Eye-Tracking Explains Cognitive Test Performance in Schizophrenia

Mila Marinković
Jure Žabkar
mila.marinkovic@fri.uni-lj.si
jure.zabkar@fri.uni-lj.si
University of Ljubljana,Faculty of Computer and Information Science, Ljubljana, Slovenia

Abstract

Schizophrenia is associated with cognitive impairments that are difficult to assess with traditional neuropsychological tests, which are often lengthy and burdensome. Eye-tracking (ET) provides objective, minimally invasive measures of visual attention and cognitive processing and may complement shorter assessments. This study investigated whether ET features recorded during three computerized tasks could distinguish patients with schizophrenia from healthy controls. Using the Explainable Boosting Machine (EBM), we achieved an accuracy of 0.86, and balanced sensitivity and specificity, with an area under the curve exceeding 0.9. Features related to fixation patterns, saccadic dynamics, and temporal engagement emerged as the most informative. These findings indicate that ET features collected during brief cognitive tasks can provide clinically relevant markers of schizophrenia. Incorporating ET into short test batteries may reduce patient burden while enhancing diagnostic value, supporting the development of scalable and practical screening tools.

Keywords

schizophrenia, eye-tracking, cognitive tasks, machine learning

1 Introduction

Schizophrenia is a severe and chronic neuropsychiatric disorder that affects about 1% of the population worldwide and is characterized by disturbances in thought, perception, and behavior [1]. In addition to positive and negative symptoms, patients experience pronounced cognitive impairments, including deficits in attention, working memory, and executive functioning, which substantially affect everyday life outcomes [2, 3]. Cognitive assessment is therefore central for both diagnosis and monitoring of schizophrenia. However, traditional neuropsychological testing is lengthy, cognitively demanding, and often exhausting for patients, limiting its feasibility in clinical practice. Shorter test batteries reduce the burden but often fail to provide sufficiently informative data for reliable diagnosis

Eye-tracking (ET) offers a promising avenue for addressing this challenge. ET provides objective, real-time measures of visual attention, oculomotor control, and information processing strategies [4]. Numerous studies have shown that patients with schizophrenia exhibit abnormalities in smooth pursuit eye movements, antisaccades, and fixation stability [5, 6, 7]. These alterations are considered potential endophenotypes of the disorder, as they are also observed in first-degree relatives who do not have schizophrenia [6, 7]. More recent work has extended ET beyond basic oculomotor paradigms by embedding it in cognitive tasks. For example, Okazaki et al. [8] combined ET metrics with digit-symbol substitution tests and showed improved discrimination between patients and controls. Yang et al. [9] reported that abnormal gaze patterns during reading tasks—such as longer fixation durations and increased saccade counts—enabled high diagnostic accuracy when analyzed with machine learning models. Similarly, Morita et al. [10] demonstrated the feasibility of portable tablet-based ET combined with cognitive assessments for schizophrenia screening. Collectively, these studies highlight that combining ET with cognitive testing enriches diagnostic value and provides insights into the cognitive mechanisms underlying gaze abnormalities.

Building on this prior work, the present study investigates whether ET features recorded during a small set of computerized cognitive tasks can serve as reliable markers of schizophrenia. Participants completed three tasks (digit span, picture naming, and n-back), each divided into phases of instruction reading, video demonstration, and test execution. From these tasks, we extracted 117 ET features, including fixation measures, saccadic dynamics, gaze entropy, and recording duration. We then applied machine learning methods to evaluate the discriminative power of these features. By focusing on only three short tasks, our aim is to test whether ET provides sufficient additional information to overcome the limitations of brief cognitive testing, ultimately supporting the development of less burdensome but more informative screening approaches.

2 Methods

2.1 Participants

The study involves 126 individuals, including 58 patients diagnosed with schizophrenia (SP) and 68 healthy controls (HC). All participants were adults, aged 18 years or older. Patients were recruited and tested at the University Psychiatric Hospital Ljubljana. The control group was matched to the patient group on age and gender.

Eligibility criteria required fluency in Slovenian and excluded individuals with intellectual disability, organic brain disorders, or a history of substance abuse. Additional exclusion criteria for the HC group included any past or current psychiatric disorder. At the time of assessment, all SP participants were receiving stable doses of antipsychotic medication.

Demographic characteristics of the two groups are presented in Table 1 and were analyzed to ensure that the groups were comparable in terms of age and gender. While educational attainment differed between groups, further analyses confirmed that within each education level there were no significant differences between SP and HC participants, indicating that education was unlikely to confound the comparisons.

Information Society 2025, Ljubljana, Slovenia

https://doi.org/10.70314/is.2025.skui.9583

Table 1: Demographic characteristics of participants.

Measure	SP	HC
Counts		
Total participants	58	68
Male sex	29	35
Female sex	29	33
Continuous		
Age (mean years)	46.1	46.7
Categorical		
Most common education level	Primary school	High school

HC: Healthy Controls; SP: Patients with Schizophrenia

The study was approved by the Medical Ethics Committee of the Republic of Slovenia (approval number: 0120-51/2024-2711-4). All participants received a detailed explanation of the study procedures and provided written informed consent prior to participation.

2.2 Testing Procedure

Eye-tracking data were collected using a Tobii Pro Spectrum [11] eye tracker integrated into a 24-inch monitor with a resolution of 1920×1080 pixels. Recordings were made at 1200 Hz in the "human" tracking mode, with a stimulus presentation latency of approximately 10 ms. The display frame rate was 30 FPS. Participants sat ~55cm from the monitor, in a upright position with seating adjusted for comfort and optimal tracking.

Before each task, participants were seated comfortably, and the Tobii Pro Lab [12] interface provided a live preview (see Fig. 1) to verify that both eyes were detected and that the viewing distance was within the recommended range (displayed as a green zone, typically around 55 cm). Once this was confirmed, a standard five-point calibration was performed, during which participants followed a moving dot across the screen. Calibration served both to align gaze tracking and to ensure that the participant had not moved their head between tasks. If the system indicated suboptimal accuracy, the calibration was repeated.

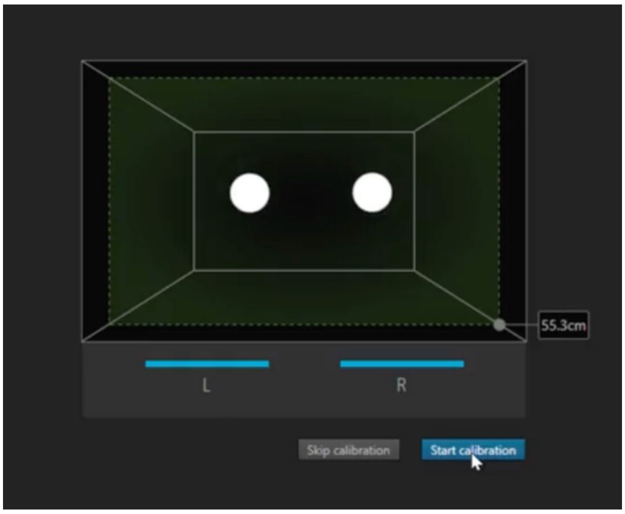

Figure 1: Calibration interface in Tobii Pro Lab. The preview window ensures both eyes are detected and the participant is seated at an appropriate distance (green zone, approximately 55 cm) before calibration and testing begin.

Participants completed three computerized cognitive tasks in a fixed order: digit span (DS), n-back (NB), and picture naming (PS). A short break was provided between tasks, with the duration determined by each participant. All tasks were presented within the Tobii Pro Lab application, which also stored the raw data. After recording, the data were exported and processed using a custom Python program for feature extraction and analysis.

Each task followed the same three-phase structure:

(1) **Reading instructions.** Written instructions were displayed on the screen. Participants could read them at their own pace and advanced to the next phase with a mouse click.
(2) **Video example.** A short instructional video was presented once, demonstrating the task procedure.
(3) **Test execution.** The participant began the task when ready. Task duration depended on individual performance.

The procedure was identical for all participants, ensuring standardization across groups. Only the test execution phase varied in length, as it was determined by each participant's performance. Group-level descriptive statistics of fixation durations for all tasks and phases are reported in the Results section (Table 3).

2.3 Feature Extraction

We extracted a total of 117 ET features from three computerized cognitive tasks. As described in Section 2.2, each task was divided into three phases: instruction reading (BN), video demonstration (GN), and test execution (T).

Each participant contributed a single data point to the ML analysis. For every task (DS, PS, NB) and every phase (BN, GN, T), we computed the 13 eye-tracking features listed in Table 2. Each feature was calculated over the entire duration of the given phase (e.g., the number of fixations refers to the total count during that phase, while mean fixation duration refers to the average across all fixations in that phase). These were then concatenated across all tasks and phases, yielding 117 features per participant. Thus, the unit of analysis was the participant, not individual trials or task phases.

2.4 Data Analysis

We trained and evaluated several machine learning models using these features. We applied stratified 10-fold cross-validation at the subject level to ensure that all features from a given participant were assigned exclusively to either the training or test set, thereby preventing data leakage across folds. In each iteration, the model was trained on nine folds and tested on the remaining one, and the reported metrics represent averages across all folds. Performance was assessed using accuracy, sensitivity, specificity, and the area under the receiver operating characteristic curve (AUC). The final results were reported as the average across all folds.

We evaluated a diverse set of ML models (logistic regression, random forest, gradient boosting, extreme gradient boosting, and the explainable boosting machine) to cover both linear and non-linear approaches with varying levels of interpretability. EBM was selected as the primary model because it consistently achieved the highest overall performance while providing inherently interpretable feature importance, which is particularly valuable in clinical contexts. We did not pursue deep neural networks in this study, as the dataset size (126 participants) is relatively small and does not provide sufficient power to train high-capacity models without overfitting.

Table 2: Eye-tracking features extracted from each task and phase.

Feature	Description
num_fixations	Total number of fixations during the interval.
avg_fixation_duration	Mean duration of fixations (ms), indicating fixation stability.
std_fixation_duration	Standard deviation of fixation duration, reflecting variability in fixation times.
num_saccades	Total number of saccadic eye movements.
avg_saccade_distance	Mean distance of saccades, reflecting amplitude of eye shifts.
avg_saccade_velocity	Mean velocity of saccades, indicating how quickly gaze shifts occurred.
avg_saccade_angle	Average angular change of saccades, reflecting directional scanning patterns.
gaze_entropy	Entropy of gaze distribution, quantifying dispersion vs. concentration of gaze.
recording_duration_ms	Total duration of recording for the phase (ms).
unique_squares	Number of unique spatial areas (AOIs) visited during the interval.
num_changes	Number of transitions between distinct gaze areas.
missing_left_percent	Percentage of missing data from the left eye.
missing_right_percent	Percentage of missing data from the right eye.

Note: All features are computed as aggregates over the entire task phase for each participant.

3 Results

To characterize task engagement and potential variability between groups, we compared fixation durations across all tasks and phases (Table 3). SP showed longer fixations than HC, especially during instruction reading and video phases, with smaller but consistent differences during execution. This indicates altered attention even outside active task solving.

Table 3: Mean fixation duration in ms per task and phase.

Task	Phase	HC (Mean ± SD)	SP (Mean ± SD)
Numbers	Reading instructions	239.64 ± 47.79	283.97 ± 45.33
	Watching video	352.14 ± 81.56	400.10 ± 89.51
	Test execution	390.66 ± 83.92	407.60 ± 98.53
Pictures	Reading instructions	228.44 ± 52.49	267.78 ± 60.79
	Watching video	302.40 ± 69.06	368.93 ± 81.42
	Test execution	301.97 ± 49.91	319.36 ± 58.07
Square	Reading instructions	229.36 ± 45.41	286.70 ± 63.42
	Watching video	309.41 ± 89.45	352.08 ± 79.37
	Test execution	394.91 ± 115.50	406.24 ± 99.36

SD: Standard deviation; HC: Healthy controls; SP: Schizophrenia patients

The ML models were trained on 117 extracted eye-tracking features and achieved strong performance in distinguishing SP from HC. The key cross-validation performance metrics are summarized in Table 4.

Table 4: Cross-validation performance metrics for different models. The Explainable Boosting Machine (EBM) achieved the best overall performance across all metrics.

Model	Accuracy	Sensitivity	Specificity	AUC
EBM	**0.86**	**0.84**	**0.86**	**0.93**
LR	0.85	0.77	0.91	0.92
GB	0.78	0.70	0.84	0.83
RF	0.83	0.84	0.82	0.91
xGB	0.81	0.77	0.85	0.90

EBM: Explainable Boosting Machine; LR: Logistic Regression; GB: Gradient Boosting; RF: Random Forest; xGB: Extreme Gradient Boosting

Among the tested models, the EBM achieved the highest overall performance and was therefore selected for detailed analysis. Fig 2 presents the receiver operating characteristic (ROC) curve, which confirms the model's strong discriminative ability.

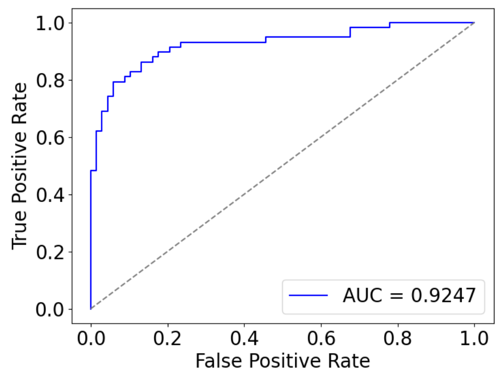

Figure 2: ROC curve for the EBM model. The mean AUC across folds was 0.92, confirming strong classification performance.

We analyzed the feature importance scores provided by EBM, focusing on the ten most informative features (Fig 3). These features were predominantly derived from the test execution phases and included measures such as recording duration, number of fixations, mean fixation duration, and saccadic counts.

4 Discussion

The present study demonstrates that eye-tracking (ET) features obtained during brief computerized cognitive tasks can effectively discriminate between individuals with schizophrenia and healthy controls. Using 117 features, the Explainable Boosting Machine (EBM) achieved strong classification performance, with accuracy, sensitivity, and specificity values around 0.85 and an AUC of 0.92. These results provide further evidence that ET-based measures capture clinically relevant differences in cognitive processing and attentional control in schizophrenia.

Our findings are consistent with previous work showing that patients with schizophrenia exhibit abnormalities in fixation behavior, saccadic dynamics, and gaze distribution during

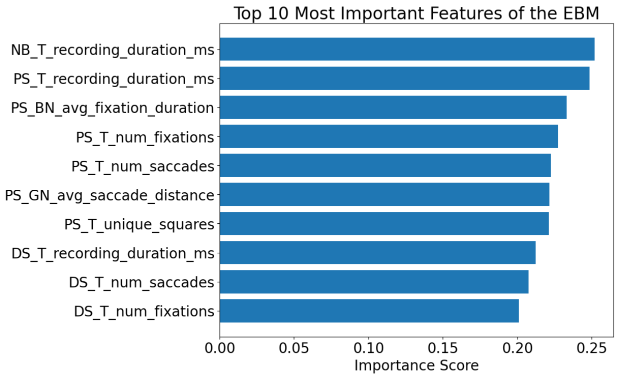

Figure 3: Top 10 most important features identified by the EBM model. The prefixes indicate the task and phase: DS = digit span, PS = picture naming, NB = n-back; BN = reading instructions, GN = watching video, T = test solving. For example, `PS_T_num_fixations` refers to the number of fixations during the test phase of the picture naming task.

both simple oculomotor paradigms and more complex cognitive tasks [5, 6, 7, 8, 9, 10]. Importantly, by embedding ET into a small set of standardized cognitive tasks, we demonstrate that group differences emerge not only during active problem solving but also in more passive phases such as reading instructions or watching a video example. This suggests that ET provides valuable information across the continuum of cognitive engagement, extending beyond traditional task performance metrics.

While prior studies have applied machine learning to ET data in schizophrenia, they have typically relied on single paradigms or isolated task conditions. The novelty of the present work lies in combining a multi-task, multi-phase design with interpretable ML within a short, clinically feasible test battery. This approach captures a broader range of cognitive and attentional processes while linking model performance to specific, clinically meaningful features.

Interpretability showed that temporal engagement, fixation stability, and saccadic activity best differentiated groups. Longer recording durations may reflect slower processing, while altered fixations and saccades align with prior reports of impaired attentional control. These findings suggest that eye-tracking captures both temporal and oculomotor aspects of task performance, supporting its potential as a clinically meaningful biomarker.

From a clinical perspective, these results are encouraging. Traditional neuropsychological assessments are lengthy and cognitively demanding, which can be exhausting for patients and limit their applicability. Our study shows that by integrating ET measures into just three relatively brief cognitive tasks, it is possible to achieve a high level of diagnostic accuracy. This approach may therefore support the development of shorter, less burdensome, and more objective screening protocols that could complement existing clinical evaluations.

Limitations and Future Work

Several limitations should be noted. First, although our sample size of 126 participants is comparable to similar studies, larger and more diverse cohorts are needed to confirm the generalizability of the results. Second, all patients were on stable antipsychotic medication, which may have influenced oculomotor behavior. Third, while we employed subject-level cross-validation to prevent data leakage, robustness checks such as leave-one-subject-out or leave-one-task-out validation could further strengthen reliability. Fourth, our analysis focused on static ET features; dynamic sequence-based or deep learning models could capture additional temporal information in gaze patterns. Finally, we only tested three tasks; future research should explore whether expanding or tailoring the task battery improves performance while still keeping the protocol brief. Replication with independent cohorts will be essential to establish clinical utility.

Conclusion

In conclusion, this study provides strong evidence that eye-tracking features embedded within short cognitive tasks can serve as robust markers of schizophrenia. Machine learning models trained on these features achieved high discriminative accuracy, with interpretable patterns that align with known attentional and cognitive impairments in the disorder. By reducing patient burden while maintaining informativeness, this approach holds promise for the development of accessible, scalable, and clinically relevant screening tools for schizophrenia.

5 Acknowledgments

This research was funded by the Slovenian Research Agency (core funding No. P2-0209). Additional support for Mila Marinković was provided through the ARIS project AI4Science (GC-0001). We thank the University Psychiatric Hospital Ljubljana for participant recruitment, asist. dr. Polona Rus Prelog, dr. med., spec. psih., and Martina Zakšek for their support, and the broader research team, clinicians, and participants for their contributions. The authors declare that they have no conflict of interest.

References

[1] S. R. Marder and T. D. Cannon, "Schizophrenia," *The New England Journal of Medicine*, vol. 381, no. 18, p. 1753–1761, 2019.
[2] W. Hinzen and J. Rosselló, "The linguistics of schizophrenia: thought disturbance as language pathology across positive symptoms," *Frontiers in Psychology*, vol. 6, 2015.
[3] L. Colle, R. Angeleri, M. Vallana, K. Sacco, B. Bara, and F. Bosco, "Understanding the communicative impairments in schizophrenia: A preliminary study," *Journal of Communication Disorders*, vol. 46, no. 3, pp. 294–308, 2013.
[4] A. Wolf, K. Ueda, and Y. Hirano, "Recent updates of eye movement abnormalities in patients with schizophrenia: A scoping review," *Psychiatry and Clinical Neurosciences*, vol. 75, pp. 104–118, 2021.
[5] P. S. Holzman, L. R. Proctor, and D. W. Hughes, "Eye-tracking patterns in schizophrenia," *Science*, vol. 181, no. 4095, pp. 179–181, 1973.
[6] L. Deborah, H. Philip, M. Steven, and M. Nancy, "Eye tracking and schizophrenia: A selective review," *Schizophrenia Bulletin*, vol. 20, no. 1, pp. 47–62, 1994.
[7] U. Ettinger, "Smooth pursuit and antisaccade eye movements as endophenotypes in schizophrenia spectrum research," PhD Thesis, Department of Psychology, Goldsmiths College, University of London, 2002.
[8] K. Okazaki, K. Miura, J. Matsumoto, N. Hasegawa, M. Fujimoto, H. Yamamori, Y. Yasuda, M. Makinodan, and R. Hashimoto, "Discrimination in the clinical diagnosis between patients with schizophrenia and healthy controls using eye movement and cognitive functions," *Psychiatry and Clinical Neurosciences*, vol. 77, pp. 393–400, 2023.
[9] H. Yang, L. He, W. Li, Q. Zheng, Y. Li, X. Zheng, and J. Zhang, "An automatic detection method for schizophrenia based on abnormal eye movements in reading tasks," *Expert Systems With Applications*, vol. 238, p. 121850, 2024.
[10] K. Morita, K. Miura, A. Toyomaki, M. Makinodan, K. Ohi, N. Hashimoto, Y. Yasuda, T. Mitsudo, F. Higuchi, S. Numata, A. Yamada, Y. Aoki, H. Honda, R. Mizui, M. Honda, D. Fujikane, J. Matsumoto, N. Hasegawa, S. Ito, H. Akiyama, T. Onitsuka, Y. Satomura, K. Kasai, and R. Hashimoto, "Tablet-based cognitive and eye movement measures as accessible tools for schizophrenia assessment: Multisite usability study," *JMIR Mental Health*, vol. 11, p. e56668, 2024.
[11] M. Nyström, D. Niehorster, R. Andersson, and I. Hooge, "The tobii pro spectrum: A useful tool for studying microsaccades?" *Behavior Research Methods*, vol. 53, 07 2020.
[12] Tobii AB, "Tobii pro lab," Computer software, Danderyd, Stockholm, 2024. [Online]. Available: http://www.tobii.com/