Vietnam National University, Hanoi
International School

==========



M.A Thesis

Early Detection Autism Spectrum Disorder In Children Using Machine Learning Techniques

NGUYEN DINH THANH

Field: **Master of Informatics and Computer Engineering**

Code: **8480111.01QTD**

Hanoi - 2026

Vietnam National University, Hanoi
International School

==========



M.A Thesis

Early Detection Autism Spectrum Disorder In Children Using Machine Learning Techniques

NGUYEN DINH THANH

Field: **Master of Informatics and Computer Engineering**

Code: **8480111.01QTD**

Supervisor: **Dr. Ha Manh Hung**

Hanoi - 2026

CERTIFICATE OF ORIGINALITY

I, the undersigned, hereby certify my authority of the study project report entitled *"Early Detection Autism Spectrum Disorder In Children Using Machine Learning Techniques"* submitted in partial fulfillment of the requirements for the degree of Master Informatics and Computer Engineering. Except where the reference is indicated, no other person's work has been used without due acknowledgement in the text of the thesis.

Hanoi, June 12, 2026

Nguyen Dinh Thanh

Approved by

SUPERVISOR

Dr. Ha Manh Hung

Date: June 12, 2026

ACKNOWLEDGEMENTS

First of all, I would like to express my thanks to my advisor, PhD. Ha Manh Hung for his direction, valuable feedbacks and constant support in this research.

Secondly, I appreciate the contribution from my teachers and colleagues at International School, Vietnam National University who offered valuable opinions and advice in improving my research. In addition, I want to express my sincere gratitude to my family and friends for always supporting me.

Lastly, I would like to express my appreciations to all people and organizations who contributed to my research project.

Hanoi, June 12, 2026

Nguyen Dinh Thanh

ABSTRACT

Autism Spectrum Disorder (ASD) is a multifaceted neurodevelopmental disorder characterized by difficulties related to socialization, communication, and behavior with different degrees of severity. Early detection of autism is crucial for proper diagnosis and enhanced developmental progress. However, existing diagnostic procedures focus much on behavioral analysis that is both subjective and time-consuming and requires highly trained clinicians. In recent times, researchers have shown keen interest in computerized methods involving eye-tracking technologies and machine learning algorithms that make the detection process quantitative. Eye-tracking provides information about the different types of attention mechanisms that are quite distinct in ASD cases compared to those of TD individuals.

This paper explores the development of a machine learning algorithm for autism spectrum disorder (ASD) diagnosis based on eye tracking in combination with computational saliency analysis and object/face detection. Specifically, the Saliency4ASD dataset was employed in this research. This dataset includes the fixation and scanpath information obtained from subjects while they were looking at images across various social and non-social categories. Based on this eye gaze data, several visual exploration measures such as fixation-based measures, saliency map descriptors, and regional attention measures were computed for each subject. In terms of supervised learning models, RF, XGB, SVC, Adaboost, Catboost, and MLP were compared.

The meta-learner ensemble classifier using Naive Bayes was found to yield the best accuracy with 71.7% and F1-score with 70.6%. For the single classifiers, the accuracies attained by XGBoost and Random Forest were 70.1% and 68.6%, respectively, which indicates that tree-based algorithms function well with detecting non-linear dependency. The soft-voting ensemble generated a fairly balanced performance score of 64.6% precision and 84.8% recall, while SVC and CatBoost achieved high levels of recall score with 88.4% and 86.4%, respectively.

The MLP classifier had the recall rate of 86.4% and F2-score of 80.0%, which

is quite an impressive result for detecting gaze behavior typical of individuals with ASD despite the lower accuracy of 66.9%.

The feature importance measure determined that fixation time, average saliency of fixated points, and percentage of fixations on face-related objects are the most important features used in classification. These findings are in line with previous research, which showed that typical adults spend more time fixing their gaze on social stimuli, such as faces, while persons with ASD exhibit a greater dispersion of fixation times and decreased focus on centrally located or socially relevant stimuli. Thus, the use of eye-tracker data and visual features is effective for recognizing ASD gaze abnormalities.

Keywords: Autism Spectrum Disorder, Eye-Tracking, Machine Learning, Saliency Map, Ensemble Learning

TABLE OF CONTENTS

CERTIFICATE OF ORIGINALITY i

ACKNOWLEDGEMENTS ii

ABSTRACT iii

TABLE OF CONTENTS v

LIST OF FIGURES viii

LIST OF TABLES ix

CHAPTER 1 INTRODUCTION 1

1.1 Aims and Objectives 4

1.2 Contributions 5

1.3 Thesis Structure 6

CHAPTER 2 BACKGROUND AND RELATED WORK 7

2.1 Autism Spectrum Disorder 7

2.2 Eye-Tracking Technology 8

2.2.1 Fixations 9

2.2.2 Saccades 9

2.2.3 Scanpaths 9

2.2.4 Saliency Maps 10

2.3 Eye-Tracking for Autism Research 10

2.4 Autism Detection Using Machine Learning 11

2.4.1 Traditional Machine Learning Approaches 11

2.4.2 Deep Learning Approaches 12

2.4.3 Comparative Analysis of Existing Studies 12

2.5 Evaluation Metrics 13

2.5.1 Accuracy 13
2.5.2 Precision 14
2.5.3 Recall 14
2.5.4 F1-Score 14
2.5.5 ROC-AUC 14
2.6 Research Gap and Motivation 14
CHAPTER 3 METHODOLOGY 16
3.1 Dataset 16
3.2 Preprocessing and Feature Engineering 18
3.3 Base Models and Training Procedure 19
3.3.1 Implementation Environment and Pipeline Architecture 20
3.3.2 Participant-Level Data Splitting Strategy 20
3.3.4 Detailed Implementation of the Base Models 21
3.3.5 Probability Calibration and Threshold Optimization 22
3.4 Ensemble Methods 23
3.5 Deep Neural Networks 24
3.5.1 Multi-Layer Perceptron Architecture 25
3.5.2 Training Strategy for Small-Scale Data 25
CHAPTER 4 EXPERIMENTAL RESULTS 27
4.1 Performance Analysis 27
4.1.1 Performance of Tree-Based Models 27
4.1.2 Performance of Margin-Based and Boosting Models 29
4.1.3 Performance of Ensemble Models 30
4.1.4 Performance of the Deep Learning Model 32

4.1.5 Comparative Discussion 34
4.2 Relationship Between Visual Features and ASD-Specific Viewing Behavior...35
4.3 Implications for Clinical Screening and Early ASD Detection 35
CHAPTER 5 CONCLUSIONS AND FUTURE WORK.....37
REFERENCES 39

LIST OF FIGURES

Figure 1.1: Reported prevalence of ASD in the US from 2000 to 2022.....1
Figure 3.1: Overview of the proposed multimodal ASD detection framework.....17
Figure 4.1: Random Forest performance analysis. 28
Figure 4.2: XGBoost performance analysis.....28
Figure 4.3: Support Vector Classifier (SVC) performance analysis. 29
Figure 4.4: AdaBoost performance analysis.....30
Figure 4.5: CatBoost performance analysis.....30
Figure 4.6: Stacking ensemble with logistic regression meta-learner. 31
Figure 4.7: Stacking ensemble with Naive Bayes meta-learner. 31
Figure 4.8: Stacking ensemble with KNN meta-learner.....32
Figure 4.9: Soft voting ensemble performance analysis.....32
Figure 4.10: MLP performance analysis.....33
Figure 4.11: MLP training loss over epochs.....33

LIST OF TABLES

Table 2.1:Comparison of ASD Classification. 12
Table 4.1: Performance comparison of evaluated models. 27

CHAPTER 1 INTRODUCTION

Autism Spectrum Disorder is a multifaceted, heterogeneous neurodevelopmental disorder characterized by differences in social interaction, communication, and behavior to various extents of severity. The spectrum of the disorder denotes the heterogeneity among the manifestations of the disease in patients. It should be noted that according to the latest statistics provided by the U.S. Centers for Disease Control and Prevention, around 1 in 31 or about 3.2% of eight-year-old children have been diagnosed with ASD in 2022. There is evidence that there has been a steady increase in the percentage of children diagnosed with ASD since the first available statistic from the CDC which indicated 1 in 150 children in 2000; the statistics in subsequent years were 1 in 59 in 2014 and 1 in 54 in 2016, as represented in Fig. 1.1.

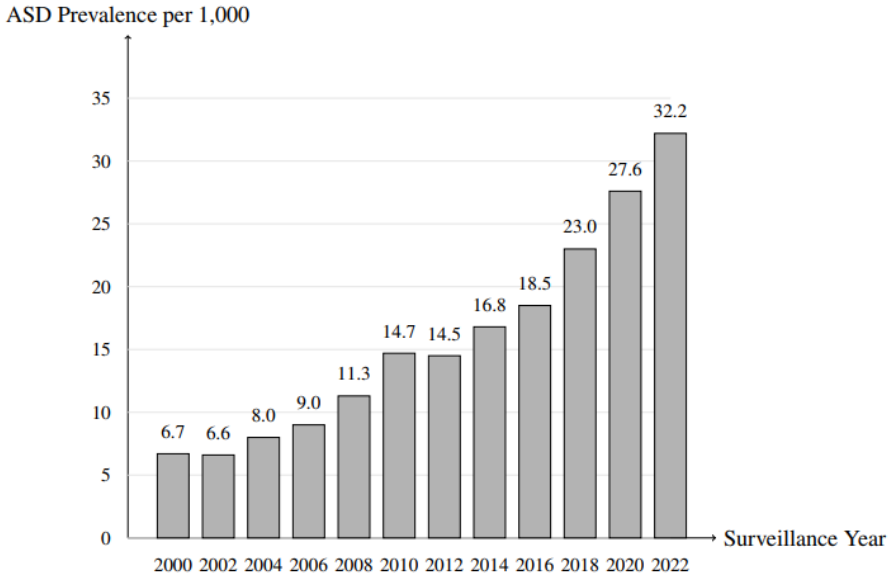


Figure 1.1: Reported prevalence of ASD in the US from 2000 to 2022.

In the last two decades, many approaches have been studied by researchers in order to help detect ASD and understand the reasons behind it. While the exact cause of ASD is not known for the majority of patients, recent studies have shown that genetics and the environment play a major role. Twin studies have found high heritability, meaning that genetic factors are very important in the formation of ASD.

On the other hand, there is no certainty regarding the environmental factors that are involved.

Diagnostic criteria for ASD have proved elusive and hard to establish. Currently, there is no medical or lab test that can provide a definitive diagnosis for this condition. Diagnosis of autism depends on specific tests aimed at evaluating behavior. In particular, the CDC recommends that diagnostic tools be used in combination since they tend to produce inaccurate results when applied separately. Examples of popular diagnostic tools include the Autism Diagnostic Interview–Revised (ADI-R), which examines communication and social skills across all developmental stages; Autism Diagnostic Observation Schedule– Generic (ADOS-G), which examines social and play behavior; Childhood Autism Rating Scale (CARS), which offers a quick assessment of characteristic features; and Gilliam Autism Rating Scale–Second Edition (GARS-2), which allows clinicians, educators, and parents to evaluate behaviors and symptoms. According to the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) published by the American Psychiatric Association, the primary diagnostic criteria focus on social communication problems and repetitive/restrictive behaviors.

Due to the complicated nature of symptoms associated with autism spectrum disorders (ASD), the issue of diagnosis is still a problem, especially when considering adults and people who suffer from less severe forms of the condition. One possible way to solve this problem is to use various technologies such as eye-tracking technology, electroencephalography (EEG), and functional magnetic resonance imaging (fMRI). Eye-tracking is a non-invasive approach, which allows to record a person's gaze trajectory. EEG involves recording electrical activity on the surface of the head, whereas fMRI involves measuring differences in brain activity through changes in blood oxygenation level. Among all these approaches, the one that stands out because of its advantages is eye-tracking because:

The eye tracking data can be viewed as structured records, which consist of numerical and categorical values like the positions, duration, and transitions of

fixation points. In turn, the EEG and fMRI methods result in large-scale multivariate time-series data that have to go through significant pre-processing in order to derive useful features. It is the structuredness and the high temporal resolution of the eye tracking data that have made the method extremely appropriate for statistical analyses.

The machine learning framework introduced in this thesis constitutes an exhaustive system, which utilizes different types of visual and behavioral input data for early and objective diagnosis of autism spectrum disorders (ASD). In particular, this framework utilizes a multimodal system of feature extraction using measures related to fixation-based eye tracking, saliency maps, and visual information extracted from images in relation to objects and faces. Eye tracking information provides temporal analysis in terms of gaze behaviors including the duration, dispersion, and transitions between fixations; saliency maps help identify the attention driven by stimuli and regions that naturally attract visual attention; and image-derived information can be used to provide context information about socially and non-socially salient stimuli.

As individuals suffering from autism spectrum disorder find it challenging to interact through social communication or do not focus on certain stimulus in the environment, most of the research conducted to study autism spectrum disorder using the eye-tracking approach makes use of visual stimuli, which are mostly comprised of static images or videos with social stimuli. Visual stimuli make it easy to identify any difference in the way a person with autism views or focuses on certain visual stimuli as these individuals generally have low fixations on socially relevant features, including faces and eyes, and instead have preferences for other things present around the stimuli.

Aside from the eye-tracking data, that clearly indicates the visual focus of the participant, information derived from computational saliency maps and image-level descriptors may prove helpful for providing a more complete picture. The saliency maps indicate what parts of the image are likely to capture a person's visual interest,

making it possible to compare the prediction of the visual focus with the actual gaze trajectory of the subject. Object and face detection characteristics extracted from the image similarly allow one to understand what parts of the scene are socially relevant.

The current study thus concentrates on integrating multiple visual sources, including data from eye tracking, saliency maps, and images in analyzing attention patterns among persons with autism spectrum disorder. This type of approach helps capture both the way people pay attention to various types of stimuli and the connection between their gaze pattern and the visual characteristics of stimuli. This approach is also ecologically valid in researching visual attention abnormalities among autistic persons.

1.1Aims and Objectives

The purpose of this research is to explore the problem of ASD recognition through the use of multi-modal systems consisting of eye-tracking data, computational saliency maps, and visual features from images. The general aim is to construct a model based on data analysis that would allow for characterization of unusual visual behavior in individuals with autism spectrum disorder. Three specific research goals include:

• The first research goal is to acquire a more profound knowledge about differences between the way individuals with ASD and typically developing people perceive visual stimuli. This objective can be achieved through the combination of fixation metrics with information about salient and non-salient parts of images as well as by utilizing both salience-based and image-level data. Given that effective data pre-processing is critical for the success of predictive modeling, special attention will be given to the development of a multi-modal approach that would maximize predictive power.

• The second research goal concerns investigation of the most efficient ways to analyze multi-modal data on visual attention. Traditional statistical prediction techniques as well as artificial neural network approaches will be tested as possible tools for analyzing multi-modal visual attention data. All developed models will be

assessed using the same performance metrics as well as following evaluation protocols used in other relevant studies.

• The third goal is to assess the efficiency and reliability of the presented approach during the analysis of various types of visual stimuli in the current dataset for Saliency4ASD problematics. As a result of the small number of pictures and participants, the research will be concentrated on exploring the influence of multi-modal features extracted from eye tracking results, computational saliency maps, and images' features on classification results. The main purpose of this step is to understand the working principle of the model as well as prepare for future learning approaches with less data.

1.2 Contributions

The major contribution of this study is the development and evaluation of an applied multimodal framework for ASD identification, integrating eye-tracker data, computational saliency maps, and object/face detection features through a systematic experimental comparison. This research combines all three types of distinct features along with the metrics of gaze behavior such as duration of fixations, pattern of saccades, average saliency of the point of fixation, and duration of fixations within face AOIs.

The current research entails extensive testing of several machine learning models and algorithms along with the use of some ensemble approaches. The traditional classification models of the Random Forest, XGBoost, and AdaBoost were considered alongside SVM, MLP, and stacking and voting ensemble techniques. It was observed that the use of the ensembling methods yielded satisfactory performance in terms of accuracy, precision, recall, and F-scores. The use of SVM and MLP yielded higher recall scores for the detection task. In order to increase the sensitivity of the classification process, calibration of probability values was carried out.

Results from the analysis of most discriminative features for visual attention showed consistent differences between ASD and TD subjects. For TD subjects, it was

seen that they spent longer times on fixations concentrated around central regions and socially relevant areas. On the other hand, for ASD subjects, the fixation times were less focused, and there was an increase in attending to peripheral or non-social regions. These findings support the significance of behavioral indicators using gaze or images, which help us understand the importance of using multimodal data.

Moreover, the work highlights the interpretability and prediction accuracy of the framework introduced. Feature Importance analysis revealed that the most important gaze and image-based feature set included measures such as average saliency at fixation locations and fixation duration within face-related AOIs when trying to classify participants with ASD versus those with TD. The results obtained from the comparison among different classifiers can help guide further studies in this field while developing an early-screening aid for autism spectrum disorders using artificial intelligence.

1.3 Thesis Structure

The second chapter of the paper focuses on the background and related work in eye tracking and application areas where this methodology is employed. The use of machine learning and deep learning algorithms for detecting ASDs in different forms of data are discussed in detail, such as gaze data, saliency features, and image features. The significance of saliency and image features is highlighted while making predictions about ASD.

Chapter 3, Methodology, describes the materials, apparatus, and procedures used to collect eye-tracking data. It presents the methodology developed in this study for ASD detection through machine learning models and statistical analysis of gaze and image-based features.

Chapter 4, Results, reports the differences in visual attention patterns between ASD and TD participants. It evaluates and compares the performance of multiple machine learning models applied to the multi-modal gaze and image data.

Chapter 5, Discussion, interprets the findings of the study, emphasizes the significance of using machine learning for ASD detection, and discusses the limitations of the current research as well as potential directions for future work

CHAPTER 2 BACKGROUND AND RELATED WORK

The theoretical foundation and literature review on the early detection of Autism Spectrum Disorder (ASD) using machine learning methods will be outlined in this chapter. Initially, the nature and features of ASD along with the traditional way of detecting ASD will be explained. Secondly, eye-tracking technology and its use in autism studies will be explained. Subsequently, a review of the literature on ASD detection studies using eye-tracking information along with machine learning will be outlined. Moreover, publicly available ASD datasets and evaluation methods will be identified. Lastly, current limitations in ASD studies will be discussed.

2.1 Autism Spectrum Disorder

Autism Spectrum Disorder is a condition that affects a person's social interaction and communication and involves restricted repetitive behavior. The use of the term "spectrum" indicates that there is a lot of variability in how autism presents in individuals. Some children have problems with cognition and language acquisition. Other individuals show normal or even superior intellectual skills but still struggle with socialization and behavioral issues.

ASD appears in early childhood and impacts a wide range of developmental domains, such as social interactions, language acquisition, communication skills, learning capabilities, and behavioral functioning. The typical symptoms of the disorder include lack of eye contact, delay in language development, poor responsiveness to social interaction, repetitive behaviors, insistence on routine, and hyper- or hyposensitivity to sensory inputs. As a result of the described variability in the manifestation of autism, the disorder is considered one of the most heterogeneous neurodevelopmental conditions.

It is also essential due to the fact that any intervention performed during sensitive developmental stages may significantly enhance the results for an individual affected by autism. It was found out that children whose behavioral and educational problems were detected early could reach better achievements concerning such skills as communication, socialization, and adaptation than other patients. Therefore, much

attention has been paid to creating effective techniques for early diagnosis of ASD.

Conventional ASD diagnosis involves mainly observational studies, in addition to using standardized tests. Popular methods used to diagnose autism include the Autism Diagnostic Observation Schedule (ADOS), the Childhood Autism Rating Scale (CARS), the Autism Diagnostic Interview-Revised (ADI-R), and the Vineland Adaptive Behavior Scales (VABS). While such diagnostic measures have been shown to be effective in clinical settings, they are generally known to involve extensive training and evaluation time. Additionally, diagnoses made using such approaches tend to be highly subjective in nature, thereby delaying diagnoses.

In order to mitigate the weaknesses of these diagnostic techniques, there has been increasing interest in the use of biomarkers to aid in the objective measurement of behaviors related to autism. Among other measures, one measure that has gained popularity is eye-tracking technology based on the differences in attention behaviors exhibited by TD children and ASD children.

2.2 Eye-Tracking Technology

Eye tracking is a technique that captures and analyses the movement of eyes so that it is possible to determine where the person is looking at and how his or her visual attention was directed throughout the process. The modern methods of eye tracking include infrared lighting and fast video cameras that can capture gaze points with both great spatial and temporal accuracy. Since the movements of our eyes have much in common with brain activity, eye tracking became very popular in psychology and other spheres.

In general, the goal of eye tracking is the analysis of eye movements when observing visual stimuli. Individuals observe visual material such as photos, videos, web sites or social interaction situations, and at the same time their eye movements are recorded. The result is a dataset containing a lot of data concerning visual attention and cognitive processing.

There are some basic features that can be found in the recordings of eye movements. They include fixations, saccades, scanpaths, and blinks

2.2.1 Fixations

The fixation is described as a phase during which the focus of the observer's eyes stays fairly fixed on one point. During fixation, there is a gathering and processing of visual information. Examples of variables associated with the fixation include duration, frequency, and spatial distribution of fixation.

It is important for the analysis of fixation behavior for the purpose of research about ASD due to its ability to identify preferred visual regions. Earlier research has found that children with ASD demonstrate less interest towards the socially-relevant areas like face or eyes compared to non-social areas like objects and backgrounds.

2.2.2 Saccades

Saccades are fast eye movements which arise while changing gaze from one spot to another spot. As opposed to fixations, visual information is not extracted efficiently while making saccades because of the quickness of eye movements. Saccadic characteristics include amplitude, latency, velocity, and transition rate.

Saccadic behavior can give interesting information about how individuals explore scenes and manage their attentional focus. There are reports that saccadic behaviors of children with ASD are significantly different from those of neuro-typical children.

2.2.3 Scanpaths

Scanpath is described as the order of fixation and saccades during visual exploration. Unlike standalone gaze metrics, scanpaths include both temporal and spatial aspects of visual attention. Therefore, by analyzing scanpaths, it becomes possible to study the dynamics of changes in visual attention between various areas of a visual scene.

Scanpaths have already been used in autism studies to show disparities in gaze strategies among ASD and TD children. Indeed, children with ASD tend to display less structured and less social scanpaths when observing scenes that include other people.

2.2.4 Saliency Maps

A saliency map is an image that depicts attention distribution obtained through the analysis of eye-tracking results. It shows the probability of each image area being attended to visually. The process of generating a saliency map generally involves collecting gaze fixations from several subjects and using spatial smoothing algorithms.

The use of saliency maps makes it easier to compare the visual attention patterns of subjects with ASD and TD participants since differences in saliency maps often demonstrate decreased attention to socially relevant image areas in the case of ASD children. Another advantage of saliency maps is that they can be considered as visual input in ML and DL applications.

2.3 Eye-Tracking for Autism Research

Eye tracking has become one of the most commonly employed tools for studying gaze abnormalities linked to Autism Spectrum Disorder. Several studies have confirmed that ASD patients tend to show unusual patterns of looking at socially significant stimuli, including faces, social situations, biological motion, and naturalistic pictures. At the same time, many of the peculiarities linked to eye-tracking of ASD sufferers appear early enough to be used for early autism detection.

The fact that autistic individuals pay little attention to faces and eyes is one of the most common findings in autism research. Typically, healthy kids automatically pay attention to socially salient cues, including facial features and eye contact. In contrast, ASD patients tend to look less at such areas and more at various objects or other salient, yet nonsocial parts of visual scenes.

Numerous studies have used various types of experiments in an effort to study the properties of gaze behavior in individuals suffering from ASD. The list includes free-viewing tasks, facial recognition experiments, visual search tasks, observation of social interaction scenarios, and observation of natural scenes. Although these techniques differ in terms of implementation, most of the results produced by different studies appear similar when referring to gaze characteristics in people with

ASD.

The fact that eye tracking is objective and quantifiable renders eye movement data well suited for use in machine learning tasks. The computer can identify patterns related to autism based on analysis of gaze features.

2.4 Autism Detection Using Machine Learning

The growing prevalence of eye tracking information has opened up new possibilities for the automatic generation of screening tools for autism spectrum disorder (ASD). The conventional methods of diagnosis depend on the observation of behavior and involve a certain level of subjectivity that takes quite a bit of time and effort to accomplish. Machine learning presents a new option by leveraging computer models to recognize patterns in eye movements and make the distinction between ASD patients and typically developed (TD) individuals.

Machine learning techniques train their models using labeled input to generate decision boundaries that allow the classification of new samples according to what was learned. In terms of ASD detection, the input variables might range from fixation durations and counts, saccade features, gaze transitions, and other scanpath characteristics to metrics associated with the concept of saliency.

2.4.1 Traditional Machine Learning Approaches

Several conventional machine learning techniques have been considered for the task of autism spectrum disorder (ASD) classification. Out of these, SVM is one of the most popular choices as a result of its efficiency in dealing with highly dimensional data as well as the capability of constructing strong decision boundaries. Previous works show the effectiveness of applying SVM in cases where gaze related features extracted from eye tracking videos are employed.

Another algorithm that has gained popularity among researchers interested in the field of autism is the Random Forest (RF). Being an ensemble machine learning technique, RF utilizes the output of multiple decision tree classifiers in order to make a final prediction. Furthermore, this algorithm offers measures of feature importances, which helps researchers identify the gaze features contributing to ASD

detection.

Commonly employed techniques include, amongst others, Logistic Regression, K-Nearest Neighbors (KNN), Decision Trees, AdaBoost, and Gradient Boosting algorithms.

2.4.2 Deep Learning Approaches

The recent developments in the area of deep learning techniques have increased opportunities for automatic ASD diagnosis as well. Unlike the machine learning approaches which use feature extraction as an initial step for building classification models, deep learning techniques do not need feature extraction because they automatically learn hierarchical representation from the data.

CNNs have found wide application in the analysis of saliency maps, fixation maps, and heatmaps of gaze data. The capability of CNN-based models to automatically detect spatial patterns that differentiate between ASD and TD individuals allows them to be used in representing image-based gaze data.

The use of Recurrent Neural Networks (RNN) and Long Short Term Memory (LSTM) networks in ASD detection is also well reported. LSTM networks are designed for modeling temporality, which makes it well suited for scanpath analysis. The analysis of scanpaths using LSTM networks allows one to gain some insight into temporal patterns in visual attention.

Hybrid architectures that combine both CNN and LSTM parts have been proposed to leverage information from space and time simultaneously. While the hybrid network structure usually outperforms single-part networks in terms of accuracy, it requires significantly more data and computational resources.

2.4.3 Comparative Analysis of Existing Studies

Table 2.1:Comparison of ASD Classification.

Reference	Dataset	Year	Features	Model	Accuracy	Recall
Startsev & Dorr [34]	Saliency4 ASD	2019	Scanpath, saliency, image content	Random Forest	59.8%	71.7%

Arru et al. [35]	Saliency4 ASD	2019	Object, saliency, fixation	TreeBagger	59.3%	68.4%
Tao et al. [36]	Saliency4 ASD	2019	Scanpath, saliency	CNN–LSTM	57.9%	59.2%
Wu et al. [37]	Saliency4 ASD	2019	Image features, synthetic scanpaths	FCN	55.1%	63.5%
Xu et al. [38]	Saliency4 ASD	2021	Gaze deviation, fixation duration	Rule-based	51.6%	70.5%

Despite a large number of works having demonstrated positive results in terms of ASD detection by means of eye-tracking data, a direct comparison between various works is not easy since there exist many differences in terms of dataset usage, experiments, features, and evaluation techniques. Classification accuracy values reported vary greatly among the various works, from moderate up to very high levels.

What can be stated regarding the majority of works published in this topic area is the possibility of using both classical machine learning algorithms and neural networks for detecting characteristic gaze behaviors associated with ASD. Nonetheless, most works use rather small-sized datasets, which affects negatively on the generality of their results.

2.5 Evaluation Metrics

Classification model performances must be evaluated using evaluation measures that are objective. When studying ASD detection, there are some measures commonly used for evaluating machine learning methods' performances.

2.5.1 Accuracy

Accuracy refers to the number of correctly classified samples divided by the total number of samples. This can be considered the overall performance of classification and is expressed as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

2.5.2 Precision

Precision measures the proportion of correctly identified positive samples among all samples predicted as positive.

$$\text{Precision} = \frac{TP}{TP + FP}$$

2.5.3 Recall

Recall, otherwise referred to as sensitivity, is the ratio of positive samples to those identified correctly by the classifier.

$$\text{Recall} = \frac{TP}{TP + FN}$$

In scenarios like autism screening, recall is significant since failure to detect the patients could mean a delay in treatment.

2.5.4 F1-Score

The F1-score is calculated using the harmonic mean of both precision and recall.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

This metric is especially useful when class distributions are imbalanced.

2.5.5 ROC-AUC

This is because the ROC curve displays graphically the relation between the true positive rate and false positive rate at different threshold levels for classification.

The higher the AUC score, the greater the ability to discriminate between the ASD class and the TD class. This is why ROC-AUC has become one of the most cited metrics in autism classification research.

2.6 Research Gap and Motivation

However, there are still several limitations encountered in eye-tracking-based ASD detection methods. First of all, quite a lot of studies are using rather small databases that could limit the results' generalization. Secondly, there is a lot of diversity in feature extraction methods that makes comparison of studies quite

complicated. Lastly, even though deep learning techniques were extensively explored, more conventional machine learning algorithms need more thorough research.

In addition, quite little research has been done to investigate the capabilities of different ensemble learning approaches on the same eye tracking data. Ensembles have proved themselves in numerous machine learning tasks by successfully integrating the strengths of multiple learners. However, they are yet to be tested for their potential in ASD detection based on eye-tracking.

This thesis aims to examine the use of several machine learning techniques, such as Random Forest, Support Vector Machine, AdaBoost, XGBoost, CatBoost, as well as an ensemble approach, in identifying the early signs of ASD based on eye-tracking data obtained from the Saliency4ASD data collection. Apart from analyzing classification results, this research also seeks to explore the role of visual attention-based features in the context of detecting gaze behavior linked to autism.

CHAPTER 3 METHODOLOGY

The methodology developed in this thesis for ASD diagnosis by means of visual attention assessment based on gaze and saliency analysis will be described in this chapter. This model is comprised of eye tracking data processing, computational saliency modeling and object/face detection to learn discriminative attention profiles between ASD children and their typically developing peers through machine learning algorithms. Also, the inclusion of a multi-layer perceptron (MLP) extension is considered to exploit nonlinear interactions between the extracted features.

The system architecture comprises six components:

1. Data collection and organisation,
2. Preprocessing and feature extraction,
3. Base model training,
4. Probability calibration and threshold optimisation,
5. Ensemble learning (soft voting and stacking),
6. MLP extension for comparison and ablation.

3.1 Dataset

This paper utilizes the Saliency4ASD dataset [32], which is an open-source benchmark eye-tracking dataset designed specifically for studies concerning the visual attention system of individuals with Autism Spectrum Disorder (ASD). This dataset was originally proposed by Duan et al., and subsequently, became the standard benchmark dataset for the Saliency4ASD Grand Challenge held during IEEE International Conference on Multimedia and Expo (ICME) 2019.

The data set comprises 300 natural images in the size of 1024 × 768 pixels. These images have been chosen in such a way that they depict complex environments, having different types of stimuli in them such as human face stimuli, social interaction stimuli, and background.

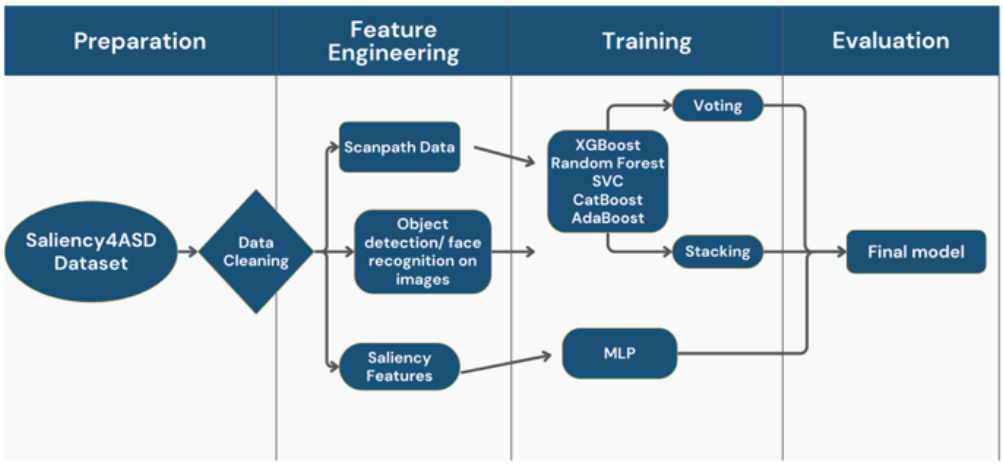


Figure 3.1: Overview of the proposed multimodal ASD detection framework.

Eye-tracking measurements were obtained from a total of 28 subjects, divided into 14 children with a diagnosis of ASD and 14 TD children. In the experimental session, the subject was free to view each picture while their eye movements were tracked using a Tobii eye-tracker. This free-viewing technique was employed to record natural behavior during visual exploration tasks.

The eye-tracking data for each stimulus picture include:

Fixation points for individual participants; Scan paths of individual subjects illustrating the gaze sequence; Fixation maps based on aggregation of fixations; Group saliency maps produced independently for individuals with ASD and TD.

Probably the most notable aspect about Saliency4ASD is the fact that both groups were exposed to the same visual inputs and therefore allowed to compare eye movement and gaze behavior patterns across individuals with ASD and typically-developing participants. This aspect renders the database highly appropriate for investigating visual exploration patterns associated with autism.

Thanks to its open nature, standardized design, and involvement of individuals with and without ASD, Saliency4ASD has proven extremely useful as a benchmark dataset in visual attention research related to ASD. The database was employed heavily in various experiments focusing on prediction of saliency, investigation of

gaze behavior patterns, modeling of attention, and development of machine learning-based ASD diagnosis models.

The use of the Saliency4ASD database is essential to provide eye-gaze-based features that can be used in this thesis for assessing the proposed machine-learning model for ASD classification. It is because the presence of eye-tracking data from subjects of both ASD and TD is a perfect starting point for such research tasks.

3.2 Preprocessing and Feature Engineering

To be able to construct a reliable machine learning pipeline to detect ASD, raw gaze data and the related visual input were structured and formatted into a structured dataset ready for analysis by an algorithm. First of all, the continuous variables were standardized based on their z-scores, since standardization removes the risk of any discrepancy in scale between the different variables used, thus reducing the effect any one variable could have on the learning of the model. The categorical variables, such as type of object, were uniformly coded to keep them consistent and avoid confusion. Roughly 43 features per observation were created.

Saliency descriptors were obtained based on computational deep saliency models, particularly the SAM framework with a ResNet architecture [1]. The saliency model gives information regarding which parts of an image attract human attention. Quantitative measures were extracted from each participant's eye movements using the different quantifiers. These quantifiers include the initial saliency, saliency ratio for which the fixation was regarded as being an example for the first time that exceeded the predefined threshold of the maximum saliency value. Furthermore, other saliency features such as the mean, sum, and maximum saliency values during fixations were considered. To capture time-dependent changes in visual attention, the duration of fixations was normalized against their saliency values, resulting in weighted durations. Moreover, spatial distance measures, such as the Kullback-Leibler distance and normalized scanpath saliency, were used to determine the similarity between the experimental fixation distribution and the expected saliency map [2], [3].

The face and object detection metrics were calculated using the integration of the MediaPipe EfficientDet Lite0 [4] for object detection and CNN based face detection assisted by the face_recognition library. Post-processing was done to make sure the classifications of the classes were right and avoid any irrelevant detections. For every image, the following features were recorded: the number of detected faces/objects; size of bounding boxes; mean confidence scores; the proportion of fixation points inside a face/object area. Thus, these features measure how much attention was paid to social aspects (face detection) and to nonsocial aspects (object detection) in images. The distinction is important because differentiating characteristics of typical versus atypical visual exploration can be established based on such differences.

Gaze metrics included fixation counts, total, average, and variance of fixation durations, and saccade amplitudes as Euclidean distances between consecutive fixations. In addition, the sum of all saccade amplitudes gave a measure of exploration through the concept of scanpath length. Further, distances between fixations and image center and mean fixation positions were computed to evaluate centrality and distribution of gaze. Ratios of fixations in regions of interest, defined as faces, objects, or background areas, were also calculated in order to study attention to social and non-social content. Together, these measures offer a broad picture of viewing behavior and make a good starting point for machine learning.

The inclusion of saliency maps, object and face detectors, and gaze metrics in one comprehensive feature matrix makes sure that the temporal and spatial characteristics of visual attention are captured. The resulting multimodal feature set can be used to learn the differences in viewing behavior of ASD individuals.

3.3 Base Models and Training Procedure

This section presents the practical implementation of the baseline machine learning models, from environment setup to advanced optimization procedures. The primary objective is to transform the theoretical foundations of these algorithms into a

reproducible and reliable ASD screening system suitable for a limited clinical dataset.

3.3.1 Implementation Environment and Pipeline Architecture

The entire framework was implemented in Python 3.11.3, utilizing well-established machine learning libraries including scikit-learn for the general pipeline, and XGBoost and CatBoost for gradient boosting algorithms. Rather than training each classifier independently, an encapsulated pipeline architecture was developed in the framework.

Each pipeline consists of two major components:

- **Preprocessing:** All numerical features were standardized using StandardScaler, which transforms each feature into a Z-score distribution. This normalization prevents variables with different measurement scales, such as fixation duration (milliseconds) and pixel coordinates, from disproportionately influencing the learning process.
- **Model Execution:** The selected classifiers were integrated with either their default configurations or slightly adjusted hyperparameters, providing a consistent baseline for subsequent model comparison and ensemble learning.

3.3.2 Participant-Level Data Splitting Strategy

To address concerns regarding the reliability of results obtained from a relatively small dataset consisting of 28 participants, model evaluation was performed using a participant-level splitting strategy rather than an image-level split.

Information Leakage Prevention: All observations (images) belonging to the same participant were assigned exclusively to either the training set or the testing set. This strategy prevented information leakage that could otherwise lead to overly optimistic performance estimates.

Dataset Organization: Data from the 28 participants were partitioned such that the independent test set contained the final 30 images, serving as previously unseen samples for objectively evaluating the model's generalization capability to new individuals.

3.3.3 Detailed Implementation of the Base Models

Five machine learning algorithms with distinct mathematical characteristics were selected to capture complementary information from gaze behavior, saliency features, and object detection metrics.

Random Forest (RF)

Random Forest was implemented as an ensemble of decision trees constructed using the bootstrap sampling technique. Within the implementation, RF was particularly effective at modeling complex nonlinear interactions among features, such as the relationship between saliency values at fixation locations and fixation duration on facial regions. Furthermore, the feature importance scores generated by RF were utilized to interpret which gaze patterns contributed most significantly to ASD prediction.

Extreme Gradient Boosting (XGBoost)

XGBoost served as the primary model for maximizing classification performance. Through the implementation pipeline, XGBoost employed second-order gradient optimization based on both gradient and Hessian information, enabling the model to assign greater emphasis to previously misclassified samples during successive boosting iterations. This optimization strategy allowed XGBoost to achieve the highest classification accuracy (**70.1%**) among all individual baseline models.

Support Vector Classifier (SVC)

The Support Vector Classifier was configured with an appropriate kernel function to identify the optimal hyperplane that maximizes the margin between ASD and

typically developing (TD) participants. In the framwork probability estimation was enabled, allowing calibrated probability outputs to be generated for subsequent threshold optimization. After calibration, SVC achieved the highest Recall (**88.4%**), making it particularly suitable for clinical screening applications where minimizing false negatives is a critical objective.

AdaBoost (Adaptive Boosting)

AdaBoost was implemented to iteratively improve classification performance by increasing the weights assigned to previously misclassified observations. Although the algorithm is relatively sensitive to noisy data, it complements the overall framework by capturing less common gaze behavior patterns exhibited by children with ASD.

CatBoost

CatBoost was included because of its ability to effectively handle categorical variables without requiring extensive manual encoding. This capability helps preserve the inherent relationships among categorical object classes, such as humans, animals, and everyday objects appearing in visual stimuli, thereby improving the representation of semantic information within the dataset.

3.3.4 Probability Calibration and Threshold Optimization

An important practical contribution of this study is the implementation of probability calibration and threshold optimization. Although these procedures are often omitted in theoretical descriptions, they are essential for developing clinically reliable prediction systems.

Probability Calibration

The probability outputs produced by classifiers such as Random Forest and SVC are often poorly calibrated, meaning that predicted probabilities do not accurately reflect the true likelihood of ASD. Therefore, probability calibration techniques

were applied to ensure that an output probability (e.g., 0.80) more closely corresponds to an actual 80% probability of ASD.

F2-Score-Based Threshold Optimization

Instead of adopting the conventional decision threshold of 0.5, a comprehensive threshold analysis was performed to determine the optimal classification threshold. Because ASD screening prioritizes identifying affected children over minimizing false positives, the threshold was selected to maximize the F2-score, which assigns twice as much importance to Recall as to Precision. This optimization strategy enables the proposed system to detect a larger proportion of children at risk for ASD while accepting a moderate increase in false-positive predictions, making it more appropriate for early clinical screening.

3.4 Ensemble Methods

In addition to boosting classification effectiveness and generalizing to new subjects, two approaches in ensemble learning were considered. They are “soft voting” and “stacking”. Ensemble learning techniques have been proven especially efficient when the individual base classifiers represent different facets of the underlying data structure, such as gaze-based multimodal features.

In Soft Voting, the output probabilities from each base classifier are averaged using a weighting factor depending on each base classifier's performance on the validation set. The weighting was optimized through an exhaustive grid search where we attempted to maximize the F2 score, favoring recall, a vital metric in clinical screening scenarios. In soft voting, we utilize the diversity of the base learners, such that each base classifier addresses certain patterns or relationships within the data. For instance, while tree-based learners such as Random Forests (RF) and Extreme Gradient Boosting (XGB) are more effective in modeling non-linear feature interaction, SVC can model more complicated decision boundaries. Boosting models such as AdaBoost and CatBoost concentrate on finding rare patterns.

The Stacking approach is an advanced form of ensemble techniques where

predictions made by base classifiers are employed as inputs to a second-level meta-learner. Stacking provides an opportunity for a meta-learner to find the optimal combination of base models' outputs, thereby reducing any possible bias and even finding some relationships between outputs of different models that would be impossible to capture using simple averaging. Three Stacking setups were analyzed, varying only in meta-learner algorithms – K-Nearest Neighbors, Naive Bayes Classifier, and Logistic Regression. The meta-learner was trained using the outputs obtained using the cross-validation technique on base learners' probability predictions, in order to ensure there was no data leakage during the process. This technique allowed us to train our meta-learners to generalize from predictions of base models, rather than memorize base models' behavior. All base classifiers were retrained using the whole training set, whereas meta-learners were fitted using validation set outputs.

Both methods of ensembling have their own benefits. Soft Voting is simple to implement and easy to understand, which also gives an indication of how confident the system is by considering the relative weights of each classifier's prediction. On the other hand, Stacking is more sophisticated and is able to exploit the dependencies of predictions made by the base models, which may result in a high level of performance and a proper balance of precision and recall. The results of empirical research show that ensembles, especially stacking combined with Naive Bayes or Logistic Regression, perform better than base classifiers.

The approach is built on the advantage of using different learning algorithms, while minimizing weaknesses caused by using each particular algorithm. When it comes to diagnosing ASD, this approach is particularly useful as it guarantees accuracy with a minimum number of false negatives.

3.5 Deep Neural Networks

Although traditional machine learning models demonstrated strong performance on the relatively small dataset, an additional deep learning branch based on a Multi-Layer Perceptron (MLP) was developed to investigate the capability of deep neural

networks to automatically learn more complex nonlinear relationships from multimodal tabular features.

3.5.1 Multi-Layer Perceptron Architecture

Instead of adopting a deep and computationally expensive neural network that would be prone to overfitting on a dataset containing only 28 participants, a lightweight yet expressive MLP architecture was designed specifically for tabular multimodal data. The network architecture consists of the following components:

- **Input Layer:** Receives the 43 standardized multimodal features extracted from gaze metrics, saliency descriptors, and object detection features.
- **Hidden Layer 1:** Contains 64 neurons with the ReLU activation function, followed by Batch Normalization to stabilize the learning process and accelerate convergence.
- **Hidden Layer 2:** Contains 128 neurons, enabling the network to learn more complex nonlinear interactions among gaze behavior and saliency-related features.
- **Hidden Layer 3:** Contains 64 neurons, serving as a feature compression stage before the final classification layer.
- **Output Layer:** Consists of a single neuron with a Sigmoid activation function, producing the probability of ASD for the binary classification task (ASD vs. TD).

3.5.2 Training Strategy for Small-Scale Data

Training deep neural networks on the Saliency4ASD dataset presents a significant challenge because of the limited number of participants, making overfitting a major concern.

During training, the learning dynamics were carefully monitored by comparing the training loss and validation loss, characterized by a continuously decreasing training loss accompanied by an increasing or unstable validation loss, was considered an indication that the model had begun to overfit.

To improve training stability, Batch Normalization was adopted in place of

Dropout in several experimental configurations, allowing the network to retain as much information as possible from the limited number of training samples. The model was trained using the Binary Cross-Entropy loss function and optimized with the Adam optimizer, whose learning rate was carefully tuned through experimentation.

Within the overall framework, the MLP served not only as an independent classifier but also as a deep learning baseline for comparison with traditional machine learning approaches. Although its overall classification accuracy was lower than that of the ensemble models because of the limited training data, the MLP achieved a Recall of 86.4% and an F2-score of 80.0%, demonstrating that deep neural networks remain highly effective in identifying subtle ASD-related visual attention patterns even in small-scale clinical datasets.

CHAPTER 4 EXPERIMENTAL RESULTS

4.1 Performance Analysis

In this section, there is a complete analysis of the machine learning models that have been proposed to discriminate between ASD and TD children. This is done through the use of confusion matrixes, ROC curve analysis, AUC, and training dynamics. All these graphs give a different outlook into the way the model works, hence making it possible to measure not only accuracy but also sensitivity, specificity, and generalization of the algorithm.

Table 4.1 summarizes the quantitative performance of all evaluated models, while Figures 4.1–4.11 illustrate their classification characteristics.

Table 4.1: Performance comparison of evaluated models.

Model	Accuracy	Precision	Recall	F1	F0.5	F2
RF [34]	59.8	57.4	71.7	63.2	59.8	68.3
RF	68.6	64.0	83.1	72.3	67.1	78.4
XGB	70.1	66.3	80.1	72.5	68.6	76.9
SVC	64.3	59.3	88.4	71.0	63.5	80.5
AdaBoost	67.1	62.7	82.0	71.1	65.8	77.2
CatBoost	66.4	61.3	86.4	71.7	65.1	79.9
Stacking (KNN)	68.7	68.7	68.7	68.7	68.5	66.8
Stacking (Logistic	70.4	70.6	68.4	69.5	70.1	68.8
Stacking (Naive Bayes)	71.7	72.5	68.7	70.6	71.7	69.4
Soft Voting	69.5	64.6	84.8	73.3	67.8	79.8
MLP	66.9	61.8	86.4	72.1	65.5	80.0

4.1.1 Performance of Tree-Based Models

Figure 4.1 shows the performance of the Random Forest algorithm, while Figure 4.2 illustrates that of XGBoost. It can be seen that both algorithms perform relatively well-balanced classification tasks, and there is a proper balance between the identification of ASD and TD individuals as indicated by the confusion matrices.

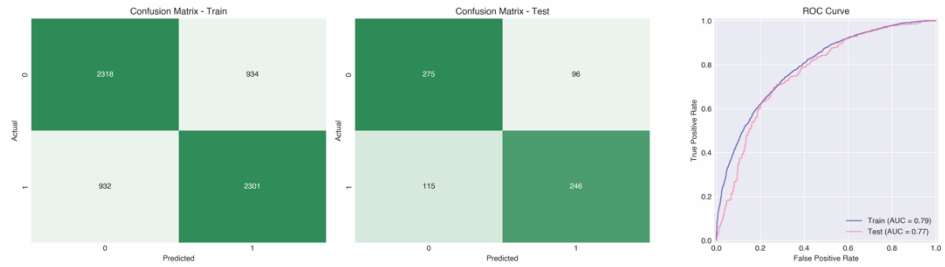


Figure 4.1: Random Forest performance analysis.

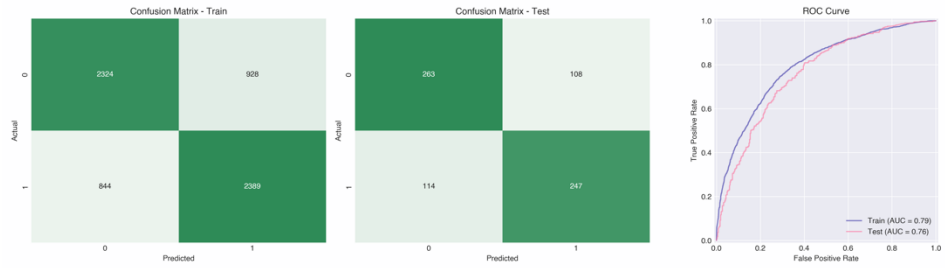


Figure 4.2: XGBoost performance analysis.

The performance metrics of the Random Forest algorithm include an accuracy of 68.6%, a recall of 83.1%, and an F2-score of 78.4%. Inspection of the confusion matrix shows that the classifier is capable of recognizing many instances of ASD patients, although it has a reasonable false positive rate. These results are expected as Random Forest relies on several decision trees and thus is capable of learning non-linear interactions between gaze, saliency, and object-based data.

Among individual machine learning algorithms, the best accuracy was observed with XGBoost algorithm, which scored 70.1%. The ROC curve obtained using XGBoost shows that there is clear discrimination between the classes (ASD, TD), implying effective ranking by this model. The precision for XGBoost and Random Forest was recorded as 66.3% and 64.0%, respectively, showing better performance of XGBoost algorithm. This is attributed to the fact that the gradient boosting technique of this model helped in lowering the number of false positive errors made by the classifier. XGBoost algorithm also exhibited high recall of 80.1%.

4.1.2 Performance of Margin-Based and Boosting Models

As for the Support Vector Classifier (Figure 4.3), the classification results obtained by this classifier were quite distinct from those of the other models. Even though its overall accuracy level was not very high (64.3%), SVC demonstrated the highest value of recall (88.4%) amongst all other considered models.

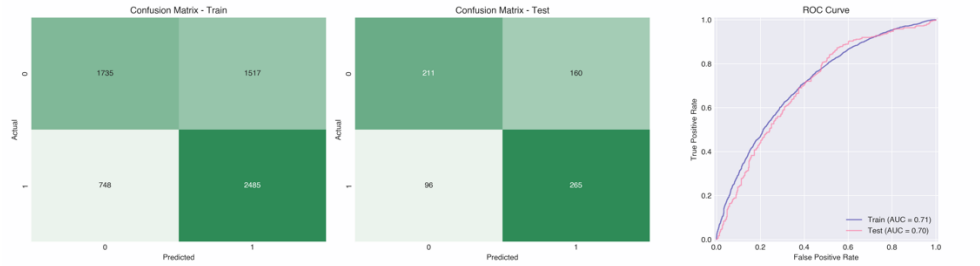


Figure 4.3: Support Vector Classifier (SVC) performance analysis.

The results of the confusion matrix analysis clearly indicate that this classifier managed to classify nearly all ASD patients, producing an increased amount of false positives relative to other machine learning models. This tendency indicates that the model's algorithm puts more emphasis on sensitivity rather than specificity. For screening purposes, this trait can prove quite useful since failing to identify ASD patients is always more dangerous than identifying TD patients.

The AdaBoost model (Figure 4.4) scored 67.1% for accuracy and 82.0% for recall. From the confusion matrix provided, the AdaBoost model demonstrates a fairly balanced classification result, but from the ROC curve graphically presented, it appears that the model may be a little less discriminative than the Random Forest and XGBoost models. However, due to AdaBoost's ability to focus on challenging samples, it will improve the ability to detect ASD-related features; on the other hand, the model's vulnerability to noise can affect its overall accuracy because it is being used in a small sample size.

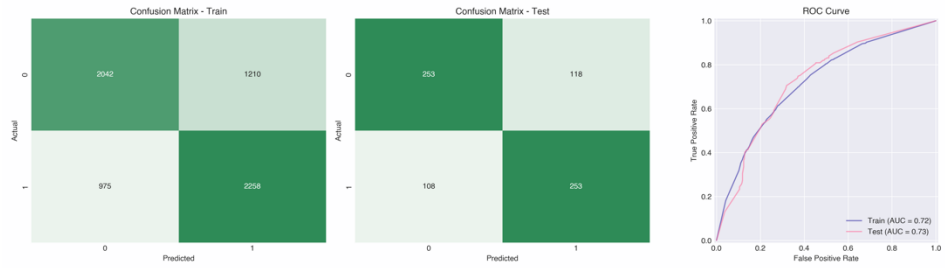


Figure 4.4: AdaBoost performance analysis.

In terms of the results provided, CatBoost showed good sensitivity, scoring a high recall rate of 86.4% and F2 score of 79.9%. The confusion matrix provided shows that the CatBoost model was able to detect many instances of ASD, albeit having a fairly balanced classification. This was probably because of CatBoost's ordered boosting mechanism, which helped prevent overfitting.

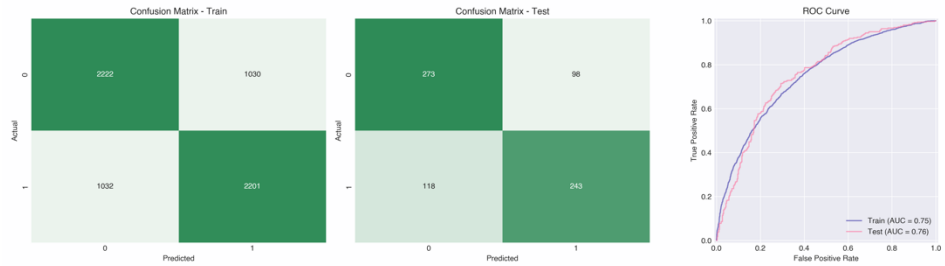


Figure 4.5: CatBoost performance analysis.

4.1.3 Performance of Ensemble Models

Some of the most intriguing results came out of the ensemble-based approaches. The goal of the ensembling technique is to exploit the advantages of various classifiers by minimizing variance.

Stacking with Logistic Regression (Figure 4.6) increased both accuracy (70.4%) and precision compared to other ensembles, being one of the best models for classifying ASD. The confusion matrix shows the more conservative way of decision making, resulting in lower false positive rate but still having a rather good sensitivity level. This is in line with the features of the used method, where decision boundaries are smoother.

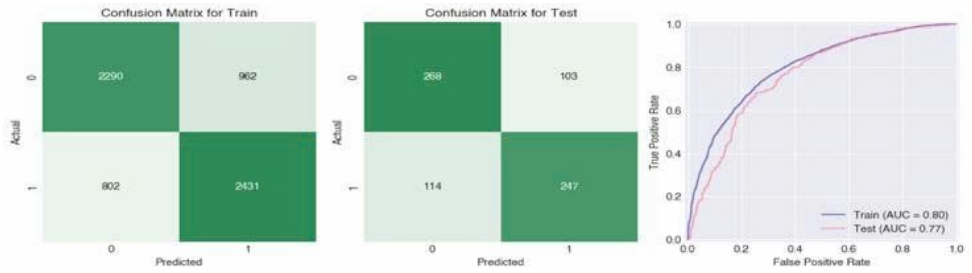


Figure 4.6: Stacking ensemble with logistic regression meta-learner.

The best-performing method was found in the Stacking ensemble with Naive Bayes as a meta-learner. The model demonstrated the highest accuracy (71.7%) and precision (72.5%), outperforming all the rest of the considered techniques. From the point of view of the confusion matrix, the method proved to work well in distinguishing between two classes, showing a good balance in terms of ASD detection and TD classification. The ROC curve also speaks about excellent discriminative capabilities of the method.

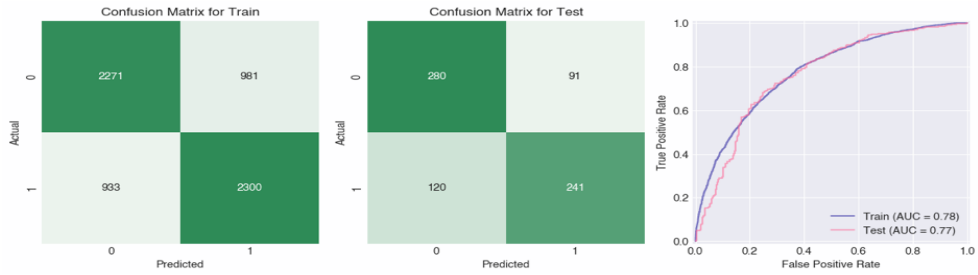


Figure 4.7: Stacking ensemble with Naive Bayes meta-learner.

The Stacking method with the meta-learner as KNN Classifier yielded an extremely balanced decision boundary since the accuracy, precision, recall, and F1-score values are all approximately equal at 68.7%. The low values for the recall and F2-score can mean that KNN as a meta-learner might not be expressive enough.

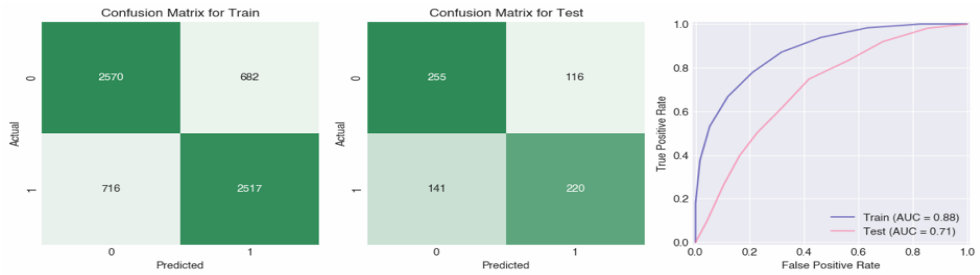


Figure 4.8: Stacking ensemble with KNN meta-learner.

Soft Voting Classifier showed a very good trade-off between its recall score and classification accuracy. Its recall was 84.8%, while F1 and F2-scores were 73.3% and 79.8% respectively. This classifier always occupied one of the leading positions in the ranking by various metrics. The analysis of the confusion matrix and ROC curves showed that the classification results were balanced. Thus, the positive outcome proves the advantage of using calibrated predictions of several classifiers.

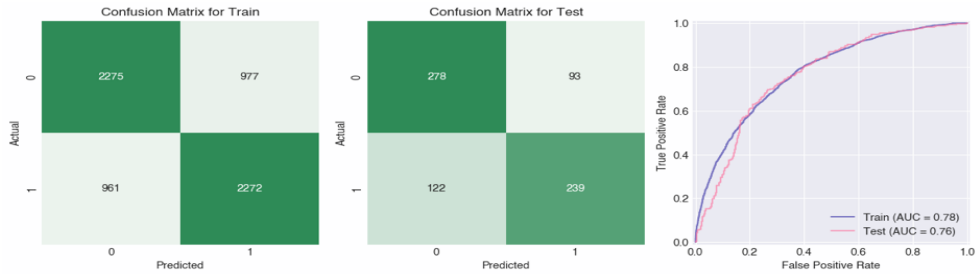


Figure 4.9: Soft voting ensemble performance analysis.

4.1.4 Performance of the Deep Learning Model

Figures 4.10 and 4.11 depict the outcomes of the multi-layer perceptron (MLP). From the confusion matrix, it is clear that the neural net was able to successfully classify a significant portion of the cases with ASD, producing an excellent recall rate of 86.4%. The sensitivity of the network is similar to that of CatBoost but lower than that of SVC by a small margin.

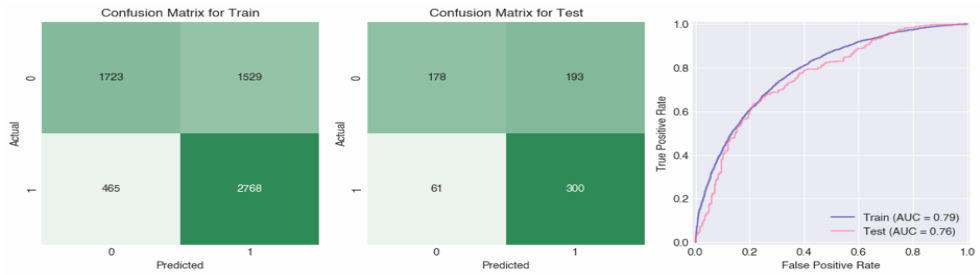


Figure 4.10: MLP performance analysis.



Figure 4.11: MLP training loss over epochs.

Surprisingly enough, while the classification performance of the neural network is moderate (at just 66.9%), its F2-score turned out to be quite high (equaling 80.0%).

Another important observation is made by analyzing the dynamics of the training. In this case, we see that the training loss decreases monotonically, showing that the network successfully discovers some patterns in the data. However, at the same time, the validation loss significantly varies and remains much larger compared to the training loss. This discrepancy is a clear sign of the presence of overfitting in the process.

There are several reasons why this situation occurs. First of all, the number of samples in the Saliency4ASD dataset is quite small, as it includes only 28 people. Furthermore, the dimensionality of the feature space is relatively large, which increases the likelihood that the model will simply learn the features of each particular participant instead of discovering general patterns related to autism.

But despite this, the performance of the MLP, measured by its recall rate and F2-score, indicates that deep learning architectures still have the potential for autism diagnosis in the future, when large datasets will be available. By incorporating new participants, using data augmentation, or applying a transfer learning technique, neural network algorithms might prove to be superior to classical machine learning techniques.

4.1.5 Comparative Discussion

Several important trends emerge from the experimental results.

Ensemble models demonstrated superior performance relative to all other machine learning algorithms evaluated. Stacking and Soft Voting models demonstrated superior results compared to most other algorithms used in this experiment, indicating that different algorithms capture complementary aspects of visual attention behavior.

Algorithms that paid more attention to recall (SVC, CatBoost, Soft Voting, MLP) performed well for the task of detecting ASD due to high F2-scores and thus can be considered good candidates for clinical implementation.

The Tree-based algorithms (Random Forest, XGBoost) showed superior performance in terms of trade-off between accuracy, precision, and interpretability. The excellent results obtained using these algorithms indicate the presence of complex nonlinear relationships in multimodal gaze and saliency features.

The analysis of conventional machine learning approaches versus deep learning techniques proves that a high level of complexity does not guarantee high performance. In spite of a high level of interaction obtained through MLP, the lack of generalization ability was observed because of the relatively low number of cases within the dataset. In turn, ensemble machine learning algorithms showed a substantially better performance rate.

The conducted studies have proved that multimodal visual attention-based features have a great potential in terms of distinguishing between people with ASD. Among all tested algorithms, the highest performance rates were demonstrated by the

Stacking (Naive Bayes). The greatest sensitivity was obtained through SVC, Soft Voting, CatBoost, and MLP. These findings motivate using ensemble machine learning methods for diagnosis purposes.

4.2 Relationship Between Visual Features and ASD-Specific Viewing Behavior

The analysis of eye-tracking engineering features allowed identifying critical differences between the behavior of children in the ASD and TD groups. It was found that feature importance analysis revealed those features related to the attention to socially relevant aspects, particularly the duration of fixations, and the percentage of face-related fixations, as the most discriminating. The TD children focused their attention more often on socially significant areas, such as human faces and central areas containing socially significant information. Conversely, children in the ASD group had shorter attention times devoted to the socially relevant stimuli.

Another interesting observation concerns the relationship between gaze and saliency maps. The correlation between fixations of TD participants' eyes and predictions based on the latter maps was greater, indicating that visual attention in these individuals depended to a greater extent on bottom-up characteristics, namely the contrast, colors, and saliency of objects. In turn, the correlation of saliency with ASD visual attention was lower. Therefore, visual attention in children with ASD may depend not on salient objects but on other, less common attention cues.

It should be noted that ASD participants differed from TD ones concerning gaze dynamics and visual scanning techniques. For one thing, the former had shorter fixation duration; for another thing, the dispersion of fixation points among ASD subjects was wider. It means that the paths of the ASD participants' eyes' movements were less consistent as compared to those of the control group members. In other words, gaze atypicality in ASD may indicate inconsistency of visual attention and its inefficiency when processing socially significant information.

4.3 Implications for Clinical Screening and Early ASD Detection

The combination of eye tracking parameters and computer vision approaches like saliency analysis and object/face detection can serve as an integral strategy for

characterizing visual attention patterns. In the proposed methodology, apart from objective criteria relying on attentional mechanisms caused by stimuli, subjective ones are also included, taking into account gaze behavior of people and their social attention.

Up to 88.4% and 86.4% recall scores for SVC model and CatBoost models, respectively, represent promising prospects for employing gaze-based biomarkers in order to assist autism spectrum disorders diagnosis at an early stage. It can be noted that visual attention features can help diagnose autism

Furthermore, the suggested framework also presents the advantage of being interpretable, which can often be missing from diagnostic approaches driven purely by data. For example, mean saliency of fixations, fixation duration, and dwell time on socially relevant AOIs represent some of the parameters that can serve as quantifiable indicators of abnormal attention processing. By converting such gaze patterns into objective measurements, the presented framework could potentially help develop user-friendly AI-powered screens for detecting early symptoms of ASD alongside regular clinical examinations.

CHAPTER 5 CONCLUSIONS AND FUTURE WORK

This thesis has demonstrated significant results showing how the use of eye-tracking metrics, computational saliency, and object/face detection features together may provide useful information on the visual attention behavior of children with Autism Spectrum Disorder (ASD). This is shown through an analysis of both conventional classification machine learning algorithms as well as ensemble-based approaches which highlight how such data-driven approaches may be used in the early screening of ASD.

A study of gaze behavior demonstrated that children diagnosed with autism spectrum disorder (ASD) display shorter and more scattered fixations with less attention paid to socially relevant features, such as faces. In contrast, TD children displayed longer fixations and higher salience towards the center and socially relevant parts of the stimuli. Additionally, the trajectories of gazes made by the ASD group demonstrated less correspondence with the computational models of saliency maps, indicating their insensitivity to the visually or semantically relevant parts of stimuli. The analysis of feature importance indicated that the salience of fixation points and fixation time in face-related areas were the most informative behavioral features.

The Support Vector Classifier model performed the most optimally in the recall metric among all other methods, having achieved an F2-score value of 80.5%. Moreover, it was found that the ensembles also yielded stable results, thus confirming that the usage of ensembles can enable the utilization of the full potential of the multimodal eye gaze data. It should be noted that the application of probability calibration and decision threshold tuning helped improve the recall metric while not compromising the precision one.

Nevertheless, some limitations need to be mentioned even in spite of promising findings that have been made. First, the sample size was small – 14 children with ASD, 14 typically developing children and 300 still pictures were utilized. Second, the analysis of images can not take into consideration the dynamics of attention, both long-term or within a single session, while the experimental setting with the use of

stationary eye-tracking devices is likely to differ from real-life scenarios such as class or household settings. Lastly, the age of participants ranged from 8 to 15 years old, and the question arises as to whether the same attention pattern exists among smaller kids.

Future work can therefore focus on addressing these limitations and scaling up the methodology in different dimensions. The development of models on datasets comprising people of diverse ages, genders, and ethnicities will enhance the generalizability and external validity of the models. In addition, using fully integrated systems where an end-to-end neural network is employed to include the eye movements, gaze saliency and image features will make it possible to capture more complex attention behaviors. Finally, the practical applications of the gaze capture techniques in medical settings or even homes should not be overlooked since these applications require high levels of efficiency, reliability, and ease of use. On the other hand, ethical issues associated with privacy and data management cannot be ignored.

In this research, it is evident that there is enormous scope for applying gaze-based cues in conjunction with computational saliency and machine learning algorithms for detecting autism spectrum disorder. Even though there might not yet be any clinically viable system due to its lack of accuracy, there have been attempts to detect behavioral traits specific to autism.

REFERENCES

[1] Marcella Cornia et al. “Predicting human eye fixations via an LSTM-based saliency attentive model”. In: *IEEE Transactions on Image Processing* 27.10 (2018), pp. 5142–5154. DOI: 10.1109/TIP.2018.2851672.

[2] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. “DeepGaze II: Reading fixations from deep features trained on object recognition”. *arXiv:1610.01563* (2016).

[3] R. J. Peters et al. “Components of bottom-up gaze allocation”. In: *Journal of Vision* (2005). DOI: 10.1167/5.8.706.

[4] Mingxing Tan, Ruoming Pang, and Quoc V. Le. “EfficientDet: Scalable and Efficient Object Detection”. In: *CVPR* (2020). DOI: 10.1109/CVPR42600.2020.01079.

[5] Warren Jones and Ami Klin. “Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism”. In: *Nature* 504.7480 (2013), pp. 427–431. DOI: 10.1038/nature12715.

[6] Mihaela Chita-Tegmark. “Social attention in ASD: A review and meta-analysis of eye-tracking studies”. In: *Research in Developmental Disabilities* 48 (2016), pp. 79–93. DOI: 10.1016/j.ridd.2015.10.011.

[7] Terje Falck-Ytter et al. “The development of gaze following in four-month-olds: Left visual field bias is modulated by face gender”. In: *Scientific Reports* 8.1 (2018), pp. 1–9. DOI: 10.1038/s41598-018-23829-y.

[8] Qi Wang et al. “Impaired organization of paired-associate learning in toddlers with autism spectrum disorder”. In: *Autism Research* 11.5 (2018), pp. 690–700. DOI: 10.1002/aur.1932.

[9] Thomas W. Frazier et al. “Development and validation of objective and quantitative eye tracking–based measures of autism risk and symptom levels”. In: *Journal of the American Academy of Child and Adolescent Psychiatry* 57.11

(2018), pp. 858–866. DOI: 10.1016/j.jaac.2018.06.023.

[10] Shuo Wang et al. “Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking”. In: *Neuron* 88.3 (2015), pp. 604–616. DOI: 10.1016/j.neuron.2015.09.042.

[11] Matthias Kümmerer and Matthias Bethge. “State-of-the-art in human scanpath prediction”. In: *Journal of Vision* 21.9 (2021), pp. 2033–2033. DOI: 10.1167/jov.21.9.2033.

[12] Romain Carette et al. “Automatic autism spectrum disorder detection thanks to eye-tracking and neural network-based data analysis”. In: *ICAART* 2 (2019), pp. 552–559. DOI: 10.5220/0007401205520559.

[13] Zhong Zhao et al. “An eye tracking study: Positive emotional interface design facilitates learning outcomes in multimedia learning?” In: *International Journal of Educational Technology in Higher Education* 18.1 (2021), pp. 1–15. DOI: 10.1186/s41239-021-00255-4.

[14] Ming Jiang and Qi Zhao. “Learning visual attention to identify people with autism spectrum disorder”. In: *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 3267–3276. DOI: 10.1109/ICCV.2017.354.

[15] Ishtiaque Ahmed et al. “A novel multi-modal machine learning framework for ASD screening using eye-tracking and behavioral data”. In: *Frontiers in Psychiatry* 13 (2022). DOI: 10.3389/fpsyt.2022.943981.

[16] Matthew J. Maenner et al. “Prevalence and characteristics of autism spectrum disorder among children aged 8 years—Autism and Developmental Disabilities Monitoring Network, 11 sites, United States, 2020”. In: *MMWR Surveillance Summaries* 72.2 (2023), p. 1. DOI: 10.15585/mmwr.ss7202a1.

[17] Karen Pierce et al. “Preference for geometric patterns early in life as a risk factor for autism”. In: *Archives of General Psychiatry* 68.1 (2011), pp. 101–109. DOI:

10.1001/archgenpsychiatry.2010.113.

[18] Ami Klin et al. “Two-year-olds with autism orient to non-social contingencies rather than biological motion”. In: *Nature* 459.7244 (2009), pp. 257–261. DOI: 10.1038/nature07870.

[19] Guobin Wan et al. “Applying eye movement modeling examples to guide novices’ attention in the comprehension of process models”. In: *Information Sciences* 482 (2019), pp. 1–19. DOI: 10.1016/j.ins.2019.01.066.

[20] Wen Liu, Ming Li, and Li Yi. “Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework”. In: *Autism Research* 9.8 (2016), pp. 888–898. DOI: 10.1002/aur.1615.

[21] Mahmoud Elbattah et al. “Learning to diagnose autism spectrum disorder from eye-tracking data using convolutional neural networks”. In: *Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management* (2021). DOI: 10.5220/0010659100003064.

[22] Adel Belkadi Dris et al. “Automatic autism spectrum disorder detection using artificial intelligence methods with MRI neuroimaging: A review”. In: *Frontiers in Molecular Neuroscience* 15 (2022). DOI: 10.3389/fnmol.2022.999605.

[23] Julia Schmidt Oliveira et al. “Computer-aided autism diagnosis using visual attention models and eye-tracking: replication and improvement proposal”. In: *BMC Medical Informatics and Decision Making* 23.1 (2023), pp. 1–12. DOI: 10.1186/s12911-023-02365-7.

[24] Guobin Wan et al. “Machine learning for early detection of autism (and other conditions) using parental questionnaires”. *arXiv:2204.06285* (2022).

[25] Laurent Itti, Christof Koch, and Ernst Niebur. “A model of saliency-based visual attention for rapid scene analysis”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.11 (1998), pp. 1254–1259. DOI: 10.1109/34.730558.

[26] Tilke Judd et al. “Learning to predict where humans look”. In: *2009 IEEE 12th International Conference on Computer Vision* (2009), pp. 2106–2113. DOI: 10.1109/ICCV.2009.5459462.

[27] Ali Borji, Dicky N. Sihite, and Laurent Itti. “Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study”. In: *IEEE Transactions on Image Processing* 22.1 (2013), pp. 55–69. DOI: 10.1109/TIP.2012.2214050.

[28] Mikhail Startsev and Michael Dorr. “A reduced-reference metric for visual quality evaluation of point clouds”. In: *arXiv preprint arXiv:1908.08089* (2019).

[29] Huiyu Duan et al. “Attentive deep regression networks for real-time visual face tracking in video”. In: *IEEE International Conference on Multimedia and Expo (ICME)*. 2019, pp. 1132–1137. DOI: 10.1109/ICME.2019.00196.

[30] Andrea Pera et al. “Identifying visual attention in children with autism spectrum disorder through eye-tracking”. In: *Applied Sciences* 12.3 (2022), p. 1443. DOI: 10.3390/app12031443.

[31] Mikhail Startsev et al. “Using eye-tracking data to predict the quality of first impressions”. In: *Scientific Reports* 13.1 (2023), pp. 1–13. DOI: 10.1038/s41598-023-28948-2.

[32] J. Gutiérrez et al., "Saliency4ASD: Challenge, dataset and tools for visual attention modeling for autism spectrum disorder," Signal Processing: Image Communication, vol. 92, Art. no. 116092, Jan. 2021.DOI: 10.1016/j.image.2020.116092.

[33] Guobin Wan et al. “Visual mismatch negativity is sensitive to symmetry as a perceptual category”. In: *European Journal of Neuroscience* 41.5 (2015), pp. 662–667. DOI: 10.1111/ejn.12825.

[34] Mikhail Startsev and Michael Dorr. “Classifying Autism Spectrum Disorder Based on Scanpaths and Saliency”. In: *Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. Shanghai, China,

2019, pp. 633–636. DOI: 10.1109/ICMEW.2019.00122.

[35] Giuseppe Arru, Partha Mazumdar, and Federica Battisti. “Exploiting Visual Behaviour for Autism Spectrum Disorder Identification”. In: *Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. Shanghai, China, 2019, pp. 637–640. DOI: 10.1109/ICMEW.2019.00123.

[36] Yuqing Tao and Mei-Ling Shyu. “SP-ASDNet: CNN-LSTM Based ASD Classification Model Using Observer ScanPaths”. In: *Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. Shanghai, China, 2019, pp. 641–646. DOI: 10.1109/ICMEW.2019.00124.

[37] Chia-Hung Wu et al. “Predicting Autism Diagnosis Using Image with Fixations and Synthetic Saccade Patterns”. In: *Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. Shanghai, China, 2019, pp. 647–650. DOI: 10.1109/ICMEW.2019.00125.

[38] Shiyu Xu, Jie Yan, and Min Hu. “A New Bio-Inspired Metric Based on Eye Movement Data for Classifying ASD and Typically Developing Children”. In: *Signal Processing* 178 (2021), p. 107787. DOI: 10.1016/j.sigpro.2020.107787.