Equivalent: Jurnal Ilmiah Sosial Teknik
Volume 8, Issue 5, 2159-2175
e_ISSN: 2775-0833
https://jurnalequivalent.id/index.php/jequi/index
DOI: doi.org/10.59261/jequi.v8i5.441

Multimodal Hybrid Deep Learning for the Detection of Autism Spectrum Disorder (ASD) Using Eye-Tracking Data

Citra Wanodya Prameswari
Universitas Telkom,
Indonesia

Erwin Budi Setiawan*
Universitas Telkom,
Indonesia

***Corresponding author:**
Erwin Budi Setiawan,
Universitas Telkom, Indonesia,
✉erwinbudisetiawan@telkomuniversity.ac.id

Article Info:
***Article history*:**
Received: August 19, 2026
Revised: October 10, 2026
Accepted: October 12, 2026
Available Online: November 04, 2026

***Keywords*:**
Autism Spectrum Disorder;
Deep Learning;
Eye Tracking;
Late Fusion;
Multimodal Learning

This is an open access article under the CC BY-SA license.

Copyright © 2026 by Author. Published by Politeknik Siber Cerdika International

Abstract

Background: Autism spectrum disorder (ASD) assessment may benefit from objective measures of visual attention, although eye-tracking classifiers remain research tools rather than replacements for clinical diagnosis.

Objective: This study compared eye-tracking representations and sequence encoders for binary classification of ASD versus typical development using the Saliency4ASD dataset.

Methods: Eye-movement sequences, fixation maps, stimulus images, and tabular features combining gaze and stimulus content were investigated in a hybrid learning pipeline that combined sequence encoders, ResNet-18 image-processing branches, and a multilayer perceptron (MLP) through feature concatenation. The sequential experiments evaluated six data-partition ratios, 15 modality combinations, seven optimization configurations, and alternative sequence architectures. The selected model was also evaluated across five additional holdout partitions grouped by stimulus image.

Results: The combination of scanpath and fixation-map representations achieved a file-level F1 score of 0.866 during modality confirmation. The long short-term memory (LSTM) + gated recurrent unit (GRU) configuration achieved a file-level F1 score of 0.9333 and a receiver operating characteristic area under the curve (ROC-AUC) of 0.9733 on the originally selected test partition comprising 30 stimulus images. Five additional image-grouped partitions yielded a mean file-level F1 score of 0.839 ± 0.015. The random forest baseline achieved an F1 score of 0.8999 on the original partition; however, it was not evaluated across the additional partitions.

Conclusion: The results demonstrate the potential of combining temporal and spatial gaze representations while highlighting sensitivity to data-partition selection. Reuse of test data during model tuning, limited subject independence, and the absence of verified generalization to previously unseen children preclude claims of independently validated diagnostic performance.

To cite this article: Prameswari, C. W., & Setiawan, E. B. (2026). Multimodal hybrid deep learning for the detection of autism spectrum disorder (ASD) using eye-tracking data. *Equivalent: Jurnal Ilmiah Sosial Teknik*, *8*(5), 2159–2175. https://doi.org/10.59261/jequi.v8i5.441

INTRODUCTION

Autism spectrum disorder (ASD) is a neurodevelopmental condition associated with differences in social communication and restricted or repetitive behaviors (Comparan-Meza et al., 2021; Tsai et al., 2020; Wang et al., 2023). In a systematic review and meta-analysis published in 2022, Salari et al. (2022) examined 74 studies comprising 30,212,757 participants and estimated a pooled global ASD prevalence of 0.6% (95% CI: 0.4%–1.0%). This estimate reflects reported prevalence across heterogeneous studies and does not constitute evidence of an increase in ASD

incidence. Earlier identification can support timely assessment and intervention, but established clinical instruments, such as the Autism Diagnostic Observation Schedule (ADOS) and the Modified Checklist for Autism in Toddlers (M-CHAT), require appropriate administration and interpretation by trained personnel. These considerations motivate research into objective behavioral indicators, including eye-tracking measures, as potential adjuncts to, rather than replacements for, clinical assessment (Salari et al., 2022).

The most relevant prior studies illustrate both the potential and the limitations of eye-tracking representations. Kanhirakadavath (2022) investigated scanpath-based ASD screening, highlighting the utility of gaze behavior but not jointly modeling sequential gaze coordinates and spatial fixation maps through the late-fusion approach examined in the present study. Han et al. (2022) integrated electroencephalography (EEG) and eye-tracking signals, demonstrating the value of complementary information sources. However, their approach fused distinct measurement modalities rather than multiple representations derived from eye-tracking data itself. Mumenin et al. (2024) modeled fixation maps and heatmaps using an involution-based architecture applied to ASD eye-tracking data, but did not evaluate the same systematic combination of scanpath sequences, fixation maps, stimulus images, and gaze–content tabular features examined here. These studies motivate a focused comparison of complementary representations without implying that all existing eye-tracking approaches have been exhaustively surveyed.

Efforts to combine multiple behavioral or physiological information sources have increasingly been reported in recent studies. However, the combined inputs generally originate from distinct signal types rather than from decomposing a single behavioral signal into several complementary representations. For example, Han et al. (2022) presented a framework that combined EEG signals with eye-tracking data using a two-stage stacked denoising autoencoder. Evaluated on data from 40 children with ASD and 50 typically developing (TD) children, the framework outperformed unimodal EEG and eye-tracking models on the same binary classification task. Zhu et al. (2023) developed a multimodal machine-learning system that combined audio and visual data collected during a response-to-name procedure. The system differentiated toddlers with ASD from non-ASD groups with an accuracy of 74.8% using computer-based assessment. The authors nevertheless emphasized that a single behavioral indicator, such as response to name, is insufficient for comprehensive ASD detection. Chen et al. (2020) proposed another multimodal framework combining eye-fixation features, facial expressions, and cognitive-level information through a hybrid fusion scheme based on a random forest classifier, achieving 91% accuracy. However, the eye-fixation features in that study were represented only by the number of fixation points within each *k*-means cluster. Neither the temporal sequence of the scanpath nor the spatial distribution of fixations represented as an image was modeled. Simon et al. (2024) explicitly discussed a fusion scheme for eye-tracking data but used the term to describe a combination of machine-learning and deep-learning methods applied to the same type of gaze data. The integration of multiple distinct eye-tracking representations was not the focus of that study. A similar pattern emerges from a broader mapping study of eight ASD-related behavioral modalities. Rakotomanana (2025) reported that multimodal combinations in the literature generally involve different types of behavioral information, such as audio and facial gestures or visual attention and kinematic data. The eye-tracking signal itself is rarely decomposed into multiple complementary representations within a single architecture.

The specific research gap concerns the joint evaluation of scanpath sequences, spatial fixation maps, stimulus images, and gaze–content tabular descriptors within a common representation-fusion pipeline, followed by controlled comparisons of modality subsets, optimization configurations, sequence encoders, and partition sensitivity. Previous studies have investigated related representations and multimodal combinations, but the cited studies do not report the same combination of sequential ablation experiments and *image_id*-grouped evaluation described here. This study investigates late fusion through the concatenation of learned branch embeddings and distinguishes its methodological contribution from any claim of a clinically validated diagnostic advance. Importantly, grouping samples by *image_id* prevents the same stimulus image from appearing in both the training and holdout partitions of a given split.

However, this procedure does not establish independence among the children represented in those partitions.

Several encoder architectures previously applied to similar tasks were considered as candidates for this study rather than selected without empirical or methodological justification. Four candidate architectures were considered for processing sequential scanpath data. Long short-term memory (LSTM) networks were included because their memory cells and gating mechanisms help mitigate the vanishing-gradient problem and capture temporal dependencies (Fan et al., 2024). Bidirectional LSTM (BiLSTM) networks extend this approach by processing sequences in both forward and backward directions, allowing the model to incorporate contextual information from the full observation sequence when generating representations (Fan et al., 2024). Gated recurrent unit (GRU) networks were considered as a computationally lighter alternative to LSTMs, although their relative performance depends on the task and experimental configuration. A one-dimensional convolutional neural network (1D CNN) was included as a non-recurrent comparator for sequence modeling. ResNet-18 was selected for image representation learning because its residual connections facilitate the training of deeper networks, its computational requirements are relatively modest, and its architecture is suitable for transfer learning when training data are limited (Alzubaidi et al., 2021). The four sequence-encoder candidates were systematically compared during the experiments to identify the best-performing architecture under the study's evaluation protocol.

The study aims to (1) evaluate six training/validation/test partition ratios using file-level F1 scores in Experiment 1; (2) compare 15 combinations of eye-tracking, stimulus-image, and gaze–content representations against unimodal alternatives using file-level F1 scores and related classification metrics in Experiment 2; (3) assess seven optimization configurations in Experiment 3; (4) compare sequence-encoder architectures and two ablation settings in Experiment 4; and (5) evaluate the selected model against classical machine-learning baselines and across five additional *image_id*-grouped holdout partitions using F1 score, accuracy, precision, recall, and, where available, the area under the receiver operating characteristic curve (ROC-AUC). The methodological contribution lies in documenting the relative performance of different input representations and examining the sensitivity of model selection to repeated use of data partitions. However, the findings do not establish clinical readiness or generalizability to independent subjects.

This study is restricted to the Saliency4ASD dataset, which comprises 28 children (14 with ASD and 14 typically developing children), 300 stimulus images, and a binary classification task using up to four candidate representations. The *image_id*-based split restricts stimulus-image overlap between partitions but does not create an independent cohort of children. According to the implementation, *subject_id* identifies a subject–image block rather than providing a stable identifier for each child across different images. Accordingly, the reported “subject-level” metric must not be interpreted as evidence of validated generalization to previously unseen children. The primary file-level evaluation uses a classification threshold of 0.5 and repeated holdout partitions, while the initially selected test partition is reused during model selection. The theoretical value of the study lies in comparing gaze representations, its methodological value in exposing partition-dependent performance, and its practical value in providing a proof of concept for further independently validated research rather than a deployable screening or diagnostic system.

METHOD

An experimental quantitative approach with a sequential ablation study design was used in this study. The best-performing configuration from each stage was carried forward to the next, as the four main experiments were conducted sequentially. After the optimal configuration had been established, the selected model was compared with a classical machine-learning baseline during the final evaluation. The configuration search space, which included data partition ratios, modality combinations, hyperparameters, and architectures, was too large to explore exhaustively through a full grid search given the available computational resources. Therefore, a sequential design was adopted as a compromise between experimental scope and computational efficiency, progressively narrowing the search space at each stage.

The primary focus of the experimental design was to examine the effect of late-fusion schemes combining multiple eye-tracking representations on classification performance relative to unimodal approaches, addressing the second research problem. The remaining experimental questions concerned modality combinations, optimization methods, sequence-encoder architectures, and validation protocols, which were investigated through Experiments 2–4 and the associated evaluation procedures.

Experiment 1 served specifically as a methodological preparation stage rather than directly addressing a single core research problem. Its purpose was to determine the data partition ratio that would be used consistently throughout the subsequent experiments, from Experiment 2 through the final evaluation. This ensured that configuration-selection stages could be compared using identical data partitions. Thus, Experiment 1 differed in purpose from Experiments 2, 3, and 4, which directly addressed the third, fourth, and fifth research problems, respectively.

The overall research workflow is summarized in Figure 1 and consists of four major stages. The first stage involved the Saliency4ASD dataset and initial preprocessing, including correlation-based filtering, StandardScaler normalization, and median imputation. The second stage involved extracting four groups of eye-tracking data representations: scanpath sequences, fixation maps, stimulus images, and tabular features derived from gaze–content fusion. The third stage comprised the four main experiments, which progressively narrowed the configuration space by evaluating data partition ratios (Experiment 1), modality combinations (Experiment 2), optimization configurations (Experiment 3), and sequence-encoder architectures (Experiment 4). The fourth stage involved the final evaluation, in which the selected deep-learning model was compared with classical machine-learning baselines. A detailed discussion of each stage, including all 15 modality combinations tested in Experiment 2, is provided in the following section.

Table 1
Research Design Summary

Aspects	Remarks
Types of research	Applied *research*
Method	Quantitative experimental, deep learning-based
Approach	Comparative
Independent variables	Split ratios, modality combinations, optimizers, learning rates, architecture
Bound variables	File-level F1, accuracy, precision, recall, AUC
Control variables	seed = 42, threshold = 0.5 (fixed), K = 5 fold, group = image_id

Figure 1 is read from left to right, following the four stages. First, the raw eye-tracking data from the Saliency4ASD dataset, comprising 300 images and 28 children, underwent preprocessing. The processed data were then used to construct the four representation groups in the second stage. In the third stage, these representations passed through the experimental funnel, in which the data partition ratio, modality combination, optimization configuration, and sequence-encoder architecture were determined sequentially through Experiments 1–4. The best-performing configuration from each experiment was carried forward to the next. Finally, the selected configuration from the third stage was evaluated in the fourth stage by comparison with classical machine-learning baselines, providing a basis for assessing the relative performance of the two approaches.

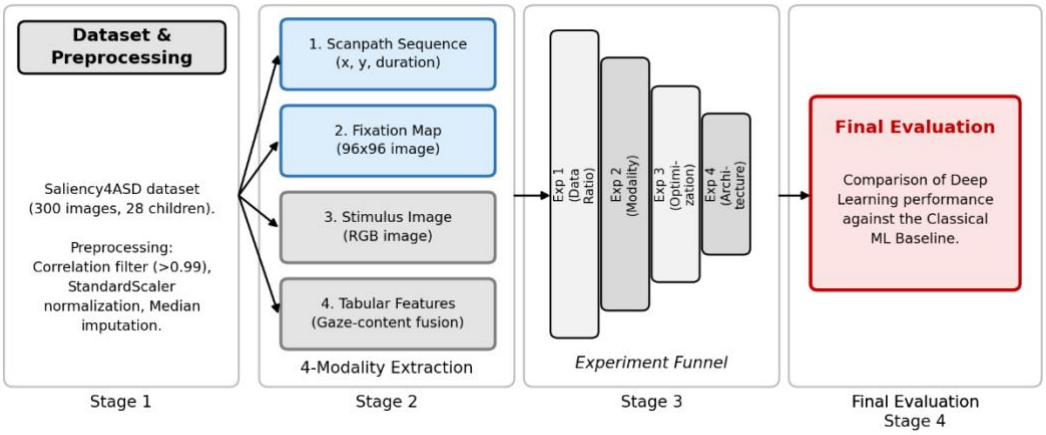


Figure 1. Overall research framework flow

The Saliency4ASD public dataset served as the data source for this study. It contains eye-tracking recordings of children with autism spectrum disorder (ASD) and typically developing (TD) children as they viewed 300 stimulus images derived from the MIT1003 dataset (Duan et al., 2019). The study population comprised 28 participants: 14 children with ASD and 14 TD children, aged 5–12 years, with a mean age of 8 years. A Tobii T120 eye tracker was used to collect the data at a screen resolution of 1280 × 1024 pixels. The viewing distance was approximately 65 cm, and each stimulus image was displayed for 3 seconds. On average, the stimulus images were viewed by approximately 14 children with ASD and 14 TD children, although the exact number of observers varied across images because some participant records were incomplete.

The class distributions at both the subject–image-block and file levels are summarized in Table 3. The relatively balanced class distribution was expected to reduce the risk of bias attributable to class imbalance, potentially obviating the need for additional oversampling or undersampling. At the file level, the class distribution was exactly balanced, with 300 files per class, because each stimulus image generated one file for the ASD group and one for the TD group by definition.

Table 2
Saliency4ASD Data Components

Components	Location	Remarks
Fixation Data	TrainingData/ASD, TD	Fixation (x, y, duration) per subject per image
Stimulus	TrainingData/Image	300 PNG images (from MIT1003) Annotations
Semantics	ImageAnnotations.csv	*Bounding box* 15 object classes
Population map	ASD_FixMaps, TD_FixMaps	Not used as an input model

Table 3
Class Distribution on Dataset

Aggregation Rate	Total	ASD	TD
Block (subject × image)	7.598	49,5%	50,5%
Files (images × class)	up to 600	300	300

The positions at which participants' eyes paused during fixation intervals were represented by the ((x, y)) coordinates recorded in the scanpath data. These coordinates corresponded to a screen resolution of 1280 × 1024 pixels, consistent with the recording configuration of the Tobii T120 eye tracker described above. Because the stimulus images were displayed in full-screen mode during recording, the recorded gaze coordinates were assumed to correspond directly to pixel locations on the displayed stimulus images, without additional geometric transformation or recalibration. This correspondence formed the basis of the gaze–content fusion process. Specifically, each fixation point was matched against the annotated object bounding boxes on the corresponding stimulus image to identify the object being viewed at that point.

The research focus was aligned with the characteristics of the dataset, which supported the extraction of eye-tracking features, visual image features, and gaze–content interaction features. Object annotations provided information about the visual content of each image, whereas gaze–content interaction features were derived by associating fixation points with the annotated object regions in the corresponding images.

The overall research workflow, from dataset acquisition and preprocessing to final evaluation, is presented in Figure 1 as four major stages. The diagram provides a simplified overview of these stages. Additional methodological details, including the fixed partition ratio, the reuse of the test partition in Experiments 2–4, and the associated risk of optimistic bias, are described in the following paragraphs and discussed further below.

The analysis involved three levels of observational units: subject–image blocks, subjects as represented by the available identifiers, and files. The block level consisted of individual subject–image observations, totaling 7,598 samples. The subject level was based on aggregation using subject_id, while the file level consisted of image–class pairs, totaling up to 600 files, as reported in Table 3. In this implementation, subject_id was uniquely assigned to each subject–image combination (block) rather than serving as a stable identifier for the same child across different images. Consequently, metrics aggregated using this column should not be interpreted as validated child-level performance. The identifier was also not used as the grouping variable for cross-validation. Instead, image_id, which identified each of the 300 stimulus images, was used to group the data during cross-validation. This ensured that the same stimulus image did not appear simultaneously in the training and validation folds, regardless of how many children had viewed it.

Among the six candidate partition ratios evaluated in Experiment 1, the 80:10:10 training/validation/test ratio yielded the highest file-level F1 score on the holdout partition, at 0.833. This ratio was subsequently used to partition the data by image rather than by individual record. Specifically, 240 images were assigned to training, corresponding to 6,093 blocks; 30 images were assigned to validation, corresponding to 757 blocks; and the remaining 30 images were assigned to testing, corresponding to 748 blocks. The test partition represented 60 image–class files. These partitions were fixed and reused consistently in Experiments 2, 3, and 4, as well as in the final evaluation.

Several measures were taken to maintain consistency and fairness in the evaluation protocol, although important limitations remained. The training, validation, and test partitions were established once at the image level using the 80:10:10 ratio selected in Experiment 1 and were then reused throughout Experiments 2–4 and the final evaluation. This ensured that the configurations could be compared using identical data partitions. However, the test partition did not function as a strictly independent, locked test set because its performance also informed the selection of modality combinations in Experiment 2, optimization configurations in Experiment 3, and architectures in Experiment 4 before the final results were reported. Repeatedly using the same test partition for configuration selection introduced a risk of implicit overfitting to that partition. Therefore, the final performance scores must be interpreted with the possibility of optimistic bias in mind rather than as unbiased estimates of generalization to independent data.

Sensitivity to the choice of data partition was also examined by retraining the selected configuration on five additional data partitions. Grouping by image_id during cross-validation was intended to prevent the same stimulus image from appearing in both the training and validation sets. This decision was motivated by evidence that conventional cross-validation methods that ignore the clustered structure of data can substantially overestimate classification accuracy when samples are correlated because they share subjects or experimental conditions (White & Power, 2023). This concern was relevant to the present study because a single stimulus image could be viewed by multiple children, producing numerous data blocks associated with the same image. Grouping by image therefore provided a more realistic estimate of generalization to previously unseen stimulus images than trial-level validation, although it did not establish generalization to previously unseen children (Nurhasan & Kasprowski, 2026).

RESULTS AND DISCUSSION

Results

Experiment 1: Data Partition Ratio

Six candidate data partition ratios were tested at this stage: 70:30, 80:20, 90:10, 70:15:15, 80:10:10, and 60:20:20. The full results are presented in Table 4.

Table 4

Experiment Result 1: Split Ratio Comparison

Ratios/Methods	FileF1	Accuracy	Precision	Recall
80:10:10	0,833	0,833	0,839	0,833
60:20:20	0,825	0,825	0,827	0,825
80:20	0,819	0,825	0,870	0,825
70:30	0,815	0,817	0,826	0,817
90:10	0,796	0,800	0,823	0,800
70:15:15	0,758	0,766	0,811	0,766

Experiment 1 yielded the highest holdout file-level F1 score for the 80:10:10 partition (0.833), followed by 60:20:20 (0.825), 80:20 (0.819), 70:30 (0.815), 90:10 (0.796), and 70:15:15 (0.758). The 80:10:10 partition was carried forward into subsequent experiments. However, the reported results do not establish an independent validation-based criterion for selecting this ratio. The test partition was subsequently reused to select modalities, optimization settings, and architectures. Consequently, these rankings should be regarded as exploratory configuration comparisons rather than independent estimates of relative generalization performance.

Experiment 2: Modality Combination Selection

Fifteen combinations of four modality groups were evaluated during the screening stage using SCREEN_CFG (16 epochs). The two highest-performing combinations were then evaluated during the confirmation stage using SEARCH_CFG (35 epochs). The screening results are presented in Table 5.

During screening and confirmation, the Eye Tracking-only combination achieved the highest file-level F1 score. This configuration comprised the scanpath sequence and FixMap branches, without stimulus imagery or tabular fusion features. It outperformed the full combination of all four modality groups, which achieved an F1 score of 0.833 during screening. By contrast, the Stimulus Image-only configuration produced near-chance performance, with an F1 score of approximately 0.471. This result is plausible because all subjects in the dataset viewed identical stimulus images; therefore, the visual content of the images alone was insufficient to distinguish ASD from typically developing (TD) participants. Findings from another study, in which a model based on a single scanpath achieved an area under the curve (AUC) above 0.90 without additional modalities, are consistent with the observation that simpler modality combinations can outperform the full combination (Kanhirakadavath & Chandran, 2022). Accordingly, most of the ASD–TD discriminative information in this dataset may be contained in gaze behavior itself. Tabular fusion features and stimulus imagery may increase model complexity without contributing commensurate information, potentially increasing the risk of overfitting in high-dimensional feature spaces with limited sample sizes.

Table 5

Experimental Results 2: Screening 15 Modality Combinations

No.	Modality Combinations	Features	File F1	Accuracy	Precision
1.	Eye Tracking	21	**0,865**	**0,867**	**0,882**
2.	Stimulus Image + Fusion Magnitude	111	0,850	0,850	0,850
3.	Eye Tracking + Fusion Magnitude	87	0,849	0,850	0,860
4.	Eye Tracking + Stimulus + Fusion Pattern	162	0,849	0,850	0,860
5.	Eye Tracking + Stimulus + Fusion Mag. + Fusion Pat. (full)	228	0,833	0,833	0,835
6.	Fusion Magnitude + Fusion Pattern	162	0,833	0,833	0,835
7.	Eye Tracking + Fusion Pattern	117	0,833	0,833	0,839

No.	Modality Combinations	Features	File F1	Accuracy	Precision
8.	Eye Tracking + Fusion Magnitude + Fusion Pattern	183	0,833	0,833	0,839
9.	Eye Tracking + Stimulus Image	66	0,817	0,817	0,817
10.	Fusion Magnitude	66	0,816	0,817	0,820
11.	Eye Tracking + Stimulus + Fusion Magnitude	132	0,799	0,800	0,805
12.	Stimulus + Fusion Magnitude + Fusion Pattern	207	0,766	0,767	0,771
13.	Fusion Pattern	96	0,514	0,533	0,540
14.	Stimulus Image + Fusion Pattern	141	0,514	0,533	0,540
15.	Stimulus Image	45	0,471	0,500	0,500

Table 6
Experiment Results 2: Confirm the Two Best Combinations

Modality Combinations	File F1	Accuracy	Precision
Eye Tracking	0,866	0,867	0,873
Stimulus Image + Fusion Magnitude	0,816	0,817	0,820

Experiment 3: Hyperparameter Optimization

Seven optimization configurations were reported during the screening stage (Table 7), followed by the confirmation of five configurations (Table 8). The screening table contains two rows labeled Adam, with F1 scores of 0.866 and 0.830, respectively. However, the manuscript does not specify the configuration that distinguishes these entries. Both results have therefore been retained without assigning an unreported parameter setting to either entry. The highest confirmed file-level F1 score was 0.900 for the LR Grid Sweep configuration. However, this comparison remains subject to the limitations associated with reusing the holdout partition for model selection.

Table 7
Experiment Results 3: Seven Methods of Optimization (Screening)

No.	Method	File F1	Accuracy	Precision
1.	Optuna TPE	0,883	0,883	0,887
2.	LR Grid Sweep	0,883	0,883	0,894
3.	Adam	0,866	0,867	0,873
4.	RMSprop	0,866	0,867	0,873
5.	Hidden Dim Sweep	0,847	0,850	0,885
6.	Adam	0,830	0,833	0,859
7.	SGD	0,799	0,800	0,805

Table 8
Experiment Results 3: Confirm the Best Method

Method	File F1	Accuracy	Precision
LR Grid Sweep	0,900	0,900	0,907
Optuna TPE	0,899	0,900	0,917
Hidden Dim Sweep	0,848	0,850	0,870
RMSprop	0,783	0,783	0,786
Adam	0,750	0,750	0,750

The top two positions in both the screening and confirmation stages were consistently occupied by LR Grid Sweep, with a learning rate of 0.001, and Optuna TPE. Their F1 scores differed by less than 0.001 during confirmation, leading to the selection of LR Grid Sweep as the baseline optimization configuration for Experiment 4. Adam and RMSprop performed well during screening, each achieving an F1 score of approximately 0.866. However, their performance declined substantially during confirmation with a longer training schedule, falling to 0.750 and

0.783, respectively. The training F1 scores for both methods remained above 0.82 during confirmation, suggesting a more pronounced generalization gap as training progressed. This pattern may indicate overfitting, although the available results do not establish that the absence of an explicit learning rate schedule was the sole cause. These findings informed the decision to use LR Grid Sweep with cosine annealing in this study.

Experiment 4: Architectural Comparison

Eight sequence-encoder architecture variants, including individual and combined architectures, were evaluated alongside two ablation models to examine the contributions of individual branches. Unlike Experiments 2 and 3, this experiment did not involve separate screening and confirmation stages. Instead, nine combinations of hidden dimension and learning rate were trained directly for each architecture using a derivative of FINAL_CFG with 55 epochs. A total of 72 complete training runs were conducted, and the results are presented in Table 9.

The LSTM+GRU architecture achieved the highest file-level F1 score (0.933333), narrowly outperforming GRU (0.933259). BiLSTM, BiLSTM+LSTM, and BiLSTM+CNN1D each achieved an F1 score of 0.933036. The F1 scores, accuracy, precision, and recall of these three architectures were identical at the reported precision. This equality may reflect identical predictions on the same test files rather than merely similar values resulting from rounding. Given the small test partition of 30 images, several distinct architectural configurations may yield identical final classification decisions despite differences in their internal training processes. The difference between LSTM+GRU and GRU is also practically negligible and cannot be interpreted as a statistically convincing advantage, because all these results were obtained from a single holdout evaluation rather than averaged across multiple independent repetitions.

The strongest results, with F1 scores of approximately 0.933, were achieved by architectures incorporating bidirectional recurrent processing or combinations of recurrent and convolutional components. By contrast, the single unidirectional LSTM and CNN1D architectures achieved lower F1 scores of 0.866518 and 0.882517, respectively. The contribution of the FixMap branch was approximately 3–4 percentage points, as indicated by the comparison between BiLSTM_SequenceOnly, the ablation model without FixMap (F1 = 0.899554), and the complete BiLSTM model (F1 = 0.933036). MLP_Tabular, which used no sequence representation, produced near-chance performance, with an F1 score of 0.471210. These findings suggest that temporal information in the scanpath, supplemented by spatial information from FixMap, provides important discriminative information for the best-performing configuration.

Table 9

Experiment Results 4: Comparison of Sequence Encoder Architecture

No.	Architecture	File F1	Accuracy	Precision	Recall
1.	LSTM+GRU	**0,933333**	**0,933333**	**0,933333**	**0,933333**
2.	GRU	0,933259	0,933333	0,935268	0,933333
3.	BiLSTM	0,933036	0,933333	0,941176	0,933333
4.	BiLSTM+LSTM	0,933036	0,933333	0,941176	0,933333
5.	BiLSTM+CNN1D	0,933036	0,933333	0,941176	0,933333
6.	BiLSTM+GRU	0,899889	0,900000	0,901786	0,900000
7.	BiLSTM_SequenceOnly (ablation, without FixMap)	0,899554	0,900000	0,907240	0,900000
8.	CNN1D	0,882517	0,883333	0,894286	0,883333
9.	LSTM	0,866518	0,866667	0,868304	0,866667
10.	MLP_Tabular (ablation, no sequence)	0,471210	0,500000	0,500000	0,500000

Final Evaluation on Test Partition

The model with the winning configuration, namely the LSTM+GRU architecture using the Eye Tracking modality combination, was trained on the training partition and evaluated on the fixed test partition. Results at three aggregation levels are presented in Table 10.

Table 10
Final Evaluation Results on LSTM+GRU Model with Combination of Eye Tracking

Metrics	File-level	Block-level	Subject-level
Accuracy	0,9333	0,6872	0,6872
Accuracy	0,9333	0,6873	0,6873
Recall	0,9333	0,6874	0,6874
F1 Score	0,9333	0,6871	0,6871
AUC-ROC	0,9733	0,7511	0,7511

At the file level, the final model achieved an F1 score of 0.9333, correctly classifying 56 of the 60 image files in the test partition. Two ASD cases were misclassified as TD (false negatives), and two TD cases were misclassified as ASD (false positives). Thus, the recall for each class was 28/30, or approximately 0.933. Each additional misclassification represents approximately 1.7 percentage points of the 60-file sample. The F1 score of 0.9333 is therefore susceptible to substantial fluctuations because of the small evaluation sample and should not be interpreted as evidence that precision to four decimal places is stable.

A considerable gap was observed between training and test performance. At the file level, training accuracy reached 0.9896, compared with test accuracy of 0.9333. At the block level, the training F1 score was 0.794, compared with a test F1 score of 0.687. These differences suggest overfitting despite the application of dropout, weight decay, and early stopping. The high file-level score may also partly reflect aggregation: each file combines multiple block-level predictions, allowing some errors to offset one another at the file level. The block-level F1 score of 0.687 is therefore more informative about performance on individual observations, defined here as one subject viewing one image, than the file-level F1 score.

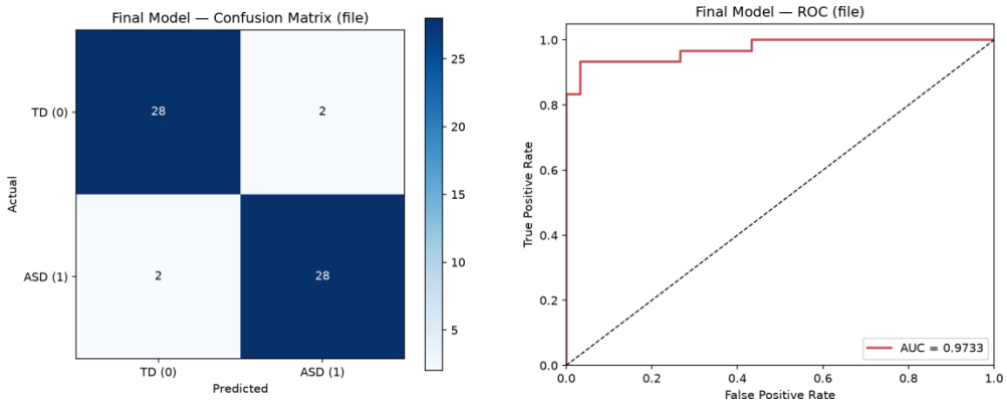


Figure 1. Confusion matrix (left) and ROC curve (right) at the file level

Experiment 5: Classic Machine Learning Baseline

Three classical machine learning models were trained using the same 228 tabular features and evaluated on the same test partition as the proposed model to facilitate comparison. The results are presented in Table 11.

Table 11
Comparison of Multimodal Models with Classic ML Baseline

Models	File F1	Accuracy	Precision
Proposed Multimodal DL (LSTM+GRU, Eye Tracking)	**0,9333**	**0,9333**	**0,9333**
Random Forest	0,8999	0,9000	0,9018
Logistic Regression	0,7818	0,7833	0,7914
kNN (k = 7)	0,6452	0,6500	0,6586

The proposed multimodal deep learning model achieved a file-level F1 score of 0.9333, exceeding the scores of the classical baselines evaluated on the same partition: Random Forest (0.8999), Logistic Regression (0.7818), and kNN (0.6452). These findings suggest that the proposed approach, which learns temporal representations of fixation sequences and spatial representations of visual attention patterns, can provide a competitive alternative to classical approaches based on tabular features. It may also capture complex patterns that are difficult to represent using aggregate statistical features. Random Forest and kNN each achieved a training F1 score of 1.0000, indicating a substantial training–test performance gap and suggesting more pronounced overfitting than in the proposed deep learning model.

However, the three baselines in Experiment 5 were not evaluated under a protocol fully equivalent to that used for the proposed model. The stability test across five different data partitions and the ROC-AUC calculation were performed only for the proposed deep learning model. Consequently, the Experiment 5 comparison should be considered supplementary evidence rather than a definitive test of the proposed model's advantages. The principal focus of this study remains the implementation of the model, the contribution of feature-level fusion, modality combinations, optimization methods, and sequence-encoder architectures.

Stability Test: Robustness Check on Five Different Partitions

The winning configuration, LSTM+GRU with the Eye Tracking modality combination, was retrained and evaluated on five different training–test partitions to supplement the final evaluation. Each partition used a different image-level holdout at the same 80:10:10 ratio.

Table 12

Robustness Check: Winning Configurations on Five Different Partition Seeds

Seed	File F1 (Test)
1	0,8124
2	0,8333
3	0,8490
4	0,8500
5	0,8500
F1 ± SD Average	0, 839 ± 0, *015*
Average Accuracy ± SD	0, 840 ± 0, 013
Average Precision ± SD	0, 848 ± 0, 009

The mean file-level F1 score across the five partitions was 0.839, with a standard deviation of 0.015. This value was substantially lower than the 0.9333 obtained during the final evaluation on a single partition and lay well outside the mean-plus-two-standard-deviations range. These findings suggest that the final evaluation result may reflect a favorable outcome on one particular partition of 30 test images rather than representative generalization performance. Although the five-partition stability test was supplementary and was not used to select a configuration, its mean and variability provide a more cautious indication of performance across the tested partitions than the single result of 0.9333.

Discussion

Contribution of Modality and Architecture

Experiment 2 showed that the Eye Tracking-only configuration, comprising the scanpath sequence and FixMap representations, outperformed the full combination of all four modality groups and other combinations incorporating stimulus imagery and tabular fusion features. This finding is consistent with a study reporting that a scanpath-based ASD classification model can achieve an AUC above 0.90 without additional modalities (Kanhirakadavath & Chandran, 2022). It supports the hypothesis that the principal discriminative information for distinguishing ASD from TD participants in the Saliency4ASD dataset lies in gaze behavior rather than in the visual content of the observed images. This interpretation is consistent with the fact that all subjects viewed identical stimulus images. Similar patterns have been reported in studies showing that visual-attention-estimation pipelines can outperform conventional handcrafted-feature

approaches by focusing on representations of observational behavior (de Belen et al., 2024). In the present study, performance decreased slightly when additional modalities were incorporated into the Eye Tracking-only configuration, contrary to the assumption that adding modalities necessarily improves performance. The small test sample of 30 images may be a contributing factor, as increasing the feature dimensionality without adding sufficient discriminative information can increase the risk of overfitting.

According to Experiment 4, the highest-performing group consisted of architectures with practically negligible performance differences: LSTM+GRU, GRU, BiLSTM, BiLSTM+LSTM, and BiLSTM+CNN1D. The single unidirectional LSTM and CNN1D architectures performed less well, with F1 scores of 0.866518 and 0.882517, respectively. This pattern is consistent with the view that bidirectional contextual information and combined representations can help capture temporal structure in sequential data more effectively than a single recurrent layer (Fan et al., 2024). However, this potential advantage was not tested across multiple independent repetitions in the present study, so its statistical significance cannot be established. The contribution of the FixMap branch was approximately 3–4 percentage points, based on the comparison between the BiLSTM_SequenceOnly ablation model without FixMap (F1 = 0.899554) and the complete BiLSTM model (F1 = 0.933036). MLP_Tabular, which lacked sequence input, achieved only near-chance performance, with an F1 score of 0.471210. These findings suggest that the combination of temporal scanpath representations and spatial FixMap representations provides important information for the winning configuration. This approach is also consistent with the use of ResNet18-based transfer learning for image representations when sample sizes are limited (Alzubaidi et al., 2021).

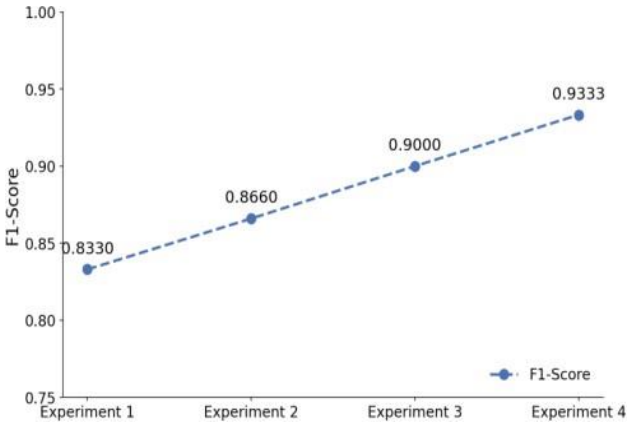


Figure 2. Best file-level F1 progression across Experiments 1–4

The attention mechanism evaluated in this study was limited to the BiLSTM encoder, one of the sequence-encoder variants tested in Experiment 4. Therefore, it should not be generalized as an inherent component of the entire sequence-encoder architecture family. The winning LSTM+GRU configuration does not use an attention mechanism. In this combination, the LSTM encoder is unidirectional and extracts a sequence representation from the final hidden state, denoted by hn[-1]. The GRU encoder is bidirectional and combines the final hidden states from the forward and backward directions through concatenation, represented as torch.cat([hn[-2], hn[-1]]). The resulting representation vectors are then combined into a joint sequence representation before being passed to the fusion head. The LSTM+GRU configuration achieved performance comparable to, and marginally higher than, that of the BiLSTM architecture reported in Table 9. This finding suggests that attention is not the only mechanism through which competitive performance can be achieved in the sequence encoders evaluated in this study. Accordingly, the results do not support a general claim that attention-based architectures are inherently superior.

Comparison with Baseline and Previous Research

In the final evaluation, the proposed model achieved a file-level F1 score of 0.9333 and an ROC-AUC of 0.9733, placing its performance above that of the classical machine learning baselines

evaluated in Experiment 5 on the same partition. Although these values appear high, this study does not claim definitive superiority over the baselines or previous research for three reasons.

First, published performance figures are included solely as contextual evidence because task definitions, input representations, participant composition, and validation procedures differ across studies. For example, the F1 score of 0.61 reported in the Grand Challenge study Mazumdar et al. (2021) and the AUC of 63.9% reported in a scanpath and saliency study Startsev (2019) do not constitute directly comparable benchmarks for the Random Forest F1 score of 0.8999 obtained in the present study. Although the proposed model and the classical baselines were evaluated on the same holdout partition, only the deep learning configuration underwent the reported five-seed stability test. Consequently, neither a cross-study ranking nor a stable advantage over Random Forest can be established from these point estimates alone.

Second, the proposed intermediate feature-level fusion scheme provides evidence of a measurable contribution from spatial information. The BiLSTM_SequenceOnly ablation model, which lacked a FixMap branch, achieved an F1 score of 0.899554, whereas the complete BiLSTM model achieved 0.933036. This difference suggests that integrating spatial representations may improve performance, although the magnitude of the improvement requires confirmation through independent evaluation. Furthermore, the Eye Tracking-only combination outperformed the full combination of four modalities in Experiment 2, indicating that modality selection in the proposed architecture has an empirical basis rather than relying solely on an initial hypothesis. The ROC-AUC of 0.9733 obtained during the final evaluation also lacks a direct baseline comparison because ROC-AUC was not calculated for the three classical machine learning baselines in Experiment 5. Their discriminative performance across the full range of classification thresholds therefore cannot be compared comprehensively.

Third, the stability test showed that the winning configuration's mean F1 score decreased to 0.839, with a standard deviation of 0.015, across five different test partitions. This value is substantially lower than the score of 0.9333 obtained on the single partition used for the main result. This variability should be considered when comparing the proposed model with other studies, including research using visual gaze direction estimated from facial images in a virtual reality environment on the Saliency4ASD dataset, which reported an accuracy of 85.88% H. Chen et al. (2025), and research using involution-based architectures with FixMap and heatmap representations, which reported accuracy above 96% (Mumenin et al., 2024). The present study's single-partition result may be optimistic relative to its performance across alternative partitions; therefore, superiority or inferiority relative to these studies cannot be established from a single evaluation score.

The F1 scores of approximately 0.82–0.85 observed across the five partitions may provide a more cautious indication of performance under the tested partitioning conditions than the single score of 0.9333. This range is broadly comparable to results reported for a hybrid CNN–LSTM architecture applied to gaze and pupil features from an eye-tracking dataset, which achieved a validation accuracy of 87.41%, an F1 score of 0.86, and an AUC of 0.9530 using participant-level data partitioning (Parikh et al., 2026). However, the comparability of the datasets and evaluation protocols cannot be established from the available information; these results should therefore be treated as contextual rather than direct comparisons. Another study reported accuracy above 99% for ASD classification using scanpath visualization images and a combination of U-Net segmentation, Inception v3 feature extraction, and LSTM on eye-tracking datasets associated with Saliency4ASD (Thanarajan et al., 2023). Differences in datasets, sample sizes, and input representations—scanpath visualization images versus raw coordinates and fixation maps in the present study—prevent direct comparison. Nevertheless, these studies demonstrate that LSTM-based architectures can be competitive for eye-tracking-based ASD classification across different data representations.

The image-ID-grouped partitioning strategy reduces overlap in stimulus images between the training and holdout sets. However, it is image-grouped rather than independently subject-grouped validation. In addition, the test partition influenced model and hyperparameter selection, introducing selection bias into the interpretation of Experiments 1–4 and the subsequent baseline comparison. The five additional partitions provide information about sensitivity to partition choice, but they cannot retrospectively convert the selected test set into an untouched evaluation set. Assessing generalization to unseen children and making definitive performance comparisons require identifiable participant groups, preprocessing fitted exclusively on training data, validation-only model selection, and a

genuinely untouched test cohort (White & Power, 2023).

Bias Analysis

The test partition was repeatedly used as a selection reference: in Experiment 2 to select modality combinations, in Experiment 3 to select optimization methods, and in Experiment 4 to select architectures. The same partition was then used to report the final evaluation result. This procedure differs from an ideal locked-test design, in which the test set remains untouched until all model and hyperparameter decisions have been finalized. Although the procedure allows different configurations to be compared on identical data, it also introduces the risk of implicitly overfitting to a particular test partition. The selected configuration at each stage may simply be the one that best matches the characteristics of the 30 test images rather than one that generalizes most effectively. The substantial performance decrease observed when alternative test partitions were used in the stability test provides evidence of sensitivity to partition choice.

Further evidence of overfitting is provided by the gap between training and test performance. At the file level, training accuracy was 0.9896, compared with test accuracy of 0.9333. At the block level, the difference was more pronounced: the training F1 score was 0.794, compared with the test F1 score of 0.687. This gap persisted despite the use of standard regularization techniques, including dropout, weight decay, early stopping, and focal loss.

Metric aggregation also contributes to the higher file-level performance. The file-level F1 score of 0.9333 was substantially higher than the block-level F1 score of 0.6871 for the same model. Aggregating probabilities across blocks within a single image may partly explain this difference because some block-level prediction errors can offset one another when probabilities are aggregated at the image level. This aggregation scheme may be methodologically appropriate if it is consistent with the original dataset's evaluation protocol. However, it does not imply that classification at the individual block level—representing one child's viewing behavior for a single image—achieved performance equivalent to the file-level score of approximately 93%. The block-level F1 score of 0.687 is more relevant when describing performance at the individual-observation level.

The small evaluation sample is an additional source of uncertainty. The test partition comprised 30 images, corresponding to 60 image files when both classes are represented. With such a small sample, each misclassification can change the file-level F1 score by approximately 1.7 percentage points. Differences of several percentage points between models, including the difference between the proposed model and Random Forest, are therefore comparable in magnitude to the effect of one or two classification errors. These differences should be interpreted with appropriate statistical caution.

The potential influence of class imbalance is comparatively limited because the dataset is relatively balanced, with class proportions of approximately 49.5% and 50.5%. The recall values for ASD and TD were also identical in the final evaluation, at approximately 0.933 for each class.

Research Limitations

The principal limitations are the small number of participants (28 children), image-ID-grouped rather than verified child-grouped validation, the selected test partition of only 30 images, repeated use of the test partition during configuration selection, and insufficient detail to audit preprocessing and probability aggregation. The mean file-level F1 score of 0.839 ± 0.015 across the five additional partitions describes performance under the tested partitioning conditions only. It is not a confidence interval for a new population or an independently validated expected performance range of 0.82–0.85. Furthermore, the classical baselines were not evaluated under the same repeated-partition protocol, preventing robust claims of superiority or equivalence across independent samples. Future research should use child-level identifiers, an untouched external test set, and a fully specified preprocessing and aggregation pipeline.

Table 13
Summary of Results, Theoretical Foundations, and Comparative Context

Experiments	Theoretical Foundations	Explanation	Compared to
Eye Tracking (sequence+FixMap) only	Scanpath as a key biomarker	The content of the stimulus is identical between the subjects so that it is not discriminatory	15 Combinations of modalities
Two-way/combination architecture top-notch	Bidirectional context representation	The difference between the top architectures is practically negligible	10 architectural variants
The proposed DL is comparable to slightly higher than the classic Machine Learning baseline	Representations are learned directly from raw data	Not proof of absolute superiority	3 basline *Machine Learning*
Singles test 0.933 drops to average 5 seed 0.839	Small holdout sensitivity	Test partition 30 images prone to fluctuations	5 different partitions
File 0.933 vs. block 0.687	Aggregation effect	Not equivalent accuracy at the individual level	Different aggregation levels

CONCLUSION

This study examined a multimodal hybrid deep learning pipeline for classifying ASD versus TD participants using eye-tracking data from Saliency4ASD. In exploratory comparisons, the scanpath-plus-fixation-map configuration outperformed the full four-representation combination, achieving a file-level F1 score of 0.866 compared with 0.833. The selected LSTM+GRU model achieved a file-level F1 score of 0.9333 and an ROC-AUC of 0.9733 on the original 30-image holdout partition. These findings suggest that combining complementary temporal and spatial gaze representations can be useful, whereas incorporating all available representations does not necessarily improve classification performance.

Across five additional image-grouped holdout partitions, the selected model achieved a mean file-level F1 score of 0.839 ± 0.015, indicating sensitivity to partition choice. The Random Forest baseline achieved an F1 score of 0.8999 on the original partition but was not evaluated under the same repeated-partition protocol. Therefore, consistent superiority over the classical baseline cannot be established. Reuse of the test partition during model selection, uncertainty concerning preprocessing implementation, and the absence of independently verified child-level validation limit the generalizability of the findings. Future research should use participant-independent data partitions, a strictly untouched test cohort, and transparent preprocessing protocols before the model's clinical applicability can be assessed.

ACKNOWLEDGMENT

The authors acknowledge the creators and contributors of the publicly available Saliency4ASD dataset, which provided the eye-tracking data analyzed in this study.

AUTHOR CONTRIBUTIONS

Citra Wanodya Prameswari contributed to conceptualization, methodology, software implementation, experimentation, data analysis, visualization, and writing the original draft. Erwin Budi Setiawan contributed to conceptualization, methodological guidance, supervision, interpretation of the results, and editing. Both authors reviewed and approved the final manuscript.

REFERENCES

Alzubaidi, L., Al-Amidie, M., Al-Asadi, A., Humaidi, A. J., Al-Shamma, O., Fadhel, M. A., Zhang, J., Santamaría, J., & Duan, Y. (2021). Novel transfer learning approach for medical imaging with limited labeled data. *Cancers*, *13*(7), 1590.

https://doi.org/10.3390/cancers13071590

Chen, H., Zhang, X., Chen, Z., Ren, Y., & Liu, R. (2025). Auxiliary diagnostic method for children with autism spectrum disorder based on virtual reality and eye-tracking technology. *Scientific Reports*, *15*(1), 40552. https://doi.org/10.1038/s41598-025-24243-w

Chen, J., Liao, M., Wang, G., & Chen, C. (2020). An intelligent multimodal framework for identifying children with autism spectrum disorder. *International Journal of Applied Mathematics and Computer Science*, *30*(3), 435–448. https://doi.org/10.34768/amcs-2020-0032

Comparan-Meza, M., Vargas de la Cruz, I., Jauregui-Huerta, F., Gonzalez-Castañeda, R. E., Gonzalez-Perez, O., & Galvez-Contreras, A. Y. (2021). Biopsychological correlates of repetitive and restricted behaviors in autism spectrum disorders. Brain and behavior, 11(10), e2341. https://doi.org/10.1002/brb3.2341

de Belen, R. A. J., Eapen, V., Bednarz, T., & Sowmya, A. (2024). Using visual attention estimation on videos for automated prediction of autism spectrum disorder and symptom severity in preschool children. *PLoS ONE*, *19*(2), e0282818. https://doi.org/10.1371/journal.pone.0282818

Duan, H., Zhai, G., Min, X., Zhu, Y., Fang, Y., & Yang, X. (2019). A dataset of eye movements for the children with autism spectrum disorder. Dalam *Proceedings of the 10th ACM Multimedia Systems Conference (MMSys'19)* (255–260). https://doi.org/10.1145/3304109.3325818

Fan, Y., Tang, Q., Guo, Y., & Wei, Y. (2024). BiLSTM-MLAM: A multi-scale time series prediction model for sensor data based on Bi-LSTM and local attention mechanisms. *Sensors*, *24*(12), 3962. https://doi.org/10.3390/s24123962

Han, J., Jiang, G., Ouyang, G., & Li, X. (2022). A multimodal approach for identifying autism spectrum disorders in children. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *30*, 2003–2016. https://doi.org/10.1109/TNSRE.2022.3192431

Kanhirakadavath, M. R., & Chandran, M. S. M. (2022). Investigation of eye-tracking scan path as a biomarker for autism screening using machine learning algorithms. *Diagnostics*, *12*(2), 518. https://doi.org/10.3390/diagnostics12020518

Mazumdar, P., Arru, G., & Battisti, F. (2021). Early detection of children with autism spectrum disorder based on visual exploration of images. *Signal Processing: Image Communication*, *94*, 116184. https://doi.org/10.1016/j.image.2021.116184

Mumenin, N., Yousuf, M. A., Nashiry, M. A., Azad, A. K. M., Alyami, S. A., Liò, P., & Moni, M. A. (2024). ASDNet: A robust involution-based architecture for diagnosis of autism spectrum disorder using eye-tracking. *IET Computer Vision*, *18*(5), 666–681. https://doi.org/10.1049/cvi2.12271

Nurhasan, A. A., & Kasprowski, P. (2026). Application of eye movement analysis in medicine: A review across neurodevelopmental, neurological, and neurodegenerative disorders. *Applied Sciences*, *16*(5), 2548. https://doi.org/10.3390/app16052548

Parikh, J., Goyal, S., Malhotra, Y., & Leong, W. Y. (2026). A spatiotemporal hybrid neural network for robust autism spectrum disorder detection via eye tracking feature fusion. *Discover Artificial Intelligence*, *6*, 495. https://doi.org/10.1007/s44163-026-01458-y

Rakotomanana, H., & Rouhafzay, G. (2025). A scoping review of AI-based approaches for detecting autism traits using voice and behavioral data. *Bioengineering*, *12*(11), 1136. https://doi.org/10.3390/bioengineering12111136

Salari, N., Rasoulpoor, S., Rasoulpoor, S., Shohaimi, S., Jafarpour, S., Abdoli, N., Khaledi-Paveh, B., & Mohammadi, M. (2022). The global prevalence of autism spectrum disorder: A comprehensive systematic review and meta-analysis. *Italian Journal of Pediatrics*, *48*, 112. https://doi.org/10.1186/s13052-022-01310-w

Simon, J., Kapileswar, N., Datchinamoorthi, M., Muthukumar, S., & Keerthana Devi, G. (2024). Gaze-assisted autism spectrum disorder identification: A fusion of machine learning and deep learning approaches for preemptive identification. Dalam *Proceedings of the 5th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)* (525–529). https://doi.org/10.1109/ICMCSI61536.2024.00082

Startsev, M., & Dorr, M. (2019). Classifying autism spectrum disorder based on scanpaths and saliency. Dalam *Proceedings of the IEEE International Conference on Multimedia & Expo Workshops (ICMEW)* (633–636). https://doi.org/10.1109/ICMEW.2019.00122

Thanarajan, T., Alotaibi, Y., Rajendran, S., & Nagappan, K. (2023). Eye-tracking based autism spectrum disorder diagnosis using chaotic butterfly optimization with deep learning model. *Computers, Materials & Continua*, *76*(2), 1995–2013. https://doi.org/10.32604/cmc.2023.039644

Tsai, C. H., Chen, K. L., Li, H. J., Chen, K. H., Hsu, C. W., Lu, C. H., ... & Huang, C. Y. (2020). The symptoms of autism including social communication deficits and repetitive and restricted behaviors are associated with different emotional and behavioral problems. Scientific reports, 10(1), 20509. https://doi.org/10.1038/s41598-020-76292-y

Wang, L., Wang, B., Wu, C., Wang, J., & Sun, M. (2023). Autism spectrum disorder: neurodevelopmental risk factors, biological mechanism, and precision therapy. International journal of molecular sciences, 24(3), 1819. https://doi.org/10.3390/ijms24031819

White, J., & Power, S. D. (2023). k-fold cross-validation can significantly over-estimate true classification accuracy in common EEG-based passive BCI experimental designs: An empirical investigation. *Sensors*, *23*(13), 6077. https://doi.org/10.3390/s23136077

Zhu, F.-L., Wang, S.-H., Liu, W.-B., Zhu, H.-L., Li, M., & Zou, X.-B. (2023). A multimodal machine learning system in early screening for toddlers with autism spectrum disorders based on the response to name. *Frontiers in Psychiatry*, *14*, 1039293. https://doi.org/10.3389/fpsyt.2023.1039293