Journal of Science and Technology – Smart Systems and Devices
Volume x, Issue x, Month Year, ppp-ppp

Schizophrenia Detection Using Eye Movement Data: A Machine Learning–Based Comparative Study

Minh-Phuong Vu, Mai Huong Nguyen, Tra-My Pham Nguyen, Thanh-Hang Vu, Thanh-Hai Tran*

School of Electrical and Electronic Engineering,
Hanoi University of Science and Technology, Ha Noi, Vietnam
**Corresponding author email: hai.tranthithanh1@hust.edu.vn*

Abstract
Schizophrenia is a severe mental disorder associated with impairments in perception, cognition, and visual exploration behavior. Objective and non-invasive screening approaches based on eye movement analysis have emerged as promising tools to complement traditional clinical assessment. This study investigates the feasibility of detecting Schizophrenia using eye tracking data through a comparative machine learning framework. Two complementary feature-engineering strategies are developed and evaluated at the subject level. The first strategy focuses on a compact set of clinically interpretable biomarkers derived from fixation dynamics, saccadic characteristics, pupil variability, and spatial distribution complexity. The second strategy constructs a high-dimensional behavioral representation by modeling gaze transitions, scanpath topology, and oculomotor statistics, followed by a structured feature selection procedure to improve generalization. Both pipelines are assessed using stratified cross-validation on a training cohort and final evaluation on an independent held-out test set. The experimental results on a benchmark dataset demonstrate that eye movement patterns contain discriminative information sufficient for automated Schizophrenia recognition. Our two approaches achieved the highest accuracy of 85.42% and 81.25% respectively. While the high-dimensional representation benefits from non-linear modeling after feature selection, the low-dimensional biomarker approach achieves competitive performance with reduced complexity and stronger interpretability. These findings highlight the trade-off between representational richness and clinical transparency and suggest that feature-centric machine learning models can provide practical and scalable support for data-driven mental health screening. The study contributes methodological insights that may facilitate translational development of interpretable, eye movement–based decision support systems in psychiatric assessment.

Keywords: Eye tracking, graph representation, machine learning, schizophrenia detection.

1. Introduction

Schizophrenia is a severe psychiatric disorder characterized by disturbances in perception, cognition, and behavior. In addition to its clinical symptoms, patients with Schizophrenia consistently exhibit abnormalities in eye movement control, including altered fixation patterns, atypical saccadic transitions, and irregular pupil dynamics. These oculomotor disturbances provide quantifiable behavioral markers that reflect impairments in cognitive control, visual attention, and sensorimotor integration. The development of objective screening tools for Schizophrenia remains an important research challenge. Current diagnosis relies primarily on clinical interviews and symptom-based assessments, which may involve subjective interpretation and variability across evaluators [1, 2].

Because eye tracking provides precise, non-invasive, and high-temporal-resolution recordings of gaze behavior, it has become a valuable modality for quantitative investigation. [3]. Machine learning enables supervised classification at the subject level by learning discriminative patterns from high-dimensional behavioral measurements [4–6]. Eye movement data are particularly suitable for this purpose due to their structured spatial and temporal characteristics, which can be transformed into feature representations for subject-level classification.

Existing computational approaches for Schizophrenia detection from eye tracking generally fall into two methodological categories. The first emphasizes compact, clinically interpretable biomarkers. In this paradigm, statistics of fixation duration, saccade amplitude, scanpath dispersion, or pupil variability are aggregated per subject and used as inputs to classical classifiers [7–9]. Such models are computationally efficient and facilitate physiological interpretation, making them attractive for moderate-sized clinical datasets. The second category focuses on richer behavioral representations, modeling scanpaths as structured spatiotemporal sequences. Techniques such as region-of-interest discretization, transition modeling, graph-based descriptors, or other high-dimensional encodings aim to capture detailed gaze dynamics [10, 11]. While these representations may encode more comprehensive behavioral information, they increase dimensionality and require careful regularization to ensure generalization.

Despite recent progress in both high-dimensional behavioral modeling and compact biomarker-based approaches, systematic comparisons under a unified

p-ISSN 3093-3285
e-ISSN 3093-3315
https://doi.org/10.51316/jst.xxx.ssad.xxxx.xx.x.x
Received: Mar 1, 2026; Revised: Apr 11, 2026;
Accepted: Jun 12, 2026; Online: Jul 22, 2026

evaluation protocol are still scarce. Consequently, it remains unclear whether increasing representational complexity consistently improves generalization, or whether carefully designed low-dimensional biomarkers can achieve comparable performance while offering better interpretability and robustness. Understanding this trade-off is crucial for the practical deployment of clinical decision-support systems. Another major challenge lies in the limited availability of eye-tracking datasets for Schizophrenia detection. Recently, Song et al. introduced a benchmark dataset for evaluating Schizophrenia detection algorithms [12], helping to standardize experimental comparisons.

In this study, we conduct a comparative machine learning analysis for Schizophrenia detection using eye movement data from the EMS benchmark dataset [12]. Two complementary pipelines are implemented at the subject level. The first pipeline, named SFE (**S**tatistical **F**eature **E**ngineering), employs a compact set of clinically interpretable biomarkers derived from fixation dynamics, saccadic characteristics, pupil variability, and spatial distribution complexity. The second, named STA (Graph based **S**patial-**T**emporal **A**nalysis), constructs a high-dimensional behavioral signature that integrates scanpath transition patterns, graph-theoretic descriptors, and statistical oculomotor features, followed by a structured feature selection procedure to obtain a stable subset of informative variables. Both approaches are evaluated under comparable subject-level validation settings, including stratified cross-validation and held-out testing.

By systematically comparing these two feature paradigms, this work aims to quantify the trade-off between representational richness and interpretability in eye movement–based Schizophrenia recognition and to provide methodological guidance for the development of reliable and scalable screening models. Our main contribution is two-fold: i) we propose two approaches exploring and representing eye movement signals; ii) we conduct a comparative study of the performance of the two proposed approaches on a benchmark dataset. Experimental results show promising results with both studied approaches.

2. Related Works

Research on Schizophrenia (SZ) diagnosis using eye-tracking technology has evolved significantly, generally categorized into three main streams: statistical feature-based methods, deep learning approaches, and sequence/graph analysis.

2.1. Statistical Features

Early studies primarily focused on extracting explicit oculomotor features. Common metrics include fixation duration, saccade amplitude, and spatial dispersion. Pioneers like Holzman *et al.* (1974) established the fundamental link between eye-tracking dysfunctions and Schizophrenia [7]. Later, researchers such as Benson *et al.* (2012) [8] and Tseng *et al.* (2013) [9] utilized these statistical features combined with classical machine learning classifiers like Support Vector Machines (SVM) and Random Forest. While these methods provide clinical interpretability, they often rely on basic manual feature design and may fail to fully capture the complex temporal dynamics and physiological nuances of the entire scanpath.

2.2. Deep Learning Models

Recent advancements have shifted towards Deep Learning (DL) to automatically learn latent representations directly from raw gaze data. Notably, Song *et al.* (2024) released the EMS dataset and proposed the Mean-Shift Network (MSNet), which integrates a mean-shift algorithm with Convolutional Neural Networks (CNNs) to extract cluster centers as subject features [12]. Other works have utilized Long Short-Term Memory (LSTM) networks or 1D-CNNs to process gaze data as time-series sequences. Pushing the boundaries for SZ classification, Dang et al. (2025) proposed RAnet-ET, a ResNet-based attention network that transforms gaze trajectories into visual heatmaps, achieving an exceptional accuracy of 96.37% [13].

Beyond Schizophrenia, the efficacy of DL in eye-tracking has been extensively validated across other psychiatric domains. For instance, recent models have successfully assessed executive function decline using LSTM-Attention [14], and achieved high performance in detecting depression and suicidal ideation using multi-input CNN frameworks [15], sentence-reading gaze patterns [16], and 2D Inception architectures [17].

However, despite their high classification performance, these "black-box" models often suffer from a severe lack of interpretability. Furthermore, they typically require massive datasets to generalize effectively, presenting a significant challenge in clinical settings where the number of patient samples is often limited.

2.3. Graph-based Sequence Analysis

A third, emerging category treats scanpaths as topological graphs or linguistic sequences. By modeling Regions of Interest (ROIs) as nodes and saccades as edges, or by encoding gaze paths into N-gram sequences (Coutrot et al., 2018), researchers can capture structural and sequential transition patterns that pure statistical methods often miss. Our work builds upon and synthesizes these promising directions [11].

3. Our Proposed Framework

As mentioned in the introduction section, we explore two main approaches for data analysis and disease classification. The first approach (SFE: **S**tatistical **F**eature **E**ngineering) aims to produce a compact feature set from the raw eye movement data, while the

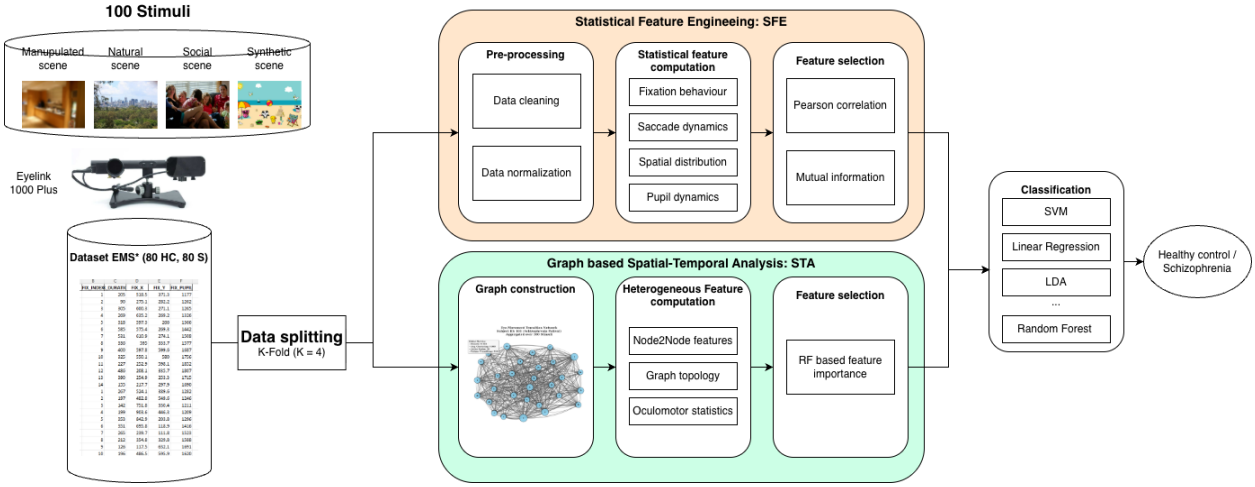

Fig. 1. General framework for Schizophrenia detection using eye movement

second (STA: Graph based **S**patial-**T**emporal **A**nalysis) represents the raw data in graph form and extracts features from this representation.

3.1. Pre-processing Steps

Eye-tracking data collected at 1000 Hz is inherently noisy due to blinks, head movements, loss of tracking, and fluctuations in pupil diameter. During the EMS experiment, stimuli were displayed on a 1024×768 screen; therefore, any ***fixation coordinates*** falling outside this range are physically impossible and must be discarded. After this step, a total of 4,058 invalid spatial points were removed, resulting in significantly cleaner fixation distributions and eliminating distorted heatmap artifacts.

Fixation duration reflects the amount of time the gaze remains stable at a location. A duration of 0 ms cannot represent a valid fixation and usually indicates a recording error or tracker dropout. Only one such case appeared in the dataset and was removed. No high-duration thresholding was applied because longer fixations may reflect meaningful cognitive patterns, especially in Schizophrenia subjects.

Pupil diameter is sensitive to illumination changes and blinking. Although the dataset contained no zero-pupil readings, several extreme values were present. To remove such artifacts, a Z-score threshold of 3 was applied. This is a widely used criterion in physiological signal cleaning to eliminate highly improbable values while preserving genuine variability. A total of 1,356 pupil outliers were removed.

3.2. SFE: Statistical Feature Engineering

This study focuses on transforming raw gaze recordings into a structured set of behavioural descriptors suitable for downstream modelling. The feature engineering process is designed to capture the major components of visual attention, including fixation behaviour, saccadic exploration, spatial distribution of gaze, and pupil dynamics. All features are aggregated at the subject level to obtain a compact yet informative representation of individual viewing patterns across the experiment.

3.2.1. Feature computation

For each subject, let N_f denote the total number of valid fixations after preprocessing. Each fixation i is characterized by spatial coordinates (x_i, y_i), fixation duration d_i, and pupil diameter p_i. All features are computed by aggregating fixation-level measurements across the entire set of 100 stimuli.

Fixation behaviour Fixation statistics describe temporal attention stability and viewing persistence [18]. Let μ_d denote the mean fixation duration:

$$\mu_d = \frac{1}{N_f} \sum_{i=1}^{N_f} d_i \tag{1}$$

The variability of fixation duration is computed as:

$$\sigma_d = \sqrt{\frac{1}{N_f - 1} \sum_{i=1}^{N_f} (d_i - \mu_d)^2} \tag{2}$$

The coefficient of variation is defined as:

$$CV_d = \frac{\sigma_d}{\mu_d} \tag{3}$$

To quantify prolonged attention, the proportion of long fixations (threshold $\theta = 400$ ms) is calculated as:

$$P_{\text{long}} = \frac{1}{N_f} \sum_{i=1}^{N_f} \mathbb{I}(d_i > \theta) \tag{4}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. The total fixation count is simply

$$F_{\text{count}} = N_f \tag{5}$$

Saccade dynamics Saccade features characterize spatial exploration between consecutive fixations [19]. After sorting fixations temporally, the amplitude between two consecutive fixations is defined as:

$$a_j = \sqrt{(x_{j+1} - x_j)^2 + (y_{j+1} - y_j)^2}, \quad j = 1, \ldots, N_f - 1. \tag{6}$$

The mean saccade amplitude is computed as:

$$\mu_a = \frac{1}{N_f - 1} \sum_{j=1}^{N_f - 1} a_j \tag{7}$$

The number of saccades is $S_{\text{count}} = N_f - 1$, while the variability of amplitudes is:

$$\sigma_a = \sqrt{\frac{1}{N_f - 2} \sum_{j=1}^{N_f - 1} (a_j - \mu_a)^2} \tag{8}$$

Here, $N_f - 2$ appears because the variance is computed over the $N_f - 1$ saccade amplitudes derived from consecutive fixation pairs.

The maximum saccade amplitude is defined as:

$$a_{\max} = \max_{1 \le j \le N_f - 1} a_j \tag{9}$$

Spatial distribution Spatial descriptors reflect how gaze points are distributed across the stimulus area [20]. Horizontal and vertical spreads are computed as:

$$\sigma_x = \sqrt{\frac{1}{N_f - 1} \sum_{i=1}^{N_f} (x_i - \bar{x})^2}, \quad \sigma_y = \sqrt{\frac{1}{N_f - 1} \sum_{i=1}^{N_f} (y_i - \bar{y})^2} \tag{10}$$

To quantify global spatial dispersion, the stimulus area is discretized into K bins. Let p_k denote the proportion of fixations falling into bin k. The spatial entropy is defined as:

$$H = -\sum_{k=1}^{K} p_k \log p_k \tag{11}$$

Additionally, spatial compactness is computed as the average radial distance from the centroid:

$$C = \frac{1}{N_f} \sum_{i=1}^{N_f} \sqrt{(x_i - \bar{x})^2 + (y_i - \bar{y})^2} \tag{12}$$

Pupil dynamics Pupil-based features capture cognitive load and physiological fluctuations during visual processing. The variability of pupil diameter is computed as:

$$\sigma_p = \sqrt{\frac{1}{N_f - 1} \sum_{i=1}^{N_f} (p_i - \bar{p})^2} \tag{13}$$

The coefficient of variation is defined as:

$$CV_p = \frac{\sigma_p}{\bar{p}} \tag{14}$$

To measure short-term fluctuations, first-order temporal differences are calculated:

$$\Delta p_i = p_{i+1} - p_i, \quad i = 1, \ldots, N_f - 1 \tag{15}$$

The variability of these derivatives is:

$$D_p = \sqrt{\frac{1}{N_f - 2} \sum_{i=1}^{N_f - 1} (\Delta p_i - \overline{\Delta p})^2} \tag{16}$$

Similarly, the denominator $N_f - 2$ is used because the variability is estimated from the $N_f - 1$ first-order pupil differences.

In total, sixteen candidate features are computed before redundancy analysis and feature selection. These candidate features were not all retained for modelling; instead, a reduced subset was subsequently selected based on redundancy analysis and consistency across multiple ranking methods.

3.2.2. Feature selection

Correlation analysis was performed to detect redundant and collinear features within the statistical feature set. Strong positive correlations were observed within several feature groups, particularly fixation-related measures (e.g., MEAN_DURATION, STD_DURATION, CV_DURATION), saccade-related measures (e.g., SAC_MEAN_AMP, SAC_STD_AMP, SAC_MAX_AMP), and spatial descriptors (e.g., SPATIAL_STD_X, SPATIAL_STD_Y, HM_ENTROPY). These dependencies indicate that multiple features encode overlapping behavioural characteristics, motivating a systematic reduction of the feature space (Fig. 2).

To assess feature relevance, three complementary ranking methods were applied: Mutual Information (MI), Random Forest (RF) importance, and minimum Redundancy Maximum Relevance (mRMR). MI estimates statistical dependency between each feature and the class label, RF captures non-linear discriminative power through ensemble learning, and mRMR balances relevance and redundancy during selection (Fig. 3).

The resulting rankings revealed a high degree of agreement across the three methods. In particular, six features (MEAN_DURATION, PROP_LONG_FIX, FIX_COUNT, SAC_COUNT, PUPIL_STD, and PUPIL_DERIV_STD) were consistently identified as highly informative across all three methods. Three additional features (SAC_MEAN_AMP, PUPIL_CV, and HM_ENTROPY) were selected by at least two of the three ranking procedures, providing further support for their inclusion in the final subset.

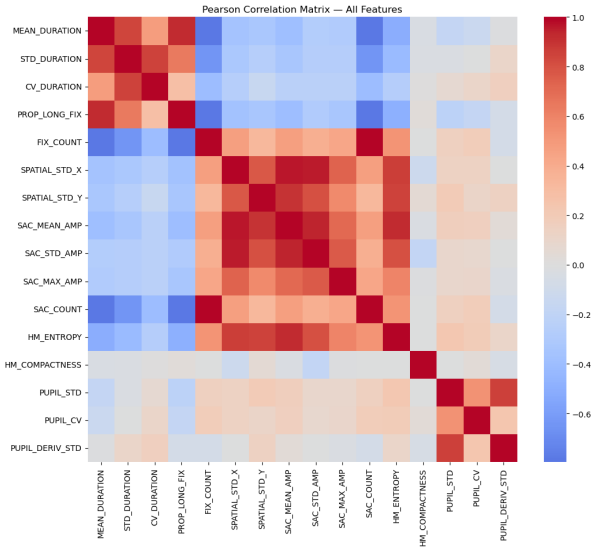

Fig. 2. Pearson correlation heatmap of the engineered SFE features

Although certain features such as SPATIAL_STD_X and SPATIAL_STD_Y exhibited non-negligible discriminative signals, they were excluded due to their strong redundancy with HM_ENTROPY, which provides a more comprehensive representation of spatial gaze distribution. Similarly, within the saccade-amplitude group, SAC_MEAN_AMP was retained as a stable representative, whereas SAC_STD_AMP and SAC_MAX_AMP were excluded due to weaker ranking consistency.

Based on the combined evidence from correlation analysis and multi-method feature ranking, a final subset of nine features was selected: MEAN_DURATION, PROP_LONG_FIX, FIX_COUNT, SAC_COUNT, SAC_MEAN_AMP, PUPIL_STD, PUPIL_CV, PUPIL_DERIV_STD, and HM_ENTROPY. This reduced feature set provides a compact yet expressive representation of oculomotor behaviour and serves as the input to all subsequent classification models. Table 1 summarizes the full set of engineered features and highlights the final selected subset.

3.3. STA: Graph based -Spatial Temporal Analysis

In the second approach, we extract features from eye movement images. These images have resolution similar to the original ones (e.g. 1024 × 768). The pixel value represents the number of times the gaze falls on that spatial location, forming a spatial attention map that summarizes the viewer's visual focus.

3.3.1. Building graph of eye fixation

Let the eye-movement scanpath of a subject across M stimuli be represented as:

$$\mathscr{S} = \{(P_1^1,\ldots,P_{k_1}^1),(P_1^2,\ldots,P_{k_2}^2),\ldots,(P_1^M,\ldots,P_{k_M}^M)\}, \quad (17)$$

where P_i^j denotes the i-th fixation point on the j-th stimulus and k_j is the number of fixation points for stimulus j and $j = 1,\ldots,M$. Let R denote the average number of fixation points per stimulus. The parameter $R = 30$ was chosen for this study. This choice is not arbitrary but is informed by the distribution of fixation points in our dataset. Specifically, the number of fixation points for most subjects falls within the range of 27 to 39, with a strong concentration around 30. Therefore, setting $R = 30$ allows the constructed graph to effectively capture the fixation structure for the majority of subjects without excessive truncation or padding. To further support this choice, we conducted additional empirical validation across multiple machine learning models by varying R from 20 to 35 (i.e., 20, 25, 30, and 35). The results consistently demonstrate that most models achieve their highest accuracy at $R = 30$.

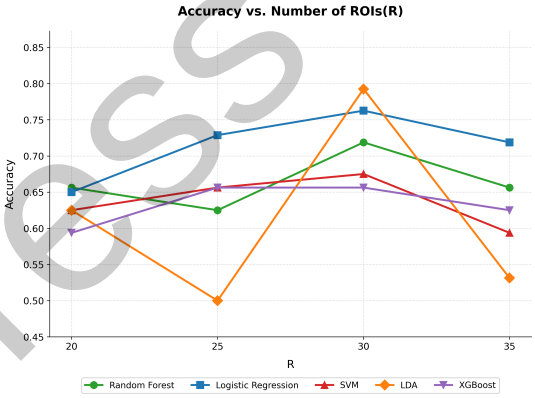

Fig. 4. Empirical evaluation of five machine learning models across varying numbers of ROIs (R) for parameter selection

We construct a directed graph $G = (V,E)$ with $|V| = R$ nodes. The node indices may correspond to fixation indices or be reassigned during graph construction. For each fixation point P_i^j, we compute its Euclidean distance to the existing nodes:

$$d(P_i^j, v) = \sqrt{(x_i^j - x_v)^2 + (y_i^j - y_v)^2}, \quad v \in V \quad (18)$$

If the minimum distance satisfies

$$\min_{v \in V} d(P_i^j, v) > \theta_d \quad (19)$$

a new node is created at location P_i^j (until the maximum number R is reached). Otherwise, P_i^j is merged into the nearest node v^*, and the visit frequency of that node is increased by one.

To encode temporal dynamics, two nodes v_p and v_q are connected by a directed edge $e_{pq} \in E$ if the gaze transitions from fixation p to fixation q in chronological order. The edge weight is defined as the Euclidean distance between the two nodes:

$$w_{pq} = d(v_p, v_q) \quad (20)$$

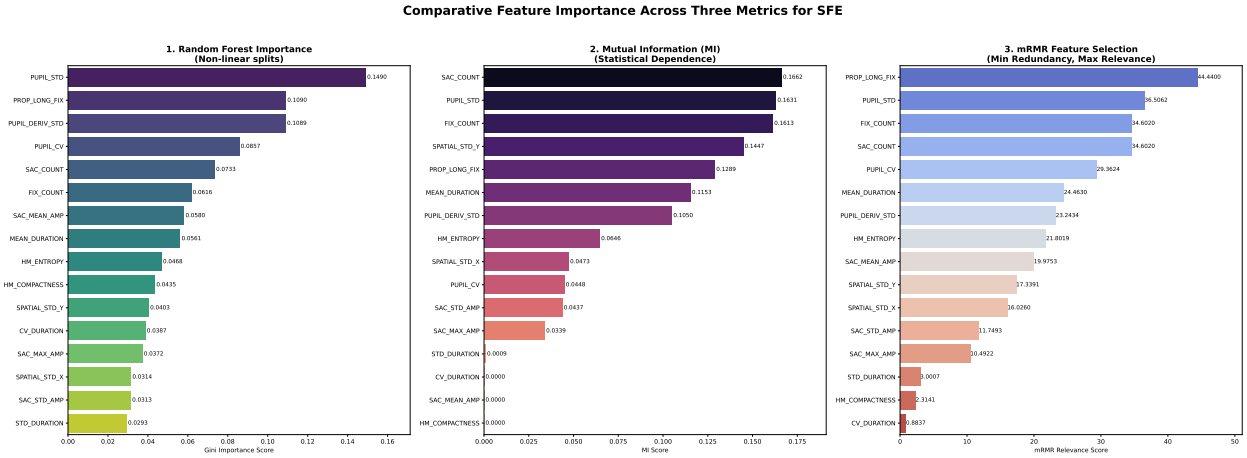

Fig. 3. Comparative ranking of SFE features across three metrics

Table 1. The full set of sixteen engineered SFE features, with the final 9 selected features shown in bold

No.	Feature	Description	Formula
1	**MEAN_DURATION**	Mean fixation duration.	$\mu_d = \frac{1}{N_f}\sum_{i=1}^{N_f} d_i$
2	STD_DURATION	Standard deviation of fixation duration.	$\sigma_d = \sqrt{\frac{1}{N_f-1}\sum_{i=1}^{N_f}(d_i-\mu_d)^2}$
3	CV_DURATION	Coefficient of variation of fixation duration	$CV_d = \frac{\sigma_d}{\mu_d}$
4	**PROP_LONG_FIX**	Ratio of long fixations.	$P_{\text{long}} = \frac{1}{N_f}\sum_{i=1}^{N_f}\mathbb{I}(d_i > \theta)$
5	**FIX_COUNT**	Total number of fixations	N_f
6	**SAC_COUNT**	Total number of saccades	$S_{count} = N_f - 1$
			$a_j = \sqrt{(x_{j+1}-x_j)^2+(y_{j+1}-y_j)^2} \quad j = 1,\ldots,N_f-1$
7	**SAC_MEAN_AMP**	Average saccade amplitude	$\mu_a = \frac{1}{N_f-1}\sum_{j=1}^{N_f-1} a_j$
8	SAC_STD_AMP	Standard deviation of saccade amplitude	$\sigma_a = \sqrt{\frac{1}{N_f-2}\sum_{j=1}^{N_f-1}(a_j-\mu_a)^2}$
9	SAC_MAX_AMP	Max saccade amplitude	$a_{\max} = \max_{1\le j\le N_f-1} a_j$
10	SPATIAL_STD_X	Horizontal spreads	$\sigma_x = \sqrt{\frac{1}{N_f-1}\sum_{i=1}^{N_f}(x_i-\bar{x})^2}$
11	SPATIAL_STD_Y	Vertical spread	$\sigma_y = \sqrt{\frac{1}{N_f-1}\sum_{i=1}^{N_f}(y_i-\bar{y})^2}$
12	**HM_ENTROPY**	Spatial entropy of fixation distribution	$H = -\sum_{k=1}^{K} p_k \log p_k$
13	HM_COMPACTNESS	Spatial compactness	$C = \frac{1}{N_f}\sum_{i=1}^{N_f}\sqrt{(x_i-\bar{x})^2+(y_i-\bar{y})^2}$
14	**PUPIL_STD**	Variability of pupil size	$\sigma_p = \sqrt{\frac{1}{N_f-1}\sum_{i=1}^{N_f}(p_i-\bar{p})^2}$
15	**PUPIL_CV**	Normalized pupil variability	$CV_p = \frac{\sigma_p}{\bar{p}}$
16	**PUPIL_DERIV_STD**	Fast fluctuations in pupil changes	$D_p = \sqrt{\frac{1}{N_f-2}\sum_{i=1}^{N_f-1}(\Delta p_i - \overline{\Delta p})^2}$

This formulation allows the constructed graph to capture both spatial clustering (node density) and temporal transitions (directed weighted edges) of eye-movement behavior. Fig.5 illustrates one graph built from eye movement data of one subject over 100 stimuli.

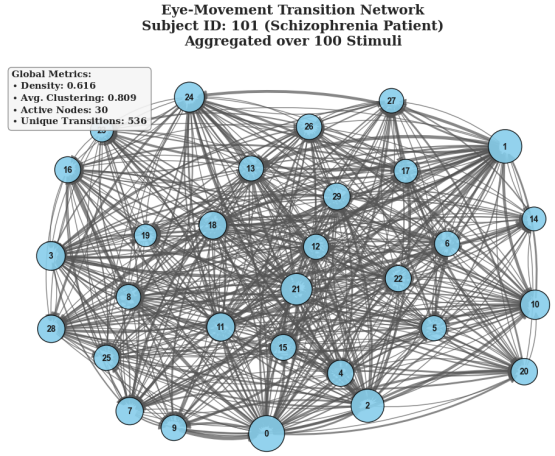

Fig. 5. Illustration of the graph built for a subject

3.3.2. Feature extraction from the constructed graph

From the constructed graph V, we generate a transition matrix $\mathbf{F}$ of size $(R \times R)$, where $\mathbf{F}_{ij}$ represents the number of times the gaze moves from node i to node j.

To ensure a uniform feature dimension across all samples for the subsequent machine learning models, we fix the maximum number of nodes to $R = 30$. For instances where the generated graph contains fewer than R nodes, the remaining rows and columns in the matrix $\mathbf{F}$ are zero-padded, which mathematically reflects the absence of transitions to non-existent nodes. Conversely, if a scanpath produces more than R nodes, the graph is truncated to retain only the first R nodes, thereby capturing the initial and most salient visual exploration phase.

The normalized matrix is then flattened into an R^2-dimensional feature vector. Besides these features, we extract eight more global metrics of the graph following [21]. The first four features represent the general information about the eye movement: 1) scanpath length L; 2) convex hull area A_{hull}; 3) fixation entropy (H); 4) mean fixation duration (μ_{dur}). The last four features represent the characteristics of the graph: 5) network density D; 6) average clustering coefficient; 7) strongly connected components SCC; and 8) degree assortativity r. Their formulas and meanings are presented in the following and summarized in Tab. 2.

1) *Scanpath Length (L):* The total spatial distance traveled by the gaze, computed as the sum of Euclidean distances between consecutive fixation points. According to [18, 22], a shorter path length often indicates restricted scanning, a hallmark of negative symptoms in Schizophrenia (SZ).

$$L = \sum_{i=1}^{R-1} \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2} \tag{21}$$

2) *Convex Hull Area (A_{hull}):* The resulting hull area measures the overall spatial spread of the subject's visual attention. Patients with Schizophrenia (SZ) tend to exhibit a smaller A_{hull}, reflecting reduced exploration and a tendency to avoid peripheral facial regions [22].

$$A_{hull} = \frac{1}{2} \left| \sum_{i=1}^{k-1} (x_i y_{i+1} - x_{i+1} y_i) + (x_k y_1 - x_1 y_k) \right| \tag{22}$$

where:

- A_{hull} denotes the area of the convex hull.
- k represents the number of vertices on the boundary of the convex hull.
- (x_i, y_i) are the Cartesian coordinates of the i-th vertex, ordered sequentially along the polygon's perimeter.
- (x_1, y_1) and (x_k, y_k) are the coordinates of the first and last vertices, respectively, closing the polygon.

3) *Fixation Entropy (H):* To measure the randomness of gaze distribution across $R = 30$ nodes, we calculated the Shannon Entropy based on the visit frequency p_k of each node. Lower entropy indicates a highly repetitive, rigid scanning pattern, whereas higher entropy indicates chaotic viewing. Modeling these transitions via N-grams and entropy has proven highly effective in distinguishing cognitive states [10, 23].

$$H = -\sum_{k=1}^{R} p_k \log p_k \tag{23}$$

4) *Mean Fixation Duration (μ_{dur}):* The average time spent per fixation, reflecting the speed of visual information processing. Prolonged fixations are strongly correlated with cognitive processing difficulties in SZ [3, 7].

$$\mu_{dur} = \frac{1}{R} \sum_{i=1}^{R} d_i \tag{24}$$

5) *Network Density (D):* The ratio of actual transitions to all possible transitions:

$$D = \frac{|E|}{|V|(|V| - 1)} \tag{25}$$

6) *Average Clustering Coefficient (C):* Computed on the undirected version of G to measure local cohesiveness (the tendency of ROIs to form tightly knit viewing clusters).

$$C = \frac{1}{R}\sum_{i=1}^{R}\frac{2e_i}{k_i(k_i-1)} \tag{26}$$

where:

- k_i is the degree of node i (the number of its neighbors).
- e_i is the actual number of edges between the neighbors of node i.

7) *Strongly Connected Components (SCC):* The number of subgraphs where every node is reachable from every other node. A higher number of isolated SCCs reveals a fragmented visual processing strategy, consistent with abnormal graph topology in SZ [11].

$$SCC(G) = \{V_c \subseteq V \mid \forall u, v \in V_c, \exists \text{ path } u \to v \text{ and } v \to u\} \tag{27}$$

where:

- V is the set of all nodes in the directed gaze graph G.
- V_c is a maximal subset of nodes where every pair (u,v) has a bidirectional path.
- The feature used in the model is $|SCC(G)|$, representing the total count of such components.

8) *Degree Assortativity (r):* Measures the tendency of a node to connect with other nodes of similar degree.

$$r = \frac{\sum_{jk} jk(e_{jk} - q_j q_k)}{\sigma_q^2} \tag{28}$$

where:

- j,k are the degrees of the nodes at the ends of an edge.
- e_{jk} is the joint probability distribution of the degrees of the two nodes at either end of a randomly chosen edge.
- q_j, q_k are the distributions of the remaining degree of nodes.
- σ_q is the standard deviation of the distribution q.

3.3.3. Feature selection

Extracting oculomotor, graph metrics, and R-gram transition frequencies yielded a massive 908-dimensional feature vector. Given our sample size ($N = 160$), this high dimensionality inherently triggers the "curse of dimensionality" and risks overfitting [24]. To overcome this, we select features based on their importance. Similar to the SFE, we also utilize Random Forest, Mutual Information and mRMR for feature selection.

RF calculates the *Feature Importance* based on the Mean Decrease in Gini Impurity (MDG) across its ensemble of decision trees:

$$Importance(f) = \frac{1}{T}\sum_{m=1}^{T}\sum_{t \in T_t}\Delta i(t,f) \tag{29}$$

where T is the number of trees and $\Delta i(t,f)$ is the decrease in Gini impurity at node t when split by feature f. By setting a strict threshold (max_features=160), we filtered the space down to the top-160 most discriminative features. Although alternative selection criteria such as Mutual Information (MI) and Minimum Redundancy Maximum Relevance (mRMR) were also evaluated during preliminary screening, RF-based importance was ultimately retained as the primary selection mechanism due to its superior stability and direct alignment with downstream ensemble classification. Fig. 6 illustrates the top discriminative features ranked by their relative importance scores. The results reveal a compelling distribution of predictive power across different feature categories. Notably, alongside traditional physical metrics (e.g., fixation duration), several advanced topological attributes (such as Network Density) and specific spatial-temporal transitions (N-grams) emerged among the highest-ranking predictors. This empirical finding strongly validates the hypothesis that integrating graph-based spatial metrics with temporal sequence analysis provides a more holistic and discriminative representation of schizophrenic viewing behaviors.

3.4. Proposed Fusion Method

To comprehensively capture the complex nature of schizophrenic viewing behaviors, we propose a fusion framework that seamlessly integrates traditional Eye Movement Synthesis (SFE) with Spatial-Temporal Analysis (STA). The fusion process consists of the following sequential stages:

1. Data integration and scaling: The graph-based metrics (STA) and traditional eye-tracking features (SFE) are initially concatenated horizontally for each subject. To mitigate the scale variations between disparate feature types (e.g., fractional transition probabilities vs. large fixation durations), a `RobustScaler` is applied.

2. Feature selection: We employ a two-stage feature selection strategy to identify the optimal feature subset.

- *Stage 1 - RF-based pre-filtering:* A Random Forest ($T = 100$, max depth $= 5$) ranks features via Mean Decrease in Gini, retaining 75–90 features to remove noisy and irrelevant variables.
- *Stage 2 - MI-based refinement:* Mutual Information is then used to select the top N features, preserving those with the highest relevance to the target ($N = 20$–35).

Table 2. Additional 8 features and top-5 features selected (in bold) for final evaluation

No.	Feature	Description	Formulation
1	**PathLen**	Total spatial distance covered by the eye.	$L = \sum_{i=1}^{R-1} \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2}$
2	**HullArea**	The maximum spatial spread of the subject's visual attention.	$A_{hull} = \frac{1}{2}\left\lvert \sum_{i=1}^{k-1} (x_i y_{i+1} - x_{i+1} y_i) + (x_k y_1 - x_1 y_k) \right\rvert$
3	**Entropy**	The randomness of gaze distribution across $R = 30$.	$H = -\sum_{k=1}^{R} p_k \log p_k$
4	**Duration**	The average time spent per fixation (speed of processing).	$\mu_{dur} = \frac{1}{R} \sum_{i=1}^{R} d_i$
5	**Density**	The ratio of actual transitions to all possible transitions.	$D = \frac{\lvert E \rvert}{\lvert V \rvert (\lvert V \rvert - 1)}$
6	Clustering	The tendency of ROIs to form tightly knit viewing clusters.	$C = \frac{1}{R} \sum_{i=1}^{R} \frac{2e_i}{k_i(k_i - 1)}$
7	NumComponents	The number of subgraphs where every node is reachable.	$SCC(G) = \{V_c \subseteq V \mid \forall u, v \in V_c, \exists u \leftrightarrow v\}$
8	Assortativity	The tendency of ROIs to connect with similar degree nodes.	$r = \frac{\sum_{jk} jk(e_{jk} - q_j q_k)}{\sigma_q^2}$

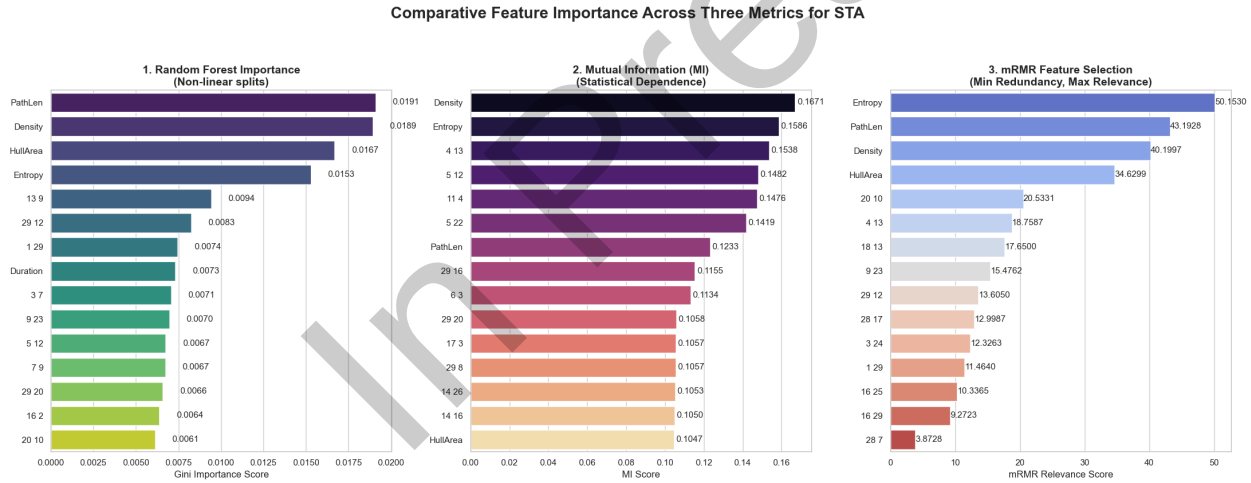

Fig. 6. Comparative ranking of STA features across three metrics

3. Model optimization and classification: The selected features are fed into multiple classifiers (Logistic Regression, SVM, Random Forest, LDA, and XGBoost). The entire pipeline is optimized using `GridSearchCV` with 4-fold stratified cross-validation on the training set (70%), ensuring robust performance and preventing data leakage.

4. Experiments

4.1. Dataset

In this paper, we utilized a benchmark dataset of eye movement EMS [12]. The eye movement data was collected in a controlled environment within a therapy room using a desktop-mounted EYELINK 1000 plus eye tracker. The system operated at a high-precision sampling rate of 1000 Hz. Participants were seated at a fixed distance of 0.6 meters from a 19-inch display screen with a resolution of 1024×768 pixels (Fig. **??**). To ensure data stability and minimize head movement artifacts, each subject's head was stabilized using a specialized chin support.

Data were recorded specifically from the eye that exhibited the smallest calibration error to ensure the highest quality of the resulting dataset. All procedures were strictly governed by ethical protocols approved by the ethics committee of the Shanghai Mental Health Center.

A total of 100 stimuli were presented, selected from the MIT and OSIE datasets, as well as supplementary internet sources. Each image was displayed for a duration of 5 seconds in a randomized order to prevent sequence bias. Each person has one data .csv file that follows

Table 3. Overview of the EMS* Eye-Tracking Dataset

Item	Value
Subjects	160 (balanced: 80 HC / 80 SZ)
Stimuli images	100
Sampling rate	1000 Hz
Total fixations	225,159 (raw)
Clean fixations	219,744
Features available	X, Y, duration, pupil size

a structured tabular format with the index of fixation, the duration of each gaze in milliseconds, the spatial X and Y coordinates in pixels, and the pupil size. Fig.9 illustrates four stimuli belonging to social, natural, synthetic, and manipulated scenes (from left to right).

Fig. 7 visualizes the average pupil size of two groups relative to fixation index in all stimuli images. We observe that the control group has larger pupil sizes at most of fixation indexes. Fig. 8 shows that Schizophrenia patients have longer fixation duration.

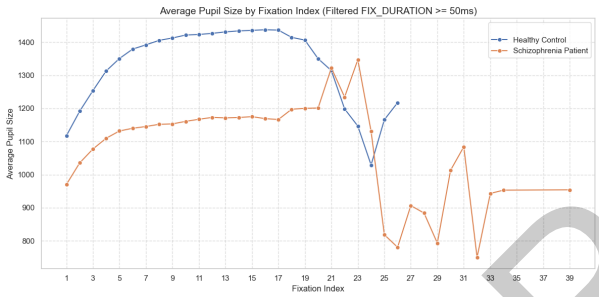

Fig. 7. Average Pupil Size of two groups relative to fixation index in all stimuli images

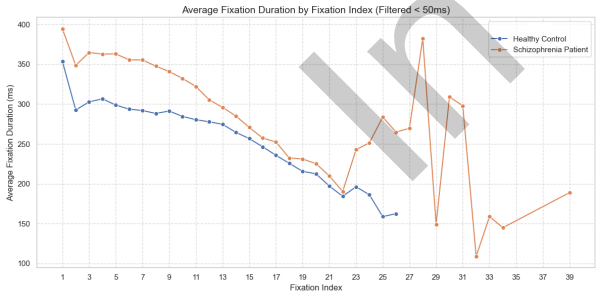

Fig. 8. Average Fixation Duration of two groups in time axis in all stimuli images

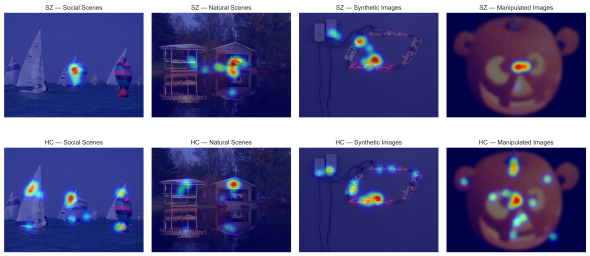

Fig. 9. Fixation density heatmaps overlaid on stimulus images. The SZ observer displays narrower and more concentrated attention compared to the HC observer

In the original EMS dataset, 104 healthy controls (HC) and 104 patients participated in the data collection. However, only 160 subjects are publicly annotated. In the original study [12], these 160 samples were used for training and validation with k-fold cross-validation ($k = 4$), while an additional 48 samples were reserved for testing. Since the labels of these 48 test samples are not provided in the EMS dataset, in this paper we use only the annotated set of 160 subjects for both training and evaluation. We name it as EMS* dataset. Tab. 3 summarizes the EMS* dataset.

We randomly split the 160 samples into 112 samples for training and validation and 48 samples for testing. The training portion is further evaluated using stratified k-fold cross-validation with $k = 4$ for model selection and hyperparameter tuning. Because this data split is not identical to that used in the original EMS study and results in a smaller effective training set, the performance comparison should be considered for reference purposes only.

4.2. Implementation Details

For machine learning algorithms, we set hyperparameters based on previous works. The parameters are summarized in Tab. 4. For the binary classification output, a standard decision threshold of $T = 0.5$ was applied. A subject was classified into the Schizophrenia class if the predicted probability $P(Y = 1|X) \geq 0.5$.

4.3. Experimental Results

The analysis was conducted on the processed EMS* dataset of 160 subjects, equally divided between the Schizophrenia (SZ) and Healthy Control (HC) groups. All behavioral features were computed at the subject level before classification.

Table 5 compares the classification performance of the two feature extraction strategies (SFE and STA) alongside the existing methods reported by Song *et al.* [12] on both Validation and Test sets. Regarding the independent **Test Set**, the **Random Forest** classifier using the **STA** approach achieves the highest overall robustness with an accuracy of **81.25%**, an F1-score of **81.63%**, and a sensitivity of **83.33%**. This performance notably surpasses both the SFE baseline (75.00% Acc) and the method proposed by Song *et al.* (72.92% Acc) for this specific classifier. These results indicate that the graph-based structural features selected via the STA method effectively capture discriminative gaze patterns that generalize well to unseen data.

In contrast, the **SFE** approach demonstrates superior performance for linear models on the test set, with both **LogReg** and **LDA** achieving a peak accuracy of **85.42%** and a high sensitivity of **91.67%**. However, the **STA** method proves to be more effective in terms of discriminative stability, consistently

Table 4. Hyperparameter configurations of machine learning models for the feature extraction strategies (STA, SFE, and Fusion)

Model	**Method**	**Key Hyperparameters**
Logistic Regression	STA	$C = 0.1$, solver = liblinear
	SFE	penalty = L_1, $C = 10$, StandardScaler + Pipeline
	Fusion	penalty = L_2, $C = 0.08$, solver = liblinear, mi_k = 32
SVM	STA	kernel = RBF, $C = 0.5$, $\gamma =$ scale, class_weight = balanced
	SFE	kernel = RBF, $C = 10$, $\gamma = 0.1$, StandardScaler + Grid Search
	Fusion	kernel = linear, $C = 0.05$, $\gamma =$ scale, mi_k = 25
Random Forest	STA	n_estimators = 500, max_depth = 8, min_samples_leaf = 3
	SFE	n_estimators = 300, max_depth = None, min_samples_leaf = 5
	Fusion	n_estimators = 40, max_depth = 4, max_features = sqrt, min_samples_leaf = 2, criterion = gini, mi_k = 35
LDA	STA	solver = lsqr, shrinkage = auto
	SFE	solver = svd, shrinkage = None
	Fusion	solver = lsqr, shrinkage = 0.6, mi_k = 25
XGBoost	STA	n_estimators = 500, learning_rate = 0.02, max_depth = 3
	SFE	n_estimators = 300, learning_rate = 0.03, max_depth = 3, subsample = 0.8, colsample_bytree = 1.0
	Fusion	n_estimators = 100, learning_rate = 0.08, max_depth = 3, subsample = 0.7, colsample_bytree = 0.6, mi_k = 30

Table 5. Comparison of performance (%) by two methods SFE , STA and the Fusion(SFE+STA) with existing methods on validation and test sets

Model	**Method, Year, Dataset**	**Validation Set**					**Test Set**				
		Acc	Prec	Sens	F1	AUC	Acc	Prec	Sens	F1	AUC
LogReg	Our SFE, 2026, EMS*	78.57 ± 4.37	77.74 ± 4.77	80.36 ± 5.92	78.91 ± 4.48	88.14 ± 3.63	85.42	81.48	**91.67**	86.27	89.58
	Our STA, 2026, EMS*	**80.36 ± 8.18**	81.11 ± 8.48	**80.36 ± 8.18**	**80.25 ± 8.2**	**87.76±4.52**	64.58	64.00	66.67	65.31	80.21
	Our Fusion, 2026, EMS*	**81.25 ± 9.24**	**82.89 ± 7.14**	81.25±9.24	80.69±10.03	87.50 ± 7.74	**87.50**	**88.57**	87.50	**87.41**	**91.32**
LDA	Our SFE, 2026, EMS*	78.57 ± 6.68	79.58 ± 5.57	76.79 ± 10.56	77.92 ± 7.39	86.35 ± 4.17	**85.42**	**81.48**	**91.67**	**86.27**	**90.10**
	Our STA, 2026, EMS*	**81.25±5.85**	**83.94±5.96**	**83.04±5.85**	**81.11±7.36**	**88.01±4.9**	72.92	70.37	79.17	74.51	82.99
	Our Fusion, 2026, EMS*	80.36±8.18	81.04±8.10	80.36±8.18	80.23±8.22	85.46±4.90	72.92	73.28	72.92	72.81	78.99
XGBoost	Our SFE, 2026, EMS*	75.89 ± 6.38	76.33 ± 5.50	75.00 ± 11.85	75.31 ± 7.47	84.57 ± 2.05	**79.17**	**79.17**	**79.17**	**79.17**	81.60
	Our STA, 2026, EMS*	75.89±10.83	76.99±11.41	75.89±10.83	75.66±10.93	**84.82±6.05**	72.92	70.37	**79.17**	74.51	81.08
	Our Fusion, 2026, EMS*	**80.36±5.36**	**81.47±4.63**	**80.36±5.36**	**80.12±5.57**	84.44±4.58	77.08	78.31	77.08	76.83	**85.07**
SVM (RBF)	ESR_SVM [25], 2020, EMS	79.38	78.59	**79.46**	78.89	84.98	75.00	77.27	70.83	73.91	77.60
	Our SFE, 2026, EMS*	**80.36 ± 6.44**	82.76 ± 9.32	78.57 ± 8.75	**80.00 ± 6.25**	87.63 ± 3.56	81.25	85.71	75.00	80.00	86.63
	Our STA, 2026, EMS*	72.32±5.28	**76.81±3.66**	72.32±5.28	70.87±6.31	**88.52±3.69**	68.75	73.68	58.33	65.12	81.94
	Our Fusion, 2026, EMS*	78.57±8.75	78.98±8.78	78.57±8.75	78.48±8.81	83.16±8.48	**85.42**	**87.02**	**85.42**	**85.26**	**90.80**
Random Forest	ESR_RF [25], 2020, EMS	76.25	75.49	75.40	75.38	85.21	72.92	73.91	70.83	72.34	77.60
	Our SFE, 2026, EMS*	76.79 ± 3.99	**81.35 ± 11.99**	73.21 ± 5.92	76.09 ± 1.81	85.46 ± 1.99	75.00	73.08	79.17	76.00	82.81
	Our STA, 2026, EMS*	77.68±10.22	78.43±10.39	77.68±10.22	77.52±10.28	**86.61±6.12**	81.25	80.00	**83.33**	81.63	85.07
	Our Fusion, 2026, EMS*	**79.46±4.46**	80.64±4.21	**79.46±4.64**	**79.23±4.83**	86.10±1.59	**83.33**	**85.56**	**83.33**	**83.07**	**89.76**
Deep Learning	MSNet [12], 2025, EMS	83.13	85.75	80.51	82.44	89.72	81.25	80.00	83.33	81.63	88.54
	DVP [26], 2022, EMS	78.75	82.06	75.48	77.69	85.16	75.00	77.27	70.83	73.91	81.94
	GPI-GRU [27], 2024, EMS	70.63	72.19	70.45	69.83	77.98	75.00	77.27	70.83	73.91	81.25

yielding competitive **AUC** scores across several models. Specifically, STA achieves an AUC of **82.99%** for LDA and **81.94%** for SVM on the test set. While STA generally exhibits lower sensitivity in certain configurations—such as **58.33%** for SVM—its ability to maintain a robust **AUC** of **85.07%** in the Random Forest model confirms that the integration of structural transition features provides a reliable metric for Schizophrenia recognition across different probability thresholds.

Overall, SFE demonstrates stronger performance in most metrics and models, while STA contributes meaningful structural information, supporting the potential benefit of combining both feature representations.

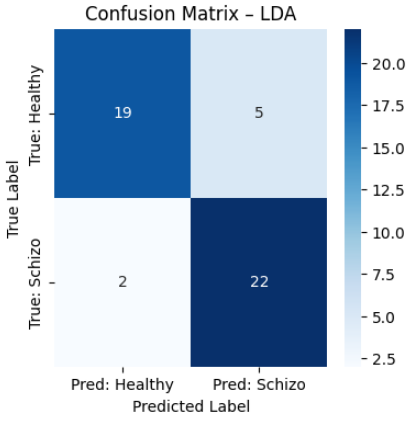

Fig. 10. SFE Confusion Matrix of LDA on test set

Fig. 10 visualizes the confusion matrix obtained by the best method (SFE+LDA). It shows that among 24 Healthy Control subjects, 5 are wrongly classified as Schizophrenia while only two Schizophrenia patients are misclassified as healthy controls. Fig. 11 illustrates

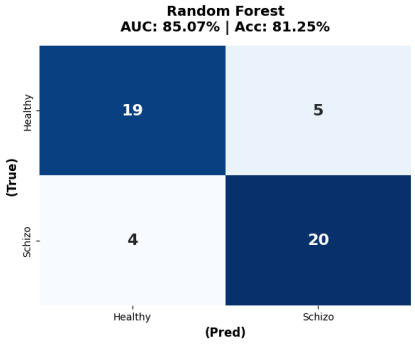

Fig. 11. STA Confusion Matrix of RF on test set

the confusion matrix for the best-performing approach (STA combined with Random Forest). The model demonstrates a robust diagnostic capability, correctly identifying 20 out of 24 Schizophrenia patients and 19 out of 24 Healthy Controls. Specifically, the misclassification involves only 5 healthy subjects being labeled as Schizophrenia (False Positives) and 4 patients being missed (False Negatives). This balanced distribution of errors confirms that the STA features provide a reliable boundary for psychiatric classification without a significant bias toward either class.

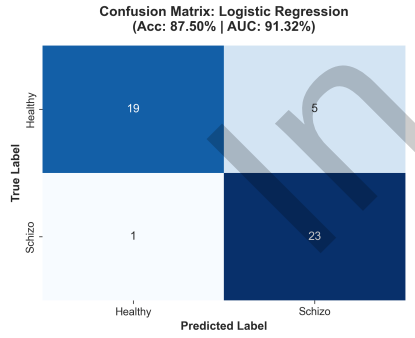

Fig. 12. Fusion(STA+SFE) Confusion Matrix of Logistic Regression on test set

Fig. 12 displays the confusion matrix for the proposed Fusion method leveraging Logistic Regression. The matrix reveals the model's exceptional sensitivity in disease detection. Specifically, it successfully identifies 23 out of 24 Schizophrenia patients, with only a single case misdiagnosed as healthy (False Negative). For the Healthy Control group, 19 out of 24 subjects are correctly classified, while 5 are misclassified as Schizophrenia (False Positives). In a clinical diagnostic context, this behavior is highly desirable; the model strongly prioritizes minimizing missed patient diagnoses (high sensitivity) while maintaining a reasonable specificity, demonstrating the practical clinical value of the fused SFE and STA features.

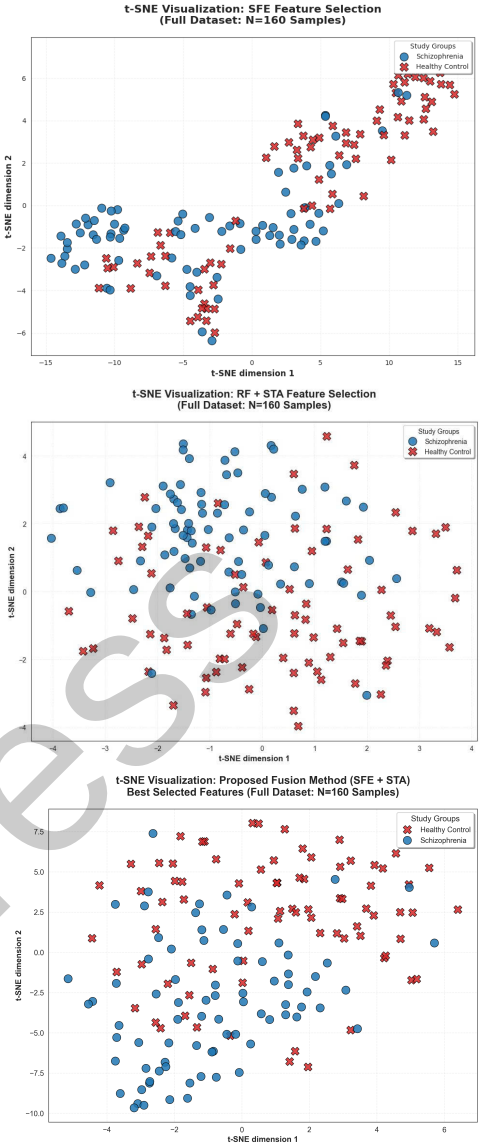

Fig. 13. t-SNE visualization comparing (first row) SFE (second row) STA and (last row) Fusion(SFE+STA) methods

5. Conclusion

This study developed a computational framework for objective Schizophrenia recognition using eye-tracking data, addressing the challenges of high-dimensional features and limited clinical samples. We proposed two complementary feature extraction methods: Statistical Feature Extraction (SFE) directly from the raw eye-tracking signals for a compact representation, and Spatial-Temporal Analysis (STA), which transforms raw fixation data into a graph structure to extract topological features. To maximize the descriptive power of the model, we also introduced a feature Fusion approach that combines both feature sets. In all approaches, feature selection was performed to reduce dimensionality using correlation analysis and Random Forest-based feature importance.

Due to the limited dataset size, the selected features were evaluated using conventional machine learning models. Experimental results show that the individual SFE and STA methods perform strongly, achieving test accuracies of 85.42% (using Logistic Regression) and 81.25% (using Random Forest), respectively. Most notably, our Fusion approach leverages the strengths of both methods to achieve the highest overall test accuracy of 87.50% with Logistic Regression. In future work, we plan to incorporate stimulus-level information into a deeper learning framework to further improve performance.

Acknowledgment

This research is funded by Hanoi University of Science and Technology (HUST) under project number T2025-PC-072.

References

[1] A. Di Cerbo, Convergences and divergences in the ICD-11 vs. DSM-5 classification of mood disorders, Turk Psikiyatri Dergisi, vol. 32, no. 4, pp. 293, Jan. 2021.

[2] M. B. First, D. E. Clarke, L. Yousif, A. M. Eng, N. Gogtay, and P. S. Appelbaum, DSM-5-TR: rationale, process, and overview of changes, Psychiatric Services, vol. 74, no. 8, pp. 869–875, Aug. 2023. https://doi.org/10.1176/appi.ps.20220334

[3] R. A. McCutcheon, R. S. Keefe, and P. K. McGuire, Cognitive impairment in schizophrenia: aetiology, pathophysiology, and treatment, Molecular psychiatry, vol. 28, no. 5, pp. 1902–1918, Jan. 2023.

[4] Z. Chen, Y. Ou, Y. Ding, Y. Wang, H. Li, F. Liu, P. Li, D. Lv, Y. Liu, B. Lang, J. Zhao, and W. Guo, Abnormal eye movement, brain regional homogeneity in schizophrenia and clinical high-risk individuals and their associated gene expression profiles, Schizophrenia, vol. 11, no. 1, pp. 64, Apr. 2025. https://doi.org/10.1038/s41537-025-00609-x

[5] K. Morita, K. Miura, A. Toyomaki, M. Makinodan, K. Ohi, N. Hashimoto, Y. Yasuda, T. Mitsudo, F. Higuchi, S. Numata, A. Yamada, A. Yohei, R. Honda Hiromitsu 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, pp. e56668, May. 2024. https://doi.org/10.2196/56668

[6] D. Zhang, L. Xu, X. Liu, H. Cui, Y. Wei, W. Zheng, Y. Hong, Z. Qian, Y. Hu, Y. Tang, C. Li, Z. Liu, T. Chen, H. Liu, T. Zhang, and J. Wang, Eye movement characteristics for predicting a transition to psychosis: longitudinal changes and implications, Schizophrenia Bulletin, vol. 51, no. 2, pp. 422–431, Jan. 2025. https://doi.org/10.1093/schbul/sbae001

[7] P. S. Holzman, L. R. Proctor, D. L. Levy, N. J. Yasillo, H. Y. Meltzer, and S. W. Hurt, Eye-tracking dysfunctions in schizophrenic patients and their relatives, Archives of General Psychiatry, vol. 31, no. 2, pp. 143–151, Aug. 1974. https://doi.org/10.1001/archpsyc.1974.01760140005001

[8] P. J. Benson, S. A. Beedie, E. Shephard, I. Giegling, D. Rujescu, and D. St. Clair, Simple viewing tests can detect eye movement abnormalities that distinguish schizophrenia cases from controls with exceptional accuracy, Biological Psychiatry, vol. 72, no. 9, pp. 716–724, May. 2012. https://doi.org/10.1016/j.biopsych.2012.04.019

[9] P.-H. Tseng, I. G. Cameron, G. Pari, D. P. Reynolds, D. P. Munoz, and L. Itti, High-throughput classification of clinical populations from natural viewing eye movements, Journal of Neurology, vol. 260, pp. 275–284, 2013.

[10] A. Coutrot, J. H. Hsiao, and A. B. Chan, Scanpath modeling and classification with hidden markov models, Behavior Research Methods, vol. 50, pp. 362–379, 2018.

[11] L. W. Stark, Abnormal patterns of normal eye movements in schizophrenia. Schizophrenia bulletin, vol. 9, no. 1, pp. 55–72, 1983. https://api.semanticscholar.org/CorpusID:38104481

[12] Y. Song, Z. Liu, G. Li, J. Xie, Q. Wu, D. Zeng, L. Xu, T. Zhang, and J. Wang, EMS: A Large-Scale Eye Movement Dataset, Benchmark, and New Model for Schizophrenia Recognition, IEEE Transactions on Neural Networks and Learning Systems, vol. 36, no. 5, pp. 9451–9462, May. 2025. https://doi.org/10.1109/TNNLS.2024.3441928

[13] R. Dang, Y. Wang, F. Zhu, X. Wang, J. Zhao, P. Shao, B. Lang, Y. Wang, Z. Pan, B. Hu, *et al.*, Classification of schizophrenia based on RAnet-ET: Resnet based attention network for eye-tracking, Journal of Neural Engineering, vol. 22, no. 2, pp. 026053, 2025.

[14] M. Kim, J. Lee, S. Y. Lee, M. Ha, I. Park, J. Jang, M. Jang, S. Park, and J. S. Kwon, Development of an eye-tracking system based on a deep learning model to assess executive function in patients with mental illnesses, Scientific Reports, vol. 14, no. 1, pp. 18186, Aug. 2024. https://doi.org/10.1038/s41598-024-68586-2

[15] S. Sugiyanto, I. K. E. Purnama, E. M. Yuniarno, and M. H. Purnomo, Early detection depression based on action unit and eye gaze features using a multi-input CNN-WoPL framework, Jurnal Nasional Pendidikan Teknik Informatika: JANAPATI, vol. 13, no. 3, pp. 642–657, Dec. 2024. https://doi.org/10.23887/janapati.v13i3.84674

[16] O. Kobo, A. Meltzer-Asscher, J. Berant, and T. Schonberg, Classification of depression tendency from gaze patterns during sentence reading, Biomedical Signal Processing and Control, vol. 93, pp. 106015, 2024.

[17] K. Avramidis, W. Jeong, A. Kommineni, S. R. Kadiri, M. Ma, C. McDaniel, M. Hughes, T. McGee, E. Kaiser, D. Byrd, *et al.*, Deep Learning Characterizes Depression and Suicidal Ideation from Eye Movements, arXiv preprint arXiv:2504.20944, 2025.

[18] S. Dowiasch, B. Backasch, W. Einhäuser, D. Leube, T. Kircher, and F. Bremmer, Eye movements of patients with schizophrenia in a natural environment, European Archives of Psychiatry and Clinical Neuroscience, vol. 266, no. 7, pp. 541–553, 2016. https://doi.org/10.1007/s00406-014-0567-8

[19] B. Zuber, L. Stark, and G. Cook, Microsaccades and the Velocity-Amplitude Relationship for Saccadic Eye Movements, Science (New York, N.Y.) vol. 150, no. 3702, pp. 1459–1460, Jan. 1966. https://doi.org/10.1126/science.150.3702.1459

[20] H. Jarodzka, K. Holmqvist, and H. Gruber, Eye tracking in Educational Science: Theoretical frameworks and research agendas, Journal of Eye Movement Research, vol. 10, no. 1, Feb. 2017. https://doi.org/10.16910/jemr.10.1.3

[21] A. Bulling, D. Roggen, and G. Tröster, Eye movement analysis for activity recognition using electrooculography, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 4, pp. 741–753, May. 2011. https://doi.org/10.1109/TPAMI.2010.86

[22] R. Yee, R. Baloh, S. Marder, D. Levy, S. Sakala, and V. Honrubia, Eye movements in schizophrenia, Investigative ophthalmology & visual science, vol. 28, no. 2, pp. 366–74, Mar. 1987.

[23] Y. Miyagi, N. Rodrigues, D. Weiskopf, and T. Itoh, Extraction of AOI transition patterns by N-grams from symbolized eye tracking scan-paths and visualization as trajectories N-gram, The Journal of the Society for Art and Science, vol. 18, pp. 125–133, Nov. 2019. https://doi.org/10.3756/artsci.18.125

[24] H. Peng, F. Long, and C. Ding, Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1226–1238, Aug. 2005. https://doi.org/10.1109/TPAMI.2005.159

[25] L. Huang, W. Wei, Z. Liu, T. Zhang, J. Wang, L. Xu, W. Chen, and O. Le Meur, Effective schizophrenia recognition using discriminative eye movement features and model-metric based features, Pattern Recognition Letters, vol. 138, pp. 608–616, Oct. 2020. https://doi.org/10.1016/j.patrec.2020.09.017

[26] C. Xia, D. Zhang, K. Li, H. Li, J. Chen, W. Min, and J. Han, Dynamic viewing pattern analysis: Towards large-scale screening of children with ASD in remote areas, IEEE Transactions on Biomedical Engineering, vol. 70, no. 5, pp. 1622–1633, Nov. 2022. https://doi.org/10.1109/TBME.2022.3223736

[27] W. Zhou, M. Yang, J. Tang, J. Wang, and B. Hu, Gaze patterns in children with autism spectrum disorder to emotional faces: Scanpath and similarity, IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 32, pp. 865–874, Feb. 2024. https://doi.org/10.1109/TNSRE.2024.3361935