a Tobii Pro Spark eye tracker at 60 Hz with a standard 9-point calibration on a single monitor.

Participants completed a series of tasks spanning three cognitive domains (referred to as task types): attention, visual-spatial, and memory (AVM). Each domain included experimentally defined easy and difficult variants, such as 2-back versus 3-back and easy versus difficult visual-difference detection. Rest blocks, consisting of passive viewing or simple fixation, were interleaved for baseline comparison. The task set provided controlled variation in cognitive demand, visual structure, and gaze behavior, allowing us to examine whether the models primarily distinguish task type, coarse task engagement, or finer easy-difficult condition differences. Tasks were presented in a fixed order for all participants,

with some paradigms of fixed duration and others terminating on the participant's response, and participants wore additional sensing equipment, including smart glasses, concurrently with the eye tracker; the dataset publication reports that this equipment occasionally occluded the line between the eye tracker and the pupil [17], a measurement constraint we return to in Section 6. Both recording sites used a fixed indoor lighting setup with natural room light and no direct sunlight exposure, kept consistent across all sessions; gaze position is the binocular average of the left- and right-eye coordinates, following the source dataset's own acquisition convention. Pupil diameter was recorded and analyzed separately for each eye; because the two eyes' pupil measurements were highly correlated, most right-eye pupil features were removed as redundant by the correlation-based pruning described in Section 3.2, leaving a mix of left- and right-eye pupil features in the final retained set (Supplementary Table S1b). Which eye's feature survived within a correlated pair was incidental to the pruning procedure rather than a deliberate choice, since the two eyes gave equivalent measurements.

The prediction targets used here are experimentally defined task conditions, not measured workload states. For each trial, participants provided subjective workload ratings using the Instantaneous Self-Assessment (ISA), normalized within participants [49], and the post-task NASA Task Load Index (NASA-TLX) [9]. These ratings were used to verify that the experimental difficulty manipulation produced the intended differences in perceived workload, and the dataset publication reports that NASA-TLX scores differed significantly between the easy and difficult versions of every task, while task performance differed for all paradigms except the Stroop and visual-memory tasks [17]; they were not used as supervised learning targets, since the analysis focused on what feature-based models learn from the experimentally defined task conditions themselves. The machine-learning analyses used three design-based targets: task type (Attention, Visual-Spatial, and Memory), binary Load versus Rest, and three-level EDR labels (Easy, Difficult, and Rest). These workload-related labels therefore represent experimental conditions rather than participant-specific measurements of experienced workload. Figure 1 summarizes the experimental design and label mappings, while Figure 2 shows the resulting class distributions with color-coded task and workload categories.

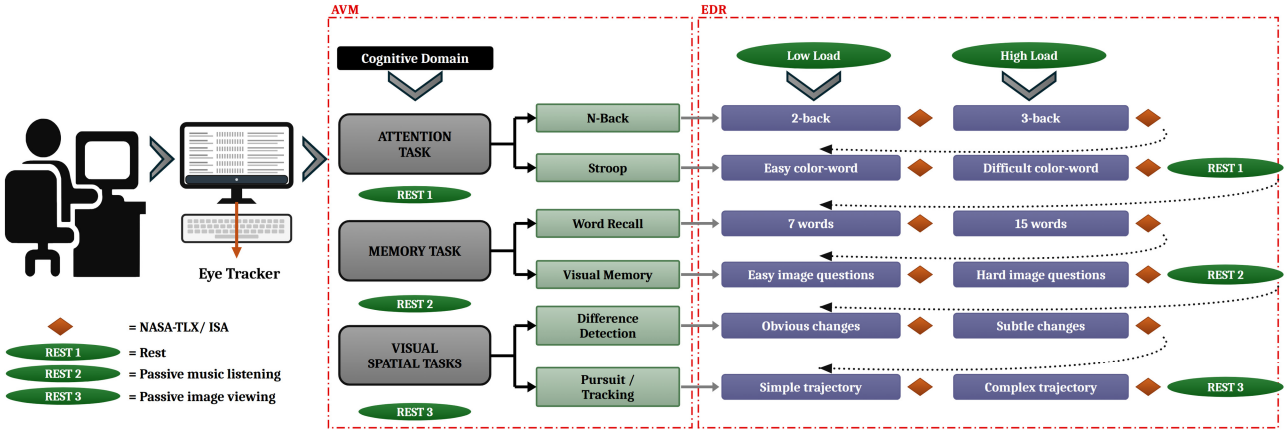

Figure 1. Overview of the experimental design. Participants completed attention, visual-spatial, and memory tasks with systematically varied difficulty (easy/difficult), interleaved with rest periods. Arrows denote structural and temporal relations rather than data flow: solid black and grey arrows indicate the decomposition of each cognitive domain into its constituent tasks and difficulty levels, whereas dotted curved arrows indicate the order in which task blocks were presented (from easy to difficult within each task, then on to the next task). Eye-tracking data were segmented into 3-s windows with 25% overlap. Labels used in analyses include AVM (task type) and EDR (Rest-Easy-Difficult) for workload-related conditions.

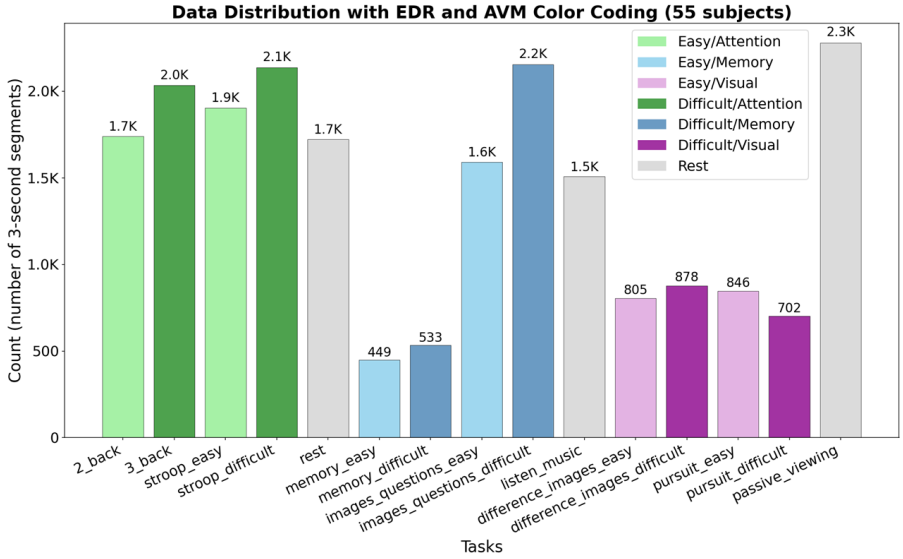

Figure 2. Distribution of all 3-s segments across individual task variants and rest conditions (pooled across all participants), color-coded by AVM task type and EDR workload level. This visualization illustrates class imbalance and the distribution of samples across both task types and workload conditions.

3.2. Preprocessing and Feature Extraction

Raw eye-tracking data consisted of timestamped horizontal and vertical gaze coordinates and pupil-diameter measurements. We removed invalid samples in cases where gaze from both eyes was invalid for longer than 400 ms. Gaze outliers were identified from the distribution of inter-sample movement amplitudes and replaced by the mean of their immediate neighbors, with large saccades and dynamic overshoots explicitly preserved rather than treated as artifacts; gaze and pupil signals were then smoothed with a Savitzky-Golay filter. Fixations were detected with the dispersion-threshold (I-DT) algorithm and blinks with a duration gate. The complete parameter set for every step is listed in Supplementary Section S1.2. To prevent data leakage, layout-dependent and subject-specific absolute gaze features (such as the values of the coordinates of fixations) were excluded.

Event-based fixation, saccade, and blink signals were derived using EyeTrackLib [50], an open-source Python 3.12.8 library developed by the authors, which implements standard dispersion-threshold identification algorithms commonly used in eye-movement parsing [51]. Because eye-movement events unfold on the order of tens to hundreds of milliseconds, we segmented all signals into 3-s windows with 25% overlap, which provides sufficiently fine temporal granularity to capture within-trial dynamics while still aggregating enough events per segment for stable feature estimates. This approach produced 21,288 windows for EDR/Binary modeling (rest: 5512; low-load: 7335; high-load: 8441) and 14,328 windows for AVM modeling (attention: 6803; memory: 4456; visual: 3069), all from 54 participants. This 3-s, 25%-overlap configuration was inherited unchanged from the CLUES acquisition protocol, which selected a short window for eye movement because ocular events unfold rapidly and because individual tasks lasted under two minutes [17]; it was not optimized on the present data, since tuning the window against the same data used for evaluation would risk an optimistic bias. Comparable work has used similar window lengths [46]. The overlap increases the number of available windows for training, while train-test separation is enforced at the participant or task-group level, so no window is shared across partitions. The class distribution across AVM and EDR labels is shown in Figure 2. To isolate within-participant variation, subject-relative features were constructed by centering window-level measures on each participant's own recording-level summary. This participant-referenced transformation does not pool information across participants

and is analytically distinct from fold-fitted scaling. All features were subsequently min-max scaled, with the scaler fitted exclusively on the training partition of each validation fold and applied unchanged to the corresponding held-out partition. In total, 488 raw oculometric features were initially extracted across fixation, saccadic, pupil, and blink categories, guided by prior literature [20,25]. The relevance of these feature families is discussed in Section 2.

To reduce redundancy and mitigate overfitting, we applied a multi-stage pruning procedure: (a) correlation-based redundancy reduction, where for each pair of highly correlated features ($|r| > 0.80$) we retained the feature judged to be more interpretable and theoretically meaningful; (b) removal of features unsuitable for modeling, comprising absolute-coordinate, identifier-like and subject-identifying columns, together with any feature retaining fewer than 60% valid values; and (c) a second, target-specific correlation pass at the same threshold. This reduced the initial 488 candidate features to a final curated set of 58 oculometric features for the binary and three-level workload targets, and 59 for the task-type target, where the target-specific pass retained additional saccade features. No dimensionality-reduction transform such as principal component analysis and no wrapper-based feature selection were applied, so every modeled feature remains directly interpretable. The leave-one-group-out experiments are defined on the pooled task set before this target-specific pass, and therefore operate on the shared 61-feature intermediate set for all three targets. Residual redundancy among the retained features, and the exact composition of each feature set, are reported in Supplementary Section S1.4. Ocular features refer to all extracted quantitative oculometric variables derived from eye-tracking data, whereas ocular markers denote the subset of these features that demonstrate statistically reliable and physiologically interpretable associations with workload-related state or task type. The complete feature taxonomy, including feature names, oculometric domain, statistical type, and brief descriptions, is provided in Supplementary Section S1.

3.3. Statistical Analysis of Ocular Features

To examine systematic differences in ocular features across experimentally defined workload conditions (EDR) and task type (AVM), we performed feature-wise one-way ANOVAs, or non-parametric permutation ANOVAs when assumptions of normality or homoscedasticity were violated. Significant omnibus effects ($\alpha = 0.05$) were followed by Bonferroni-corrected post hoc tests to identify pairwise differences (e.g., Difficult vs. Easy).

To assess the stability of these differences across participants, we conducted Friedman tests for repeated measures and Wilcoxon signed-rank tests for follow-up pairwise comparisons, corrected using the Holm-Bonferroni adjustment. Effect sizes were reported for all analyses (η^2 for ANOVA/permutation tests and r for Wilcoxon tests), providing complementary evidence for interpreting the gaze features contributing to model predictions.

3.4. Feature-Based Classification and Validation

To compare performance across the three prediction targets, we implemented a feature-based supervised machine-learning pipeline. Feature-based models were selected because the inputs consisted of compact tabular oculometric features and the study focused on model interpretation rather than representation learning. We evaluated a diverse set of classifiers widely used for workload and task-state estimation [52], including baseline models (random and majority), distance-based models (k-NN), probabilistic models (Gaussian Naïve Bayes), linear/kernel models (SVM with linear and RBF kernels), and tree-ensemble models (Random Forest, XGBoost, and LightGBM). The majority baseline always predicted the most frequent class in the training data. LightGBM was selected for the detailed analyses because it is well suited to tabular data and supports post hoc feature-attribution methods. CatBoost and other popular boosting variants (e.g., AdaBoost, HistGradientBoosting) were

not evaluated, as they largely overlap in inductive bias with the included tree-ensemble implementations. A Bayesian hyperparameter search was explored during model development but did not yield a reliable improvement over LightGBM's default configuration. All performance results reported in this manuscript therefore use the library-default configuration, applied unchanged across all validation folds; the exploratory optimized configuration is not used in any reported performance result. Full search specification, final settings, software versions, and reproducibility details are reported in Supplementary Section S2. No resampling or class weighting was applied; because the classes are imbalanced, we report macro-averaged metrics alongside a majority-class baseline so that performance cannot be inflated by the majority class.

LOSO was used as the primary participant-held-out validation protocol, while LOGO served as an additional task-group-held-out robustness assessment. The two protocols answer different deployment questions. LOSO estimates performance for a new user on task types already represented in training, the situation of a system shipped without per-user calibration. LOGO estimates performance on a task family absent from training, the situation of a system carried into a new application context; it is the stricter question because oculomotor behavior is shaped by visual layout as much as by cognitive demand. Because the parameter configuration used for the reported models was fixed rather than selected within the outer validation folds, an inner hyperparameter-selection loop was not part of the reported evaluation pipeline. Nested cross-validation would be required if model or hyperparameter selection were performed within each outer training partition. Likewise, LOSO and LOGO are exhaustive over the predefined participants and task groups rather than repeated random partitions; repeated random subject-independent evaluation therefore addresses a different source of variability. What each protocol does and does not estimate is summarized in Supplementary Section S3.4, and a stage-by-stage audit of the data dependence and leakage-control procedure for each pipeline stage is provided in Supplementary Section S2.5. For Binary and EDR classification, LOGO groups were defined at the paradigm level, with easy and difficult variants of the same paradigm withheld together with a rest condition. For AVM classification, folds were grouped primarily by difficulty level while mixing paradigms to preserve class representation; because structurally similar tasks may therefore appear in different folds, AVM LOGO should not be interpreted as strict task-independent generalization. Detailed LOGO definitions and LOTO sanity-check results are provided in Supplementary Tables S2–S4. Performance was evaluated using accuracy, macro-F1, macro-precision, macro-recall, and macro-AUC. We also performed ablation studies using fixation-only, saccade-only, pupil-only, and blink-only feature subsets and their combinations to assess the contribution of each feature family.

3.5. Post Hoc Explainability and Behavioral Validation

To explain model behavior, we applied SHAP analysis to the trained LightGBM models. SHAP values quantify each feature's contribution to predictions both locally (per window) and globally (averaged across the dataset). To complement SHAP, we also inspected LightGBM's model-native feature importances (gain and split count). SHAP-identified features were compared with the statistical results described in Section 3.3 to assess whether model reliance aligned with observed differences across prediction targets.

4. Results

The results are presented in three parts. First, we compare predictive performance across the three targets and the LOSO and LOGO validation settings. Second, we report the feature-family ablation results. Finally, we present feature-level results from the statistical analyses, model-native feature importances, and SHAP.

4.1. Predictive Performance Across Targets and Validation Settings

Table 2 summarizes LightGBM performance for task-type classification (AVM), binary workload classification (Load versus Rest), and three-level workload classification (EDR) under LOSO and LOGO validation. Results are reported as mean ± standard deviation across validation folds; bootstrap 95% confidence intervals for every metric are provided in Supplementary Section S4.3. Macro-AUC is averaged over the folds in which it is defined, since a held-out participant occasionally lacks one class. The corresponding normalized confusion matrices are shown in Figure 3.

Table 2. Predictive performance across targets and validation settings. Accuracy, macro-F1, macro-precision, macro-recall, and macro-AUC are reported for task-type classification (AVM), binary workload classification (Load versus Rest), and three-level workload classification (EDR) using LightGBM and the majority baseline under LOSO and LOGO validation.

Classification Category	Split	Classifier	Accuracy (%)	F1	Precision (%)	Recall (%)	Macro-AUC
Task Type (3-class) (AVM)	LOSO	Majority	46.4 ± 17.0	0.212 ± 0.075	16.1 ± 7.0	33.0 ± 7.5	0.500 ± 0.000
	LOSO	LightGBM	85.9 ± 10.9	0.813 ± 0.137	82.3 ± 12.1	82.4 ± 14.4	0.954 ± 0.051
	LOGO	Majority	51.0 ± 7.6	0.224 ± 0.022	17.0 ± 2.5	33.3 ± 0.0	0.500 ± 0.000
	LOGO	LightGBM	83.4 ± 1.4	0.777 ± 0.022	79.0 ± 1.1	78.1 ± 1.9	0.922 ± 0.011
Load vs. NO Load (Binary)	LOSO	Majority	76.0 ± 11.1	0.467 ± 0.154	41.7 ± 17.1	53.7 ± 13.1	0.500 ± 0.000
	LOSO	LightGBM	81.4 ± 7.9	0.695 ± 0.116	73.9 ± 12.3	68.8 ± 11.1	0.848 ± 0.099
	LOGO	Majority	55.4 ± 14.7	0.351 ± 0.061	27.7 ± 7.3	50.0 ± 0.0	0.500 ± 0.000
	LOGO	LightGBM	63.9 ± 8.6	0.594 ± 0.053	66.6 ± 5.6	62.4 ± 2.3	0.737 ± 0.065
Load Level (3-class) (EDR)	LOSO	Majority	41.4 ± 10.1	0.204 ± 0.061	14.7 ± 5.5	34.9 ± 4.8	0.500 ± 0.000
	LOSO	LightGBM	56.4 ± 8.5	0.535 ± 0.098	55.6 ± 9.5	54.6 ± 9.6	0.762 ± 0.078
	LOGO	Majority	29.5 ± 9.0	0.149 ± 0.036	9.8 ± 3.0	33.3 ± 0.0	0.500 ± 0.000
	LOGO	LightGBM	44.3 ± 4.9	0.411 ± 0.054	43.7 ± 4.7	43.3 ± 4.7	0.631 ± 0.052

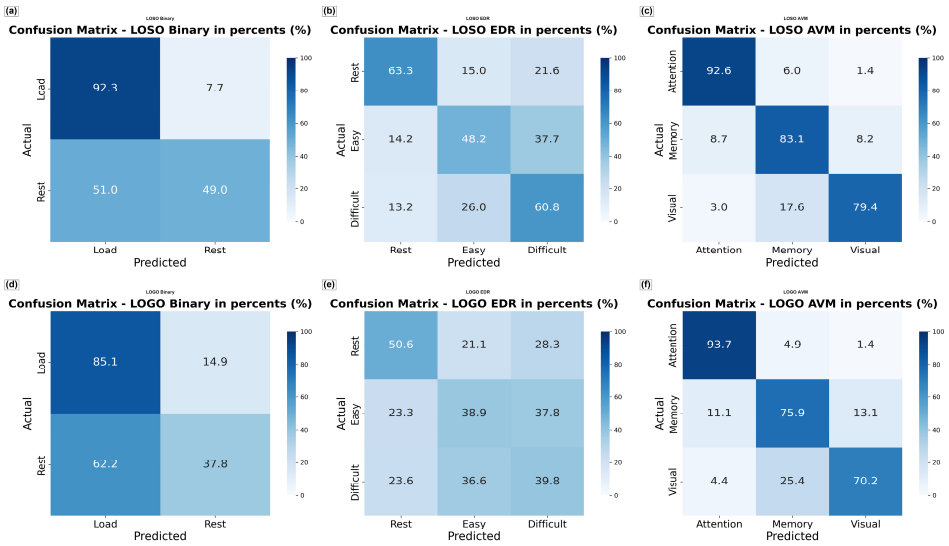

Figure 3. Normalized confusion matrices for: (**a**) binary workload classification under LOSO validation; (**b**) three-level workload (EDR) classification under LOSO validation; (**c**) task-type (AVM) classification under LOSO validation; (**d**) binary workload classification under LOGO validation; (**e**) three-level workload (EDR) classification under LOGO validation; and (**f**) task-type (AVM) classification under LOGO validation.

Task-type classification produced the highest accuracies, reaching 85.9% under LOSO and 83.4% under LOGO, compared with majority-baseline accuracies of 46.4% and 51.0%, respectively. Most LOGO errors occurred between the Memory and Visual-Spatial classes. Binary Load-versus-Rest classification achieved 81.4% accuracy under LOSO and 63.9% under LOGO, exceeding the corresponding majority baselines of 76.0% and 55.4%. Recall remained higher for Load than for Rest in both settings: 92.3% versus 49.0% under LOSO and 85.1% versus 37.8% under LOGO. Three-level EDR classification yielded the lowest accuracies, with 56.4% under LOSO and 44.3% under LOGO, compared with majority baselines of 41.4% and 29.5%. Easy and Difficult were the most frequently confused classes in both settings. Figure 3 presents the normalized confusion matrices for all three prediction targets.

4.2. Feature-Family Ablation Results

Figure 4 reports on classification accuracy for individual oculometric feature families and their combinations. The full feature set achieved the highest or near-highest accuracy across all evaluated targets and validation settings. Among the reduced feature sets, the combination of fixation and pupil features produced results closest to the full model. Under LOGO, this combination achieved 63.1% accuracy for binary workload classification, 42.9% for EDR classification, and 82.0% for task-type classification, compared with 63.9%, 44.3%, and 83.4%, respectively, for the full models. Among single feature families, fixation features achieved the highest accuracies across the three targets: fixation-only models reached 78.3% and 59.2% for binary workload classification, 51.6% and 40.9% for EDR classification, and 79.8% and 76.3% for task-type classification under LOSO and LOGO, respectively. Blink-only models produced the lowest results for EDR and task-type classification.

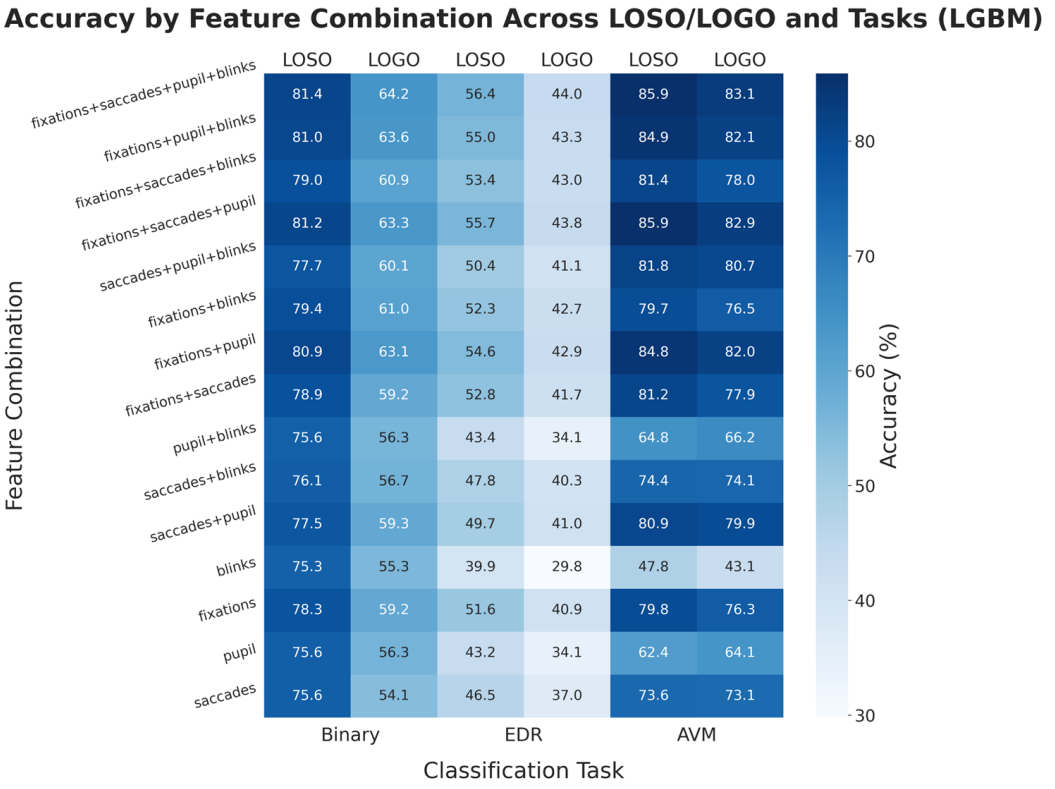

Figure 4. Accuracy of LightGBM models using individual oculometric feature families and their combinations under LOSO and LOGO validation for binary workload, three-level workload (EDR), and task-type (AVM) classification.

4.3. Feature-Level Statistical and Model-Based Results

4.3.1. Statistical Feature Differences

Feature-wise statistical analyses identified differences across task-type and workload classes. Complete results are provided in Supplementary Tables S5–S7, while selected distributions are shown in Figures 5, 6 and S1–S5. For task type, fixation dispersion showed the largest reported effect ($F = 599.2$, $p < 0.001$). Attention tasks had the lowest horizontal fixation dispersion, whereas Visual-Spatial tasks had the highest. Mean pupil size also differed across AVM classes ($F = 91.6$, $p < 0.001$), with the highest subject-relative values observed for Memory tasks.

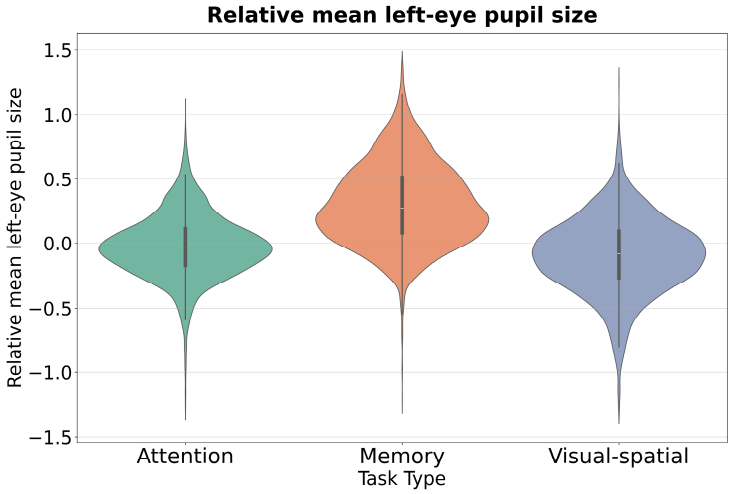

Figure 5. Subject-relative mean left-eye pupil size across task types (AVM) after outlier removal. Values were centered within participant by subtracting each participant's mean pupil size.

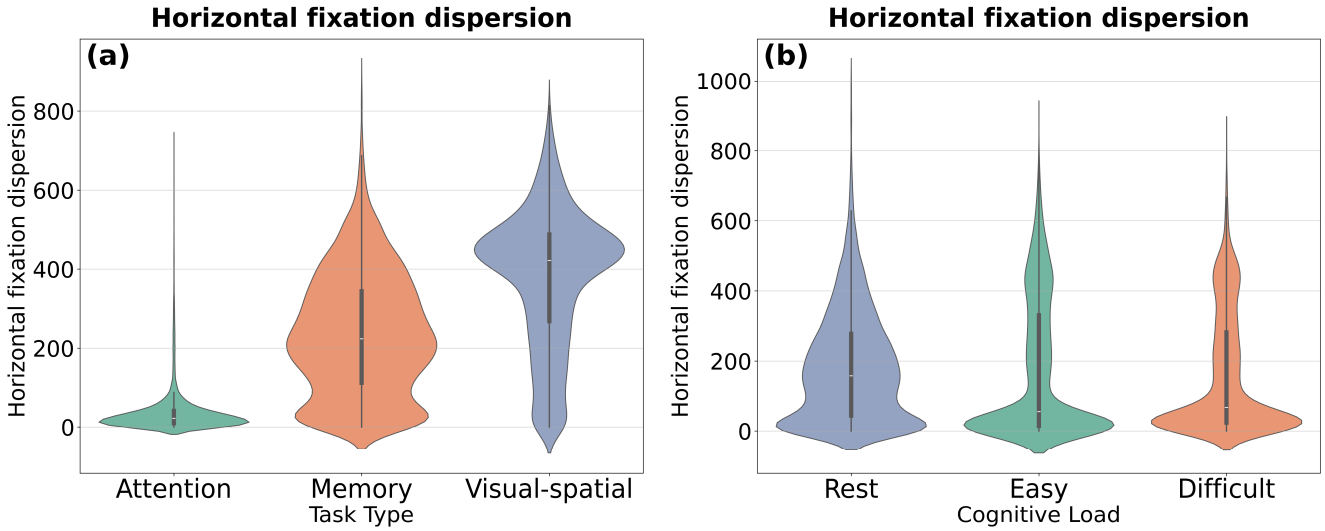

Figure 6. (**a**) Horizontal fixation dispersion across task types (AVM) and (**b**) workload classes (EDR). Horizontal lines indicate group medians.

For three-level workload classification, fixation dispersion differed across EDR classes ($F = 33.6$, $p < 0.001$), with the highest values observed during Rest. Mean pupil size also differed across EDR classes, though the effect was comparatively weak relative to fixation dispersion ($F = 3.2$, $p = 0.044$). Pairwise analyses showed larger pupil values for Difficult than Easy conditions and an increase in fixation count from Rest to Difficult (Supplementary Figure S2); despite the modest univariate effect, pupil size was subsequently found to be an important predictor in the classification models, as shown in

Supplementary Figures S10–S12. For binary workload classification, fixation dispersion differed between Load and Rest ($F = 49.8$, $p < 0.001$), whereas the univariate effect for mean pupil size was not significant ($F = 2.0$, $p = 0.161$). Missing sample counts were also higher during Rest than during Load.

4.3.2. Model-Native Feature Importance

LightGBM feature importance was examined using gain and split count. Figure 7 shows the gain-based ranking for binary workload classification under LOGO, while results for all three targets under LOSO and LOGO are provided in Supplementary Figures S10–S12. For the binary LOGO model, subject-relative mean left-eye pupil size had the highest gain, followed by vertical fixation dispersion, mean fixation duration, and horizontal fixation dispersion. Saccade-, blink-, fixation-variability-, and missingness-related features also appeared among the ten highest-ranked features. The precise ranking differed across targets, validation settings, and importance measures. Fixation-dispersion, pupil-size, and fixation-duration features nevertheless appeared repeatedly among the highest-ranked variables.

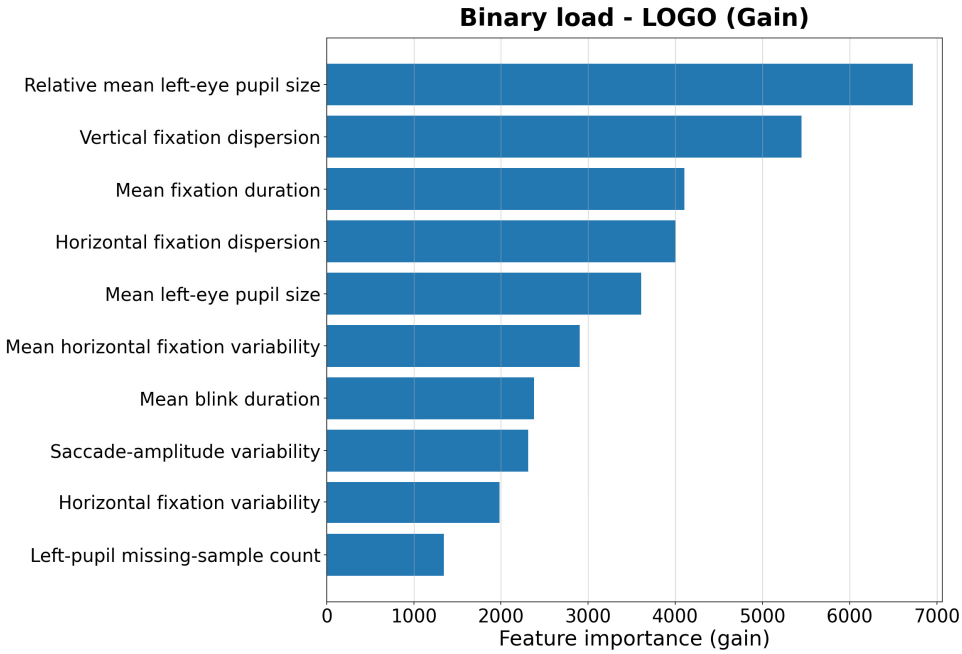

Figure 7. Gain-based LightGBM feature importance for binary workload classification under LOGO validation. The ten highest-ranked features are shown. Fixation-dispersion, pupil-size, and fixation-duration features appeared repeatedly among the highest-ranked features across different validation settings, importance measures, and targets.

4.3.3. SHAP Results and Statistical Correspondence

SHAP analyses identified fixation dispersion, subject-relative mean pupil size, and fixation duration among the highest-ranked features across the three prediction targets. Missingness-related variables also appeared among the higher-ranked features for EDR classification. Class-specific SHAP results for AVM and EDR are provided in Supplementary Figures S7 and S8. Figure 8 presents the SHAP summary for binary workload classification. Higher and lower feature values contributed differently across the Load and Rest classes, with fixation-dispersion and pupil-related features appearing among the five highest-ranked variables.

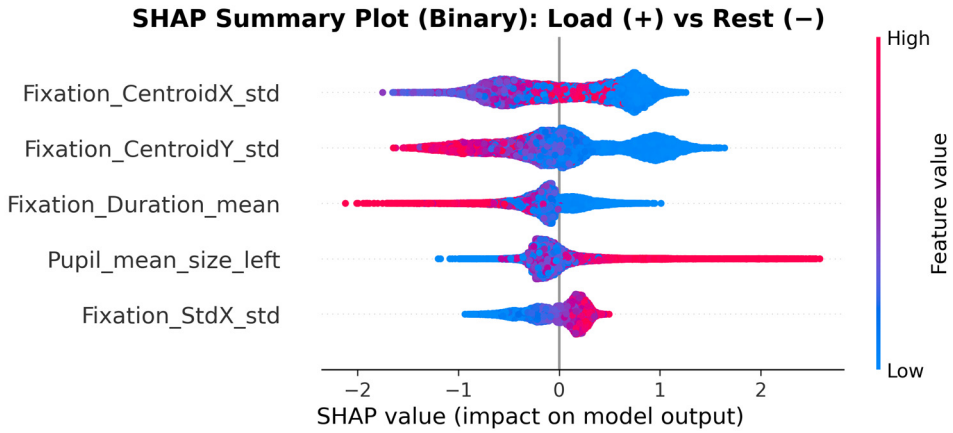

Figure 8. SHAP summary plot for binary workload classification. The plot shows the five features with the highest global SHAP importance for the Load-versus-Rest model.

Table 3 compares selected statistical results, SHAP observations, and model-native feature importances. Fixation dispersion appeared consistently across the three analyses for AVM, EDR, and binary workload classification. Pupil-size features also appeared repeatedly in the model-based analyses, although agreement with the univariate statistical results varied across targets. In particular, mean pupil size was model-relevant for binary classification despite a non-significant ANOVA result.

Table 3. Comparison of selected statistical results, SHAP observations, and LightGBM feature importances for task-type (AVM), three-level workload (EDR), and binary workload classification. Detailed statistical results are provided in Supplementary Section S4.1.

Oculometric Feature	ANOVA Results F/p	SHAP Observations	LGBM Feature Importances
		Feature family (AVM)	
Fixation dispersion	599.2/0.000	low for attention; high for visual & memory	multiple features consistently in the top 3
Mean pupil size	91.6/0.000	high for memory, low for attention/visual	consistently among top 2
		Feature family (Binary)	
Fixation dispersion	49.8/0.000	low for load, high for rest	multiple features consistently in the top 3
Mean pupil size	2.0/0.161	high for load, low for rest	strong model evidence (top split features, high-gain contributors)
		Feature family (EDR)	
Fixation dispersion	33.6/0.000	low for difficult, high for rest	multiple features consistently in the top 3
Mean pupil size	3.2/0.044	the higher the load, the bigger the pupil	strongest and most frequent (first by split)

5. Discussion

This study examined how prediction target and validation strategy jointly shape what can be inferred about task type and workload-related state from ocular metrics alone. The combined findings support our central contribution: predictive reliability decreases systematically as the target moves from directly observable task structure toward internal workload state, while a compact set of ocular markers, fixation dispersion, fixation dynamics, and pupil size, remains the consistent explanatory basis throughout. Taken together, these findings indicate that eye tracking can support low-burden adaptive human-

computer interaction systems for coarse-grained state adaptation, but that reliable fine-grained workload inference remains challenging due to task-related variability and limited discriminability in ocular signals.

5.1. Ocular Markers and Explainability Across the Prediction Hierarchy

A major strength of this work lies in the convergence between statistical analyses and model-based explainability. Across task types and load levels, three oculometric families consistently emerged as recurring explanatory markers: fixation dispersion, fixation dynamics (duration and count), and pupil size. Statistical analyses showed clear differences across tasks (Table 3). Fixation dispersion was widest in memory and visual-spatial tasks and in rest trials, reflecting more exploratory scanning. Fixation duration and count tracked task engagement and processing strategy, with attention tasks characterized by more frequent and longer fixations. Pupil size increased in memory tasks and under higher load, aligning with established task-evoked pupillary response mechanisms associated with locus coeruleus activity [27,28].

These behavioral patterns were mirrored in model explanations. SHAP analyses (Figures 8, S7 and S8; Table 3; Supplementary Section S5) consistently ranked fixation dispersion, subject-relative pupil size, and fixation duration among the strongest contributors across all three classification targets. SHAP-identified features showed substantial convergence with the statistical analyses, increasing confidence that the models relied on behaviorally interpretable ocular patterns. This convergence does not, however, exclude dataset-specific confounding or establish causal specificity. Three features recurred across both validation protocols and all three targets: horizontal and vertical fixation-centroid dispersion, subject-relative mean pupil size, and mean fixation duration. These have the strongest claim to general applicability in this dataset, although their recurrence is associational and does not establish task-invariance. Because features within an oculometric family remain moderately intercorrelated even after pruning, attribution can be distributed arbitrarily among near-equivalent members of a family; we therefore interpret feature families rather than individual rank positions, a point quantified in Supplementary Section S1.4.

While omnibus pupil-size effects were modest due to inter-subject variability, the use of subject-relative pupil features allowed the models to extract reliable within-participant effects, captured both in SHAP values and in pairwise Wilcoxon tests, providing convergent evidence that the classification decisions are associated with physiologically interpretable ocular signals. Temporal profiles of ocular metrics further indicated that workload-related markers are dynamic rather than static: fixation dispersion, pupil dilation, and fixation behavior changed over the course of a trial, revealing adaptive shifts in attention and arousal, echoing predictions from cognitive adaptation and effort-regulation theories and extending traditional static-load frameworks often used in HCI [26,27,31]. These results indicate that fixation and pupil features provide a compact and interpretable basis for workload-related and task-type inference in the present dataset.

5.2. Generalization Across Participants and Task Groups

The results show that the reliability of eye-tracking-based inference depended strongly on both the prediction target and the validation protocol. Task-type classification produced the strongest participant-held-out performance, whereas binary load-versus-rest classification showed moderate and class-asymmetric discrimination, and three-level workload (EDR) classification remained substantially more difficult. Under LOSO, the corresponding accuracies were 85.9%, 81.4%, and 56.4%, respectively. Performance decreased under LOGO for the two workload-related targets, reaching 63.9% for binary classification and 44.3% for EDR classification. The practical significance of these differences varies by target. For task

type, the drop is small (85.9% to 83.4%, 2.5 percentage points), so task-context recognition appears robust to the grouping used. For binary load, the drop is substantial (81.4% to 63.9%, 17.5 points), leaving a margin of roughly 8.5 points over the corresponding majority baseline, and the accuracy confidence intervals for model and baseline overlap under this protocol, although the macro-F1 and AUC intervals do not (Supplementary Section S4.3). For three-level workload, the drop is 12.1 points, from 56.4% to 44.3%. These results suggest that task-context recognition is the more plausible target for practical gaze-based adaptation in the present setting. Coarse load-versus-rest inference may also be useful when the relevant task families are represented during model development, whereas the observed paradigm-held-out performance indicates that fine-grained workload estimation should not yet be treated as reliable for unseen task families. An independent study transferring eye-tracking load classifiers from a working-memory task to an emergency simulation game reported 63.78–67.25% for an easy-versus-difficult distinction [44], closely matching the binary cross-task-group result obtained here.

Task-type classification achieved 85.9% accuracy under LOSO and 83.4% under the adopted AVM LOGO evaluation, indicating that task-related oculomotor patterns were relatively consistent across participants. Because the AVM LOGO folds were organized primarily by difficulty level and could contain structurally related paradigms in both training and test partitions, this result should be interpreted as robustness to the specified grouping rather than strict generalization to entirely unseen tasks; alternative LOTO analyses and their associated limitations are reported in Supplementary Sections S3.3 and S3.4. The confusion matrices in Figure 3 further illustrate the target-specific error patterns. Load windows were identified more reliably than Rest windows in the binary models, whereas Easy and Difficult conditions were frequently confused in the three-level EDR models. Attention was the most readily identified AVM class, while errors occurred more frequently between Memory and Visual-Spatial tasks.

Comparisons with other published studies should be interpreted cautiously because task design, participant populations, sensing configurations, target definitions, and validation protocols differ widely, and no published study evaluates eye-tracking-based classification on the same task battery and participant population used here. The closest available comparison point for the three-level workload target is Kaczorowska et al., who classified three experimentally manipulated workload levels, defined by digit-symbol-substitution task difficulty, from eye-tracking features using a subject-independent train/test split repeated over 200 resamples, reporting F1 up to 0.95 on the complete feature set and 0.97 after feature selection [8]. The pipeline reported here achieves a substantially lower macro-F1 of 0.535 for the three-level EDR target under LOSO. This gap most plausibly reflects task design and sample granularity rather than model quality: their three levels were separated by a large, discrete manipulation of symbol-set size and test duration and aggregated into only 87 observations at the level of a whole 90- or 180-s test, whereas the three-level target evaluated here spans several task paradigms and short, three-second windows with continuous within-condition variability. Comparisons with recent deep-learning work require similar caution: subject-independent evaluation of a multimodal driving dataset lowered accuracy by roughly 8 to 10 percentage points relative to standard 10-fold cross-validation on the same data [37], illustrating how much the validation protocol alone can move reported figures, independent of model or modality. The present eye-only approach nevertheless shows that useful task-context and coarse engagement information can be extracted without additional physiological sensors. Its principal advantage lies not in outperforming multimodal systems directly, but in reducing sensing burden while retaining semantically interpretable oculometric inputs.

The findings also clarify the limits of eye-only sensing. Performance under participant-held-out validation was consistently stronger than performance under the paradigm-held-out evaluations used for the workload-related targets, suggesting that ocular-feature relationships transfer more readily across participants than across changes in task structure. The contrast between the targets indicates that task identity is strongly encoded in gaze behavior, likely through differences in visual layout and task-specific viewing strategy, whereas experimentally defined workload conditions produce subtler and less transferable ocular differences. The error structure in Figure 3 is consistent with this reading. Easy and Difficult conditions are confused most often because they share visual structure and differ only in degree of demand, so the oculomotor signature separating them is weaker than the one separating either from Rest. Memory and Visual-Spatial tasks are confused because both encourage broad exploratory scanning, unlike the sustained central fixation that makes Attention tasks readily identifiable. Rest windows are recovered less reliably than Load windows in the binary models, reflecting both their smaller share of the data and their overlap with low-demand task windows in gaze statistics. Per-class and task-level error rates, together with local SHAP explanations for representative correct, incorrect, and borderline windows, are reported in Supplementary Section S9.

5.3. Practical, Theoretical and Ethical Implications

Ocular metrics have the potential to serve as a scalable, unobtrusive source of coarse-grained state information in everyday HCI scenarios, without requiring intrusive wearable biosensors, reducing hardware burden relative to multimodal state-monitoring systems that depend on EEG, EDA, or other physiological sensors. Because fixation dispersion and pupil-linked arousal can track coarse-level workload-related changes even under task heterogeneity, they suggest, as motivating scenarios for future validation rather than as capabilities demonstrated here, adaptive interfaces across several domains: modulating instructional pacing in learning environments [53,54], timing notifications or reorganizing information in productivity contexts [55], and detecting attentional lapses in safety-critical domains such as driving, air-traffic monitoring, and XR applications [12,14,29]. Binary engagement detection may be the most plausible near-term candidate for real-time adaptation, subject to the cross-task limitation reported in Section 5.2. Each of these scenarios would require validation in its own deployment context before any such claim could be made; none is demonstrated by the present controlled study. Grounding predictions in interpretable oculometric features supports the design of transparent, user-centered adaptive systems [39,40]. Classifier-level computational cost is unlikely to be the principal barrier in this setting. Inference requires evaluation of a gradient-boosted model over a few dozen scalar features per window; accordingly, the more consequential latency components are likely to be the 3-s analysis window and upstream signal processing and feature extraction. Measured on a CPU-only workstation, median single-window inference latency was 0.76–0.78 ms, roughly three orders of magnitude shorter than the 3 s analysis window, while mean per-fold training time ranged from 0.26 s to 1.75 s and serialized models occupied at most 1.0 MB. Full timings, memory use, a comparison against alternative classifiers, and a training-set-size scalability curve are reported in Supplementary Section S10.

From a theoretical standpoint, the observed patterns are consistent with classical and contemporary accounts of effort and visual attention [18,21,26]; fixation behavior reflects attentional breadth and integration demands, pupil dilation is associated with phasic changes in effort and arousal, and blink suppression varies with task-specific visual demand. Convergence between SHAP explanations, statistics, and psychophysiological theory provides convergent support for the behavioral interpretability of the model predic-

tions, an alignment critical for transparent, workload-aware systems and directly relevant to explainability concerns in adaptive HCI.

The ability to infer state from ocular metrics also raises ethical considerations. Because ocular signals can reveal aspects of a user's internal state without explicit awareness, workload-aware systems must ensure transparency, informed consent, and clear communication about data collection and use. Misclassification or over-interpretation of state estimates may introduce risks in high-stakes contexts, underscoring the need for human oversight and conservative deployment; privacy safeguards are essential as eye tracking becomes embedded in personal devices, vehicles, and XR headsets. Emphasizing interpretable modeling and task-general validation supports more responsible, accountable use of workload-aware technologies in real-world HCI.

6. Limitations and Future Work

Several limitations should be acknowledged to contextualize these findings. A primary one concerns ecological validity: data were collected in controlled laboratory settings using a high-quality desktop eye tracker, ensuring precise measurement and consistent labeling, but real-world environments introduce substantial noise—head movement, variable illumination, display differences that can affect gaze and especially pupil-based measures. Future work should evaluate the framework in naturalistic contexts (mobile or wearable eye tracking, XR headsets, in-the-wild screen-based tasks) to assess robustness under everyday conditions.

A second limitation concerns ground-truth labeling: the workload-related target relied mainly on task difficulty by experimental design (the EDR scheme), which, while validated with subjective ratings (ISA, NASA-TLX), does not always map uniformly onto workload state across individuals given variation in strategy, prior knowledge, motivation, or fatigue. Relatedly, missingness-derived features ranked among the higher-importance predictors for three-level workload classification (Section 4.3.3); while the feature-family ablation (Section 4.2) shows a fixation-and-pupil subset achieves performance close to the full model, we did not run a dedicated ablation excluding missingness-derived features specifically, and cannot rule out that some EDR classification performance partly reflects gaze-tracking-quality artifacts rather than physiological signal. Future work should integrate richer or continuous load measurements (secondary-task performance, physiological markers, real-time self-reports) and explicitly test performance with missingness-derived features excluded. This concern is concrete rather than hypothetical: participants wore additional sensing equipment alongside the eye tracker, which the dataset publication reports as occasionally occluding the pupil [17], so tracking-quality variation is a plausible contributor to the informativeness of missingness-derived features. We did not conduct a dedicated robustness analysis under simulated gaze noise, calibration drift, or reduced sampling rates; such an analysis would need degradation models representative of the target deployment setting rather than generic additive noise, and we regard it as a separate study.

The task design also introduces potential confounds: attention tasks required sustained central fixation, whereas visual-spatial and memory tasks encouraged broader exploratory scanning, risking a partial conflation of task identity with oculomotor demand. Future studies should examine attention paradigms without strict central-fixation requirements, or match visual-spatial and memory tasks for gaze constraints, to disentangle task effects from biological and task-driven oculomotor differences. The participant cohort, while reasonably large, was relatively homogeneous (primarily young adults), and the task suite, though spanning three cognitive domains, does not capture the full variability of real-world activity—multitasking, emotional state, environmental uncertainty that shapes workload in everyday HCI contexts. Extending analyses to more diverse populations and

more complex, dynamic tasks will be important for broader generalizability. The analyzed cohort of 54 participants, recorded in a controlled single-tracker laboratory setting, also bounds generalizability in a way the fold-level variability makes visible: per-participant accuracy varied substantially under LOSO, as the standard deviations in Table 2 and the confidence intervals in Supplementary Section S4.3 show. We found no association between per-participant performance and the participant characteristics available to us, or an index of individual gaze variability, after correction for multiple comparisons; the dataset publication likewise reports no effect of sex, education, self-reported focus, tiredness or sleep duration on task performance [17]. Visual correction and prior task familiarity were not recorded, so their influence could not be assessed and remains an open question.

Sampling design introduces two further caveats. Consecutive 3-s windows overlap by 25%, so samples within a trial are not fully independent; this is mitigated by participant- and task-group-held-out validation (LOSO/LOGO), which prevents train-test leakage at the window level, but residual within-trial correlation may still modestly inflate reported performance relative to fully independent sampling. Subject-relative features are referenced to each participant's own recording-level summary; this reduces between-participant baseline differences but should be distinguished from preprocessing parameters estimated exclusively from the outer-fold training population. A further implementation-level limitation is that the archived LOSO and LOGO analyses do not use perfectly identical feature sets: LOSO uses the final target-specific sets of 58 features for the Binary and EDR targets and 59 for AVM, whereas LOGO uses the preceding shared 61-feature set. The additional LOGO variables are highly correlated near-duplicates of retained features, so this difference primarily concerns residual redundancy rather than access to a distinct source of information; nevertheless, LOSO–LOGO performance differences should not be attributed exclusively to the validation split.

Separately, task order and data volume were not fully balanced: attention tasks contributed the largest share of data and differed structurally from memory and visual-spatial tasks, which may have influenced LOGO generalization patterns; counterbalanced presentation and more balanced domain sampling would strengthen future cross-task evaluations. Three further scope limits should be noted. Explanation stability was assessed across validation protocols, importance measures, and prediction targets rather than across random seeds, so a systematic seed-level stability analysis remains outstanding. The preliminary classifier screening was not designed as a formal statistical comparison among algorithms, and we therefore do not make claims of statistically significant superiority of LightGBM over the alternative classifiers. Finally, a stricter task-type grouping that withheld complete paradigms for every class simultaneously would yield only two folds, with correspondingly small training partitions and unstable, wide-interval estimates; designing a task battery that supports such a grouping is a worthwhile aim for future data collection.

Finally, the study focused intentionally on unimodal eye-tracking features, prioritizing explainability and scalability over multimodal integration; this isolates the contribution of ocular behavior but limits resolution for fine-grained load estimation. Multimodal fusion is well established in this literature and does improve fine-grained load discrimination [13–15,17]; combining eye tracking with complementary signals (EEG, ECG, keystroke dynamics, click behavior, environmental context) could help differentiate effort from affective arousal or fatigue and may improve challenging multi-level classification. The unimodal restriction adopted here is therefore a deliberate design choice rather than an oversight: it isolates what gaze alone supports, and the resulting figures bound what an eye-tracking component can be expected to contribute within a larger multimodal system. More broadly, since models generalized well across participants but only moderately across task groups for binary load and task type, with fine-grained load levels remaining more

context-dependent, the present results indicate that fully task-agnostic workload-related inference from ocular metrics alone remains difficult under the evaluated conditions; domain adaptation, task-invariant representation learning, personalized calibration, or hybrid modeling strategies are promising directions for improving cross-context generalization. Taken together, advancing ecological evaluation, expanding demographic and task diversity, incorporating complementary modalities, and developing adaptive or personalized modeling approaches will help build scalable, robust, and explainable ocular-based task-type and workload-related inference systems suitable for real-world HCI.

A further direction concerns the temporal horizon of inference. The present analysis, like most work in this area, estimates state from a single short window. Gaze features accumulated across sessions might instead support anticipatory inference, predicting fatigue onset, performance decline, or learning progress before it becomes behaviorally evident, in the way that models built on prior observations have been used to estimate future risk in other domains such as medical imaging [56]. Such applications would raise rather than lower the interpretability stakes, because a prediction about a future state cannot be checked against a concurrent observation. They would also inherit the concern documented here in compounded form: if a target conflates task structure with internal state within a single window, that confound accumulates across a longitudinal horizon rather than averaging out. Establishing what a gaze-based target actually encodes is therefore a prerequisite for anticipatory modeling, not an alternative to it.

7. Conclusions

This study examined how prediction-target definition and validation strategy affect what can be inferred from feature-based eye-tracking models. Using a controlled dataset, we developed and evaluated an explainable, eye-only analytical framework that combines statistical characterization, feature-based predictive modeling, post hoc explainability, participant-held-out validation, and target-specific group-held-out robustness evaluation.

The findings showed a clear target-dependent performance pattern. Task type was classified most reliably, binary load-versus-rest classification achieved moderate but class-asymmetric performance, and three-level workload classification remained substantially more challenging. This pattern indicates that gaze features more readily encode observable differences in task structure and engagement than fine-grained experimentally defined workload conditions. Predictive reliability was also validation-dependent: performance remained comparatively strong under participant-held-out evaluation but decreased under paradigm-held-out evaluation for the workload-related targets. Task-type performance remained high under the adopted AVM group-held-out scheme; however, because that scheme allowed structurally related paradigms to occur in both training and test folds, it should be interpreted as grouping robustness rather than strict generalization to unseen tasks.

Fixation dispersion, fixation duration, fixation count, and participant-relative pupil measures repeatedly contributed to model predictions across several targets and validation settings. Their recurrence in SHAP analyses and statistical comparisons supports the behavioral interpretability of these associations, although it does not establish that the features are fully task-invariant or causally specific to workload.

These findings have practical relevance for adaptive interfaces, learning technologies, driver monitoring, and human-machine teaming. In these settings, the present results motivate further evaluation of coarse, task-context-aware adaptation using eye tracking alone, subject to validation in the intended deployment setting. More broadly, the results demonstrate that conclusions drawn from gaze-based machine-learning systems depend jointly on how the target construct is operationalized, how validation groups are defined,

and how model explanations are interpreted. Eye tracking appears promising for task-context recognition and coarse engagement-related adaptation, but claims concerning task-independent, fine-grained workload inference should remain conservative. These findings provide methodological guidance for the development and evaluation of transparent human-centered AI systems based on ocular behavior.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ai7080325/s1.

Author Contributions: Conceptualization, T.B., S.G., M.G. and G.S.; methodology, T.B. and S.G.; software, T.B.; data curation, T.B.; formal analysis, T.B. and S.G.; investigation, T.B.; writing-original draft preparation, S.G.; writing-review and editing, T.B., S.G., M.L., M.G. and G.S.; visualization, T.B.; supervision, M.L., M.G. and G.S.; project administration, T.B., S.G., M.G. and G.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Swiss Government Excellence Scholarship under Grant No. 2024.0108; Swiss National Science Foundation under Grant No. 216405; and Swiss National Science Foundation under Grant No. 214991, co-funded by the Slovenian Research and Innovation Agency (ARIS) under Grant agreement N1-0319.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the Università della Svizzera italiana for the project "Cognitive load estimation from wearable devices" (SNSF project XAI-PAC), under which the CLUES dataset analysed here was collected (protocol code CE_2024_27, approved on 18 November 2024).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The analyses in this study are based on the CLUES dataset, which was collected by a broader team of authors and is currently under revision as a separate dataset publication. While the dataset is not yet publicly available, representative samples and the full dataset can be shared with researchers and reviewers upon reasonable request to the corresponding author. Public release is planned to follow the acceptance of the dataset publication. The feature-extraction library used in this work, EyeTrackLib, is already publicly available [50], and the experiment, modeling, and analysis code supporting the reported results will be released publicly upon acceptance.

Acknowledgments: We thank all participants for contributing to the data collection effort. We also acknowledge the support of the participating institutions and research facilities that enabled this study. During the preparation of this work, the authors used ChatGPT 4.o, ChatGPT 5.5, GPT-5.3-Codex, and GitHub Copilot (Claude Sonnet 4.5) to brainstorm ideas, perform sanity checks, proofread manuscript text, and assist in writing code for experiments and analyses. All content generated with these tools was reviewed and edited by the authors, who take full responsibility for the final published article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ehsan, U.; Riedl, M.O. *Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach*; Springer: Cham, Switzerland, 2020; pp. 449–466. [CrossRef]
2. Kosch, T.; Karolus, J.; Zagermann, J.; Reiterer, H.; Schmidt, A.; Woźniak, P.W. A survey on measuring cognitive workload in human-computer interaction. *ACM Comput. Surv.* **2023**, *55*, 283. [CrossRef]
3. Šola, H.M.; Qureshi, F.H.; Khawaja, S. AI eye-tracking technology: A new era in managing cognitive loads for online learners. *Educ. Sci.* **2024**, *14*, 933. [CrossRef]
4. Duchowski, A.T. *Eye Tracking Methodology: Theory and Practice*; Springer: Berlin/Heidelberg, Germany, 2017.
5. Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv* **2017**, arXiv:1702.08608v2. [CrossRef]

6. Kumar, A.; Howlader, P.; Garcia, R.; Weiskopf, D.; Mueller, K. Challenges in interpretability of neural networks for eye movement data. In Proceedings of the ACM Symposium on Eye Tracking Research and Applications, Stuttgart, Germany, 2–5 June 2020; pp. 1–5. [CrossRef]
7. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4768–4777.
8. Kaczorowska, M.; Plechawska-Wójcik, M.; Tokovarov, M. Interpretable machine learning models for three-way classification of cognitive workload levels for eye-tracking features. *Brain Sci.* **2021**, *11*, 210. [CrossRef] [PubMed]
9. Hart, S.G. NASA-task load index (NASA-TLX); 20 years later. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **2006**, *50*, 904–908. [CrossRef]
10. Marchand, C.; De Graaf, J.B.; Jarrassé, N. Measuring mental workload in assistive wearable devices: A review. *J. Neuroeng. Rehabil.* **2021**, *18*, 160. [CrossRef] [PubMed]
11. Martinez-Cedillo, A.P.; Gavrila, N.; Mishra, A.; Geangu, E.; Foulsham, T. Cognitive load affects gaze dynamics during real-world tasks. *Exp. Brain Res.* **2025**, *243*, 82. [CrossRef] [PubMed]
12. Pillai, P.; Balasingam, B.; Kim, Y.H.; Lee, C.; Biondi, F. Eye-gaze metrics for cognitive load detection on a driving simulator. *IEEE/ASME Trans. Mechatron.* **2022**, *27*, 2134–2141. [CrossRef]
13. Cabañero Gómez, L.; Hervás, R.; González, I.; Villarreal, V. Studying the generalisability of cognitive load measured with EEG. *Biomed. Signal Process. Control* **2021**, *70*, 103032. [CrossRef]
14. Fridman, L.; Reimer, B.; Mehler, B.; Freeman, W.T. Cognitive load estimation in the wild. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; pp. 1–9.
15. İşbilir, E.; Çakır, M.P.; Acartürk, C.; Tekerek, A.Ş. Towards a multimodal model of cognitive workload through synchronous optical brain imaging and eye tracking measures. *Front. Hum. Neurosci.* **2019**, *13*, 375. [CrossRef] [PubMed]
16. Pušica, M.; Mijović, B.; Leva, M.C.; Gligorijević, I. Machine learning performance in EEG-based mental workload classification across task types: A systematic review. *Front. Neuroergonomics* **2025**, *6*, 1621309. [CrossRef] [PubMed]
17. Krstevska, A.; Goyal, S.; Fiorini, L.; Bombassei De Bona, F.; Kirilenko, D.; Li, M.; Trojer, S.; Božak, T.; Anžur, Z.; Luštrek, M.; et al. CLUES: Cognitive Load Understanding Through Experimental Sensing Dataset. Available online: https://drive.google.com/file/d/1S6CxxPMzV_RsWmQoApftxPMDWOBq6Sbr/view?usp=sharing (accessed on 22 January 2026).
18. Mahanama, B.; Jayawardana, Y.; Rengarajan, S.; Jayawardena, G.; Chukoskie, L.; Snider, J.; Jayarathna, S. Eye movement and pupil measures: A review. *Front. Comput. Sci.* **2022**, *3*, 733531. [CrossRef]
19. Skaramagkas, V.; Giannakakis, G.; Ktistakis, E.; Manousos, D.; Karatzanis, I.; Tachos, N.S.; Tripoliti, E.; Marias, K.; Fotiadis, D.I.; Tsiknakis, M. Review of eye tracking metrics involved in emotional and cognitive processes. *IEEE Rev. Biomed. Eng.* **2021**, *16*, 260–277. [CrossRef] [PubMed]
20. Ktistakis, E.; Skaramagkas, V.; Manousos, D.; Tachos, N.S.; Tripoliti, E.; Fotiadis, D.I.; Tsiknakis, M. COLET: A dataset for COgnitive workLoad estimation based on eye-tracking. *Comput. Methods Programs Biomed.* **2022**, *224*, 106989. [CrossRef] [PubMed]
21. Guo, Y.; Helmert, J.R.; Graupner, S.-T.; Pannasch, S. Eye movement patterns in complex tasks: Characteristics of ambient and focal processing. *PLoS ONE* **2022**, *17*, e0277099. [CrossRef] [PubMed]
22. Bulling, A.; Ward, J.A.; Gellersen, H.; Tröster, G. Eye movement analysis for activity recognition using electrooculography. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 741–753. [CrossRef] [PubMed]
23. Bulling, A.; Weichel, C.; Gellersen, H. EyeContext: Recognition of high-level contextual cues from human visual behaviour. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France, 27 April–2 May 2013; pp. 305–308.
24. Cole, Z.J.; Kuntzelman, K.M.; Dodd, M.D.; Johnson, M.R. Convolutional neural networks can decode eye movement data: A black box approach to predicting task from eye movements. *J. Vis.* **2021**, *21*, 9. [CrossRef] [PubMed]
25. Skaramagkas, V.; Ktistakis, E.; Manousos, D.; Tachos, N.S.; Kazantzaki, E.; Tripoliti, E.E.; Fotiadis, D.I.; Tsiknakis, M. Cognitive workload level estimation based on eye tracking: A machine learning approach. In Proceedings of the 2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE), Kragujevac, Serbia, 25–27 October 2021; pp. 1–5.
26. Sweller, J. Cognitive load theory. In *Psychology of Learning and Motivation*; Elsevier: Amsterdam, The Netherlands, 2011; Volume 55, pp. 37–76.
27. Van der Wel, P.; Van Steenbergen, H. Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychon. Bull. Rev.* **2018**, *25*, 2005–2015. [CrossRef] [PubMed]
28. Rodemer, M.; Karch, J.; Bernholt, S. Pupil dilation as cognitive load measure in instructional videos on complex chemical representations. *Front. Educ.* **2023**, *8*, 1062053. [CrossRef]
29. Balta, E.; Psarrakis, A.; Vatakis, A. The effects of increased mental workload of air traffic controllers on time perception: Behavioral and physiological evidence. *Appl. Ergon.* **2024**, *115*, 104162. [CrossRef] [PubMed]

30. Borys, M.; Plechawska-Wójcik, M. Eye-tracking metrics in perception and visual attention research. *EJMT* **2017**, *3*, 11–23.
31. Zagermann, J.; Pfeil, U.; Reiterer, H. Studying eye movements as a basis for measuring cognitive load. In Proceedings of the Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; pp. 1–6.
32. Souchet, A.D.; Philippe, S.; Lourdeaux, D.; Leroy, L. Measuring visual fatigue and cognitive load via eye tracking while learning with virtual reality head-mounted displays: A review. *Int. J. Hum.-Comput. Interact.* **2022**, *38*, 801–824. [CrossRef]
33. Sáiz-Manzanares, M.C.; Marticorena-Sánchez, R.; Martin Anton, L.J.; González-Díez, I.; Carbonero Martín, M.Á. Using eye tracking technology to analyse cognitive load in multichannel activities in university students. *Int. J. Hum.-Comput. Interact.* **2024**, *40*, 3263–3281. [CrossRef]
34. Grinsztajn, L.; Oyallon, E.; Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 507–520. [CrossRef]
35. Shwartz-Ziv, R.; Armon, A. Tabular data: Deep learning is not all you need. *Inf. Fusion* **2022**, *81*, 84–90. [CrossRef]
36. Borisov, V.; Leemann, T.; Seßler, K.; Haug, J.; Pawelczyk, M.; Kasneci, G. Deep neural networks and tabular data: A survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 7499–7519. [CrossRef] [PubMed]
37. Angkan, P.; Behinaein, B.; Mahmud, Z.; Bhatti, A.; Rodenburg, D.; Hungler, P.; Etemad, A. Multimodal brain–computer interface for in-vehicle driver cognitive load measurement: Dataset and baselines. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 5949–5964. [CrossRef]
38. Rolon-Merette, T.; Hardy Joseph, G.; Karran, A.J.; Belanger, C.; Coursaris, C.K.; Sénécal, S.; Léger, P.-M. Towards a cross-participant cognitive load classification using eye tracking and deep learning. *Int. FLAIRS Conf. Proc.* **2026**, *39*, 141863. [CrossRef]
39. Abdul, A.; Vermeulen, J.; Wang, D.; Lim, B.Y.; Kankanhalli, M. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; pp. 1–18. [CrossRef]
40. Joyce, D.W.; Kormilitzin, A.; Smith, K.A.; Cipriani, A. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *npj Digit. Med.* **2023**, *6*, 6. [CrossRef] [PubMed]
41. Liu, S.; Lu, C.; Alghowinem, S.; Gotoh, L.; Breazeal, C.; Park, H.W. Explainable AI for suicide risk assessment using eye activities and head gestures. In Proceedings of the International Conference on Human-Computer Interaction, Virtual, 26 June–1 July 2022; pp. 161–178.
42. Mu, F.; Zhang, J.; Zou, C. Gaze2Atten: Analyzing explainable gaze dynamics to monitor human attention. In Proceedings of the International Conference on Intelligent Robotics and Applications, Xi'an, China, 31 July–2 August 2024; pp. 238–256.
43. Zhou, Y.; Booth, S.; Ribeiro, M.T.; Shah, J. Do feature attribution methods correctly attribute features? *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 9623–9633. [CrossRef]
44. Appel, T.; Gerjets, P.; Hoffmann, S.; Moeller, K.; Ninaus, M.; Scharinger, C.; Sevcenko, N.; Wortha, F.; Kasneci, E. Cross-task and cross-participant classification of cognitive load in an emergency simulation game. *IEEE Trans. Affect. Comput.* **2023**, *14*, 1558–1571. [CrossRef]
45. Wibirama, S.; Alfarozi, S.A.I.; Suhari, A.R.; Fristiana, A.H.; Nurlatifa, H.; Santosa, P.I. Classification of cognitive load using deep learning based on eye movement indices. *IEEE Access* **2025**, *13*, 167324–167343. [CrossRef]
46. Krejtz, K.; Duchowski, A.T.; Niedzielska, A.; Biele, C.; Krejtz, I. Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PLoS ONE* **2018**, *13*, e0203629. [CrossRef] [PubMed]
47. Chen, W.; Sawaragi, T.; Hiraoka, T. Comparing eye-tracking metrics of mental workload caused by NDRTs in semi-autonomous driving. *Transp. Res. Part F Traffic Psychol. Behav.* **2022**, *89*, 109–128. [CrossRef]
48. Sevcenko, N.; Appel, T.; Ninaus, M.; Moeller, K.; Gerjets, P. Theory-based approach for assessing cognitive load during time-critical resource-managing human–computer interactions: An eye-tracking study. *J. Multimodal User Interfaces* **2023**, *17*, 1–19. [CrossRef]
49. Tattersall, A.J.; Foord, P.S. An experimental evaluation of instantaneous self-assessment as a measure of workload. *Ergonomics* **1996**, *39*, 740–748. [CrossRef] [PubMed]
50. EyeTrackLib. 2025. Available online: https://github.com/Space0code/EyeTrackLib (accessed on 10 April 2026).
51. Salvucci, D.D.; Goldberg, J.H. Identifying fixations and saccades in eye-tracking protocols. In Proceedings of the 2000 Symposium on Eye Tracking Research & Applications, Palm Beach Gardens, FL, USA, 6–8 November 2000; pp. 71–78.
52. Gjoreski, M.; Mahesh, B.; Kolenik, T.; Uwe-Garbas, J.; Seuss, D.; Gjoreski, H.; Luštrek, M.; Gams, M.; Pejović, V. Cognitive load monitoring with wearables–lessons learned from a machine learning challenge. *IEEE Access* **2021**, *9*, 103325–103336. [CrossRef]
53. Rapp, D.N. The value of attention aware systems in educational settings. *Comput. Hum. Behav.* **2006**, *22*, 603–614. [CrossRef]
54. Van Gog, T.; Kester, L.; Nievelstein, F.; Giesbers, B.; Paas, F. Uncovering cognitive processes: Different techniques that can contribute to cognitive load research and instruction. *Comput. Hum. Behav.* **2009**, *25*, 325–331. [CrossRef]

55. Sim, G.; Bond, R. Eye tracking in child computer interaction: Challenges and opportunities. *Int. J. Child-Comput. Interact.* **2021**, *30*, 100345. [CrossRef]
56. Devnath, L.; Janzen, I.; Lam, S.; Yuan, R.; MacAulay, C. Predicting future lung cancer risk in low-dose CT screening patients with AI tools. In *Medical Imaging 2025: Computer-Aided Diagnosis*; Astley, S.M., Wismüller, A., Eds.; SPIE: Bellingham, WA, USA, 2025; Volume 13407, p. 134072L. [CrossRef]