

A MACHINE LEARNING APPROACH FOR DETECTING  
HIGH-FUNCTIONING AUTISM USING WEB-BASED EYE-TRACKING DATA

A THESIS SUBMITTED TO  
THE BOARD OF GRADUATE PROGRAMS  
OF  
MIDDLE EAST TECHNICAL UNIVERSITY, NORTHERN CYPRUS CAMPUS

BY

ERFAN KHALAJI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR  
THE DEGREE OF MASTER OF SCIENCE  
IN COMPUTER ENGINEERING

JULY 2021



## **ABSTRACT**

### **A MACHINE LEARNING APPROACH FOR DETECTING HIGH-FUNCTIONING AUTISM USING WEB-BASED EYE-TRACKING DATA**

Khalaji, Erfan

Master of Science, Computer Engineering Program

Supervisor: Assoc. Prof. Dr. Yeliz Yeşilada Yılmaz

Co-Supervisor: Dr. Şükrü Eraslan

JULY 2021, 88 pages

Autism Spectrum Disorder (ASD) is a heterogeneous neurodevelopmental disorder that causes social, communication and behavioral challenges with different severity levels. Studies report a considerable increase in ASD prevalence during the past two decades, and clinical psychologists face difficulties identifying individuals with ASD. Researchers have been using different techniques such as eye-tracking to help address ASD diagnosis. A previous study shows that training a logistic regression model with eye-tracking data gathered using web pages can be an effective method to identify high-functioning autism in adults. This thesis uses the same eye-tracking datasets, each of which includes two web-related tasks. The first dataset includes a searching task and a freely browsing task, while the second dataset includes a synthesis task and a time-restricted browsing task. Our study investigates a different data preprocessing method and evaluates various machine learning models based on that. This study obtains an accuracy of 91.6% and 76.3% for the searching task and the freely-browsing task, respectively. In contrast, the previous study obtains an accuracy of 75% for the search task and 71% for the browse task. Our study and the previous study obtain lower accuracies using the second dataset. Therefore, the results in this thesis demonstrate that tasks and pages used can affect the accuracy of machine learning algorithms. This thesis also suggests eye-tracking on the web can be a complementary practical tool for human experts to diagnose people with ASD more precisely.

Keywords: Autism Spectrum Disorder, Machine Learning, Eye-Tracking, Web

## ÖZ

### WEB TABANLI GÖZ İZLEME TARİHİ KULLANARAK YÜKSEK İŞLEVLİ OTİZMİ TESPİT ETMEK İÇİN BİR MAKİNE ÖĞRENME YAKLAŞIMI

Khalaji, Erfan  
Yüksek Lisans, Bilgisayar Mühendisliği Programı  
Tez Yöneticisi: Doç. Dr. Yeliz Yesilada Yılmaz  
Ortak Tez Yöneticisi: Dr. Şükrü Eraslan

Temmuz 2021, 88 pages

Otizm Spektrum Bozukluğu (OSB), farklı şiddet seviyelerinde sosyal, iletişim ve davranışsal zorluklara neden olan heterojen bir nörogelişimsel bozukluktur. Çalışmalar, son yirmi yılda OSB yaygınlığında önemli bir artış olduğunu bildirmektedir ve klinik psikologlar, OSBli bireyleri tanımlamada güçlüklerle karşılaşmaktadır. Araştırmacılar, OSB teşhisini ele almaya yardımcı olmak için göz izleme gibi farklı teknikler kullanıyorlar. Önceki bir çalışma, web sayfaları kullanılarak toplanan göz izleme verileriyle bir lojistik regresyon modeli eğitmenin, yüksek işlevli otizmi tanımlamak için etkili bir yöntem olabileceğini göstermektedir. Her biri web ile ilgili iki görev içeren ve aynı göz izleme veri kümeleri bu araştırmada kullanılacaktır. İlk veri seti, bir arama görevi ve serbestçe göz atma görevi içerirken, ikinci veri seti bir sentez görevi ve zaman kısıtlı bir tarama görevi içerir. Çalışmamız, farklı bir veri ön işleme yöntemini araştırmakta ve buna bağlı olarak çeşitli makine öğrenimi modellerini değerlendirmektedir. En iyi durumlarda, bu çalışma arama görevi ve serbestçe gezinme görevi için sırasıyla % 91,6 ve %76,3 doğruluk elde etmektedir. Bunun aksine, önceki çalışma arama görevi için %75 ve göz atma görevi için %71 doğruluk elde ediyor. Her iki çalışma da ikinci veri setini kullanarak daha düşük doğruluklar elde etmektedir. Bu nedenle, bu tezdeki sonuçlar, kullanılan görevlerin ve sayfaların makine öğrenimi algoritmalarının performansını etkileyebileceğini göstermektedir. Bu tez ayrıca, web sayfalarını kullanan göz izleme verilerinin, insan uzmanların OSBli kişileri daha kesin bir şekilde teşhis etmeleri için tamamlayıcı bir pratik araç olabileceğini önermektedir.

Anahtar Kelimeler: Otizm Spektrum Bozukluğu, Makine Öğrenimi, Göz İzleme, Web

I dedicate this thesis to people with autism spectrum disorder.



## **ACKNOWLEDGMENTS**

My deep and sincere gratitude to my parents for their unconditional love and help. I am grateful to my dearest and best friends, Nastaran, Alireza, Nathalie and Tina for their utmost support. I am thankful to Dr. Yesilada and Dr. Eraslan for helping me grow as a researcher. I am forever in love with our rescued animals as they changed my life for the better.

July 2021

Erfan Khalaji





## TABLE OF CONTENTS

ABSTRACT . . . . .	vi
ÖZ . . . . .	ix
ACKNOWLEDGMENTS . . . . .	xiii
TABLE OF CONTENTS . . . . .	xv
LIST OF TABLES . . . . .	xviii
LIST OF FIGURES . . . . .	xxii
LIST OF ABBREVIATIONS . . . . .	xxiv
CHAPTERS	
1 INTRODUCTION . . . . .	1
1.1 Aims and Objectives . . . . .	4
1.2 Contributions . . . . .	4
1.3 Thesis Outline . . . . .	5
2 LITERATURE REVIEW . . . . .	7
2.1 Autism Spectrum Disorder . . . . .	7
2.2 Eye-Tracking and Applications . . . . .	8
2.3 Eye-Tracking for Studying Autism . . . . .	9
2.4 Autism Detection With Eye-Tracking . . . . .	11
2.5 Autism Detection with Other Data Types . . . . .	17

2.6	Lessons Learned . . . . .	19
3	METHODOLOGY . . . . .	22
3.1	The First Dataset . . . . .	22
3.1.1	Participants . . . . .	22
3.1.2	Materials and Apparatus . . . . .	23
3.1.3	Tasks . . . . .	24
3.1.4	Procedure . . . . .	24
3.2	The Second Dataset . . . . .	24
3.2.1	Participants . . . . .	25
3.2.2	Materials and Apparatus . . . . .	26
3.2.3	Tasks . . . . .	26
3.2.4	Procedure . . . . .	26
3.3	Areas of Interests (AOI) . . . . .	26
3.3.1	Page-Specific AOIs . . . . .	27
3.3.2	Grid AOIs . . . . .	28
3.4	The Overall Method . . . . .	28
3.4.1	Data Preparation . . . . .	30
3.4.2	Feature Selection . . . . .	34
3.4.3	Model Development . . . . .	34
3.4.4	Model Evaluation . . . . .	37
4	RESULTS . . . . .	39
4.1	Exploratory Analysis . . . . .	39
4.2	Statistical Analysis . . . . .	43

4.3	Classification . . . . .	48
5	DISCUSSION AND CONCLUSION . . . . .	56
5.1	Discussion . . . . .	56
5.2	Conclusion . . . . .	62
APPENDICES		
A	APPENDIX . . . . .	64
A.1	Experiment Settings . . . . .	64
	REFERENCES . . . . .	76

## LIST OF TABLES

### TABLES

Table 2.1	Summary of our reviewed studies that use eye-tracking data for ASD detection. Evaluation metrics are reported in percent. NA stands for not applicable. In the sample size column, in each row, the first and the second values indicate the number of participants with ASD and TD, respectively. . . . .	12
Table 2.2	A summary of some of studies that use other techniques than eye-tracking for autism detection. . . . .	18
Table 3.1	The questions that the participants were asked for the searching task in the first dataset [1]. . . . .	25
Table 3.2	The web pages used in the second dataset with their level of distinguishability (D) and complexity (C), and the synthesis questions [2]. . . .	27
Table 3.3	Schematic of the initial dataset that this study uses. . . . .	33
Table 3.4	This table presents the schematic of the search dataset after preprocessing with a page-specific approach. Note that 0 represents False, and 1 represents True. If the value of Central is 0 for a record, it means that the segment was peripheral. . . . .	33
Table 3.5	This table shows the schematic of the search dataset after preprocessing and encoded with a one-hot approach. . . . .	34
Table 4.1	Significant features with respect to the total number of fixations for all web pages. . . . .	46

Table 4.2	Significant features with respect to the total fixation duration for all web pages. . . . .	47
Table 4.3	List of classifiers with the best accuracy in each experiment. The abbreviations used include XY: gaze point location, C: CNT is an attribute of the eye tracking dataset. Whenever a packet is successfully transferred (including a fixation record), CNT increases. Since it has 99% correlation with TIME, we considered CNT as the timestamp in our early experiments. S: page-specific AOI segments. T: actual timestamp, a feature of the raw eye tracking dataset which represents the timestamp. 22G: 2*2 Grid segmentation. 33G: 3*3 Grid segmentation. The best column also highlights the classifier with the highest performance accuracy in each experiment. . . . .	48
Table 4.4	The classification results using the searching task and 3*3 grid segmentation before and after feature selection. The numerical value out of the parentheses shows the result after feature selection, and the value reported within parentheses shows the result before feature selection. . . . .	53
Table 5.1	The difference between the number of data points of the first dataset used in [1] compared to the current study. . . . .	60
Table A.1	First experiment: The classification results using the browsing task dataset with gaze coordinates and media name features. . . . .	64
Table A.2	Second experiment: The classification results using the browsing task dataset with gaze coordinates, media name and CNT features. . . . .	65
Table A.3	Third experiment: The classification results using the searching task dataset with gaze coordinates and media name. . . . .	65
Table A.4	Fourth experiment: The classification results using the searching task dataset with gaze coordinates, media name and CNT features. . . . .	65
Table A.5	Fifth experiment: The classification results using the browsing task dataset with page-specific AOI segments, central and media name. . . . .	66

Table A.6 Sixth experiment: The classification results using the browsing task dataset with page-specific AOI segments, central, CNT and media name. . . . .	66
Table A.7 Seventh experiment: The classification results using the searching task dataset with page-specific AOI segments, media name, central and relevant. . . . .	67
Table A.8 Eighth experiment: The classification results using the searching task dataset with page-specific AOI segments, media name, central, relevant and CNT. . . . .	67
Table A.9 Ninth experiment: The classification results using the browsing task dataset with page-specific AOI segments, media name, central and timestamp. . . . .	68
Table A.10 Tenth experiment: The classification results using the searching task dataset with page-specific AOI segments, media name, central, relevant and TIME. . . . .	68
Table A.11 The list of the parameters with different values used in different MLP trials. . . . .	69
Table A.12 Eleventh experiment: using fully connected neural networks and the searching task dataset with page-specific AOI segment features. . . . .	70
Table A.13 Twelfth experiment: using fully connected neural networks and the browsing task dataset with page-specific AOI segment features. . . . .	70
Table A.14 Thirteenth experiment: classification results using the browsing task while only the gathered data on AVG and Apple web pages are used. . . . .	71
Table A.15 Fourteen experiment: classification results using the searching task while only the gathered data on Apple and Babylone web pages are used. . . . .	71
Table A.16 Fifteenth experiment: The classification results using the browsing task dataset with 2*2 grid segmentation, media name and timestamp. . . . .	71

Table A.17 Sixteenth experiment: The classification results using the searching task dataset with 2*2 grid segmentation, media name and timestamp. . . . .	72
Table A.18 Seventeenth experiment: The classification results using the searching task dataset with 3*3 grid segmentation, media name and timestamp. . . . .	72
Table A.19 Eighteenth experiment: The classification results using the browsing task dataset with 3*3 grid segmentation, media name and timestamp. . . . .	73
Table A.20 Nineteenth experiment: The sanity check for the obtained results using the searching task dataset. . . . .	74
Table A.21 Twentieth experiment: The sanity check for the obtained results using the browsing task dataset. . . . .	74
Table A.22 Twenty first experiment: The accuracy of the three best classifiers using the searching task data with respect to one page at a time. The values inside parentheses is the F1 score. . . . .	74
Table A.23 Twenty second experiment: The accuracy of the three best classifiers using the browsing task data with respect to one page at a time. The values inside parentheses is the F1 score. . . . .	74
Table A.24 Model accuracy percentage (%) of Decision Tree (DT), Random Forests (RF) and Extra Trees (ET) algorithm using individual web pages of the second dataset for both tasks. The names of web pages in the second study are Adobe (AD), Amazon (AZ), BBC (BC), Netflix (NF), Outlook (OL), Whatsapp (WA), Youtube (YT) and Wordpress (WP). . . . .	74
Table A.25 Model accuracy percentage (%) of Decision Tree (DT), Random Forests (RF), Extra Trees (ET) algorithm using 2*2 and 4*4 grid segmentation. The values inside the parenthesis show the F1 score. The results are reported for both the synthesis and the browsing task. . . . .	75

## LIST OF FIGURES

### FIGURES

Figure 1.1	ASD prevalence reported by CDC from 2000 to 2016 [3]. . . . .	1
Figure 3.1	Page specific AOIs. The figure is taken from [1]. . . . .	29
Figure 3.2	Grid AOIs. The figure is taken from [1]. . . . .	29
Figure 3.3	This figure shows the process model of the overall method. . . . .	30
Figure 3.4	This figure shows the flow-chart diagram of the labeling approach for the page-specific AOI segmentation. . . . .	32
Figure 3.5	This figure shows the process model of the labeling approach for the grid AOI segmentation. . . . .	32
Figure 4.1	The total number of fixations on the segments of all the pages for the ASD and TD groups with regards to the searching task. . . . .	41
Figure 4.2	The total number of fixations on the segments of all the pages for the ASD and TD groups with regards to the browsing task in the first dataset. . . . .	41
Figure 4.3	The total number of fixations on the segments of all the pages for the ASD and TD groups with regards to the synthesis task. . . . .	42
Figure 4.4	The total number of fixations on the segments of all the pages for the ASD and TD groups with regards to the browsing task in the second dataset. . . . .	42



Figure 4.5	The minimum and maximum amount of time spent on the segments of all pages by ASD and TD participants in the searching task. . .	44
Figure 4.6	The minimum and maximum amount of time spent on the segments of all pages by ASD and TD participants in the browsing task of the first dataset. . . . .	44
Figure 4.7	The minimum and maximum amount of time spent on the segments of all pages by ASD and TD participants in the synthesis task. . .	45
Figure 4.8	The minimum and maximum amount of time spent on the segments of all pages by ASD and TD participants in the browsing task of the second dataset. . . . .	45

## LIST OF ABBREVIATIONS

### ABBREVIATIONS

AB	AdaBoost
AD	Adobe
ADI	Autism Diagnosis Intervie
ADOS	Autism Diagnostic Observation Schedul
ANN	Artificial Neural Network
AQ	Autism Quotient
ASD	Autism Spectrum Disorder
AZ	Amazon
BNB	Bernoulli Naive Bayes
CARS	Childhood Autism Rating Scale
CDC	Centers for Disease Control
DOC	Degree of Cohesion
DOM	Document Object Model
DT	Decision Tree
DSM	Diagnostic and Statistical Manua
ET	Extra Trees
GARS	Gilliam Autism Rating Scale
GB	Gradient Boosting
GNB	Gaussian Naive Bayes
GP	Gaussian Process
KNN	K-Nearest Neighbors

LDA	Linear Discriminant Analysis
LR	Logistic Regression
LSTM	Long-Short Term Memory
OL	Outlook
QDA	Quadratic Discriminant Analysis
RF	Random Forests
ST	
STA	Scanpath Trend Analysis
SVM	Support Vector Machine
TD	Typically Developing
WA	Whatsapp
WHO	World Health Organization
WP	Wordpress
YT	Youtube
1FV	1-Fold Validation

## CHAPTER 1

### INTRODUCTION

Autism Spectrum Disorder (ASD) is a complex heterogeneous neurodevelopmental disability that may cause behavioral, social, or communication challenges or difficulties [4]. The term “spectrum” indicates that the severity of the symptoms varies in individuals, and symptoms also differ due to the heterogeneity of the conditions. The centers for disease control (CDC) reports that the prevalence of autism has been increasing during the past two decades [3]. According to this study, in 2000, in every 150 children, one is diagnosed with ASD in the US. In 2014 one in every 59 children and in 2016, one in every 54 children is diagnosed with ASD, as illustrated in Figure 1.1.

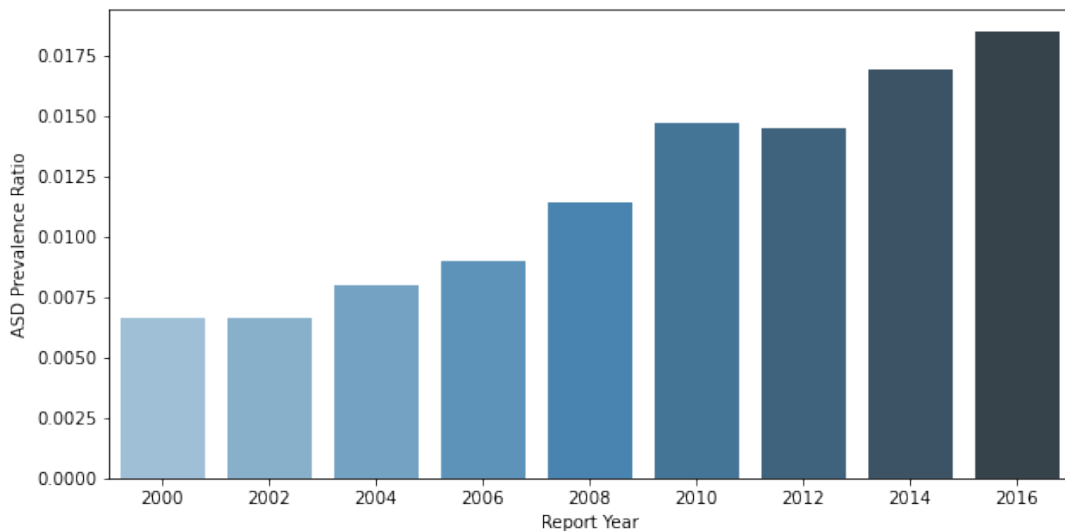


Figure 1.1 ASD prevalence reported by CDC from 2000 to 2016 [3].

For the past two decades, researchers have been using various methods to help diagnose people with ASD and understand what may cause this issue in individuals. Even

though for the majority diagnosed with ASD the reasons are still unknown, scientists believe that environmental factors and genes are likely to be playing a role in causing autism. However, it is assumed that no particular environmental factors have been identified to be causing autism yet [5]. Twin studies suggest that the ASD can be inheritable, and that is why it is assumed that genes play a pivotal role in causing ASD [6].

Methods used for diagnosing ASD differ based on the subject's age and the severity of the symptoms. In general, it is challenging to diagnose autism and there is no medical test for this purpose. There are various tools which are proposed to help diagnose people with ASD through assessments. However, CDC states that these tools should not be used alone, but a combination of these tools should be employed to avoid the diagnosis results to get biased [7]. One of the tools used by clinical psychologists is Autism Diagnosis Interview- Revised (ADI-R) [8]. This tool is used to assess autism in both adults and children, and its main focus is on behavioral areas such as communication. Autism Diagnostic Observation Schedule– Generic (ADOS-G) is another type of assessment for diagnosing people with ASD, and it focuses on the imaginative use of materials, play and communication [9]. Childhood Autism Rating Scale (CARS) is another method to focus on particular characteristics such as ability or behavior to briefly assess ASD [10]. Gilliam Autism Rating Scale – Second Edition (GARS-2) is an approach that helps clinicians, parents, and teachers identify individuals with ASD [11] and estimate the level of severity by analyzing the different aspects of the subjects personality, such as the ability to get involved in social interactions. The other tool for diagnosing autism is the American Psychiatric Association's Diagnostic and Statistical Manual, Fifth Edition (DSM-5), which tries to use criteria such as nonverbal communicative behaviors to diagnose people with ASD [12].

Due to the complex nature of the ASD, different levels of severity and the variety of the symptoms, the assessments may result in misdiagnosis, which is more common for diagnosing adults with ASD [13]. Therefore, researchers have begun to use other techniques to study ASD, including eye-tracking, electroencephalogram (EEG), and functional magnetic resonance imaging (fMRI). Eye-tracking is the process of estimating and recording the point of gaze coordinates or the motion of the eyes, and

the device used for this purpose is an eye tracker. EEG equipment is used to capture the electrical activity of the brain [14]. fMRI data also demonstrates brain activity through the blood flow [15]. While nowadays eye-trackers are portable and can be assembled on computers, fMRI and EEG equipments are relatively more expensive and they are not portable. The data that eye-trackers provide can be structured as tabular data consisting of numerical or categorical values, whereas fMRI data may consist of embedded time series which requires extra computational processes to extract potentially useful features for further use. While the volume of eye-tracking data makes it more suitable to train statistical learning models, fMRI and EEG data can be used to train both deep learning networks and statistical learning models since they provide a considerable greater volume of data.

Given that people with ASD may experience difficulties or challenges on social occasions or avoid paying attention to a particular object, most of the studies investigating autism detection by eye-tracking used videos or photos containing facial, social, or naturalistic features [16, 17, 18]. However, web pages can contain different features including web elements, and given the fact we interact with web pages on a daily basis, the web may be an effective alternative in investigating ASD using eye trackers. This thesis develops a methodology based on eye-tracking data that was previously collected while participants interacted with web pages to train machine learning models for detecting high functioning autism in adults.

The current exploration uses gaze records gathered from both typically developing (TD) and ASD subjects while they looked at several web pages and tried to complete web-related tasks. The first motivation behind using web pages is that they are widely available, and the web has become an inseparable aspect of our daily life. Therefore, web pages seem to be a useful tool for investigating ASD, given the sharp increase in the ASD prevalence and the popularity of using the web. Secondly, there is a limited number of studies that use web pages to investigate how people with ASD interact with web pages, and given the high rate of the ASD prevalence, we may expect web pages to be an effective way of studying autism.

## 1.1 Aims and Objectives

This study aims to investigate detecting high functioning autism in adults using eye-tracking data gathered the Web, and this thesis has three main objectives.

- The first objective of this study is to grasp a better understanding of how people with ASD visually process the web pages and then compare the findings with the data obtained from TD people. Given that data preparation can profoundly change the performance of machine learning models, having an efficient data preparation approach becomes first priority to improve predictive model accuracy compared to [1].
- The second objective of this study is to investigate which predictive model performs better. In addition to statistical learning approaches, we also explore the performance of multilayer perceptron models to observe the effects of using various machine learning models for detecting high functioning autism. To compare our results with the previous studies [1, 19] in a systematic manner, we use the same evaluation method.
- The third objective is that since the eye-tracking data we use in this study is gathered while performing web-related tasks, as our last objective, it is essential to explore the effects of tasks on the performance of the predictive models. Hence, we use the method we develop through the previous stages to detect autism using a different dataset which is initially collected and used in [2].

## 1.2 Contributions

Our first contribution in this study is using web pages to detect high functioning autism unlike the majority of the literature which uses images with facial, social, or naturalistic features.

This study uses a different data preprocessing approach compared to [1]. At this stage of the project, we demonstrate that in order to generate web pages segments to understand where subjects look at, we do not necessarily need to identify page-

specific Areas of Interest (AOI) for each web pages. We obtained the best results when we used grid AOI, and this also led to the dropping of fewer data points in our dataset.

In the best case, we can detect high functioning autism in adults with an accuracy of 91.6% using a Decision Tree model. Comparing this value to previous studies' model performance, we obtain the highest model accuracy using the same dataset. The evaluation we used in this study is as same as the one that [1] used. The improvement in this study is using a different data processing approach that highlights the differences between the gaze records of ASD people compared to the TD subjects. In contrast with [1], this thesis focuses on avoiding extra computations to interpret the raw data and generate a new dataset based on it. Furthermore, we explore the performance of various predictive models. In [1], the best-reported accuracy is 75%, and because we obtained a higher model performance, it can be considered a contribution of this study. However, in this study, we investigate the performance of 14 different learning models, and we report the results we obtained systematically to compare them. Among these models, we show that Decision Tree, Random Forest, and Extremely-Randomised Trees work better than the other models, and in the best case, we obtained 91.6% model accuracy in detecting participants with high functioning autism. Furthermore, we show that multilayer perceptrons do not work effectively as statistical learning models using our data. The insights we provide regarding the performance of machine learning models using eye-tracking data can help future studies, and it can be considered as a contribution to the field.

### **1.3 Thesis Outline**

**Chapter 2 Background and Related Work** reviews the history of eye-tracking and its applications in different fields of studying, analyzes and discusses articles centered around using machine learning and deep learning to address the ASD detection using various types of data, and highlights the gaps in the literature.

**Chapter 3 Methodology** explains the materials, apparatus, and procedure used to collect the eye-tracking data initially used in [20], also presents the method that this



study develops for ASD diagnosis through machine learning models and statistical data analysis.

**Chapter 4 Results** highlights the differences in visual processing patterns of people with high functioning autism versus the TD group. Furthermore, the performance of various machine learning models are evaluated and reported.

**Chapter 5 Discussion and Conclusion** aims for interpreting the obtained results of this thesis, highlighting the importance of ASD diagnosis using machine learning models and explaining limitations as well as future directions that can be investigated further as future work.

## CHAPTER 2

### LITERATURE REVIEW

In this chapter, after defining and describing autism spectrum disorder (ASD), we review the history of eye-tracking and its applications in different fields of studying, including investigating ASD. Next, we analyze and discuss articles with a focus on using machine learning and deep learning to address ASD detection using various types of data, summarized in Table 2.1 and Table 2.2. Lastly, we discuss literature's gaps and findings, as well as stating our contribution to this field of study.

#### 2.1 Autism Spectrum Disorder

ASD is a heterogeneous neurodevelopmental disorder [21]. Study presented in [22] shows that neurodevelopmental disorders are multifaceted conditions which are distinguishable by observing impairments in cognition, communication and behavior. The heterogeneity of ASD is due to its diverse symptoms and characteristics. Studies introduce numerous symptoms for ASD. For example, [23] mentioned occasional social orienting, infrequent initiation of social interactions, repetitive motor behaviors, weak engagement with the surrounding people and a limited number of gestures as early signs of ASD. The symptoms may be noticed by parents when children are younger than three years [12]. The symptoms that some children with autism may experience possibly develop throughout time. Therefore, their social and communicational skills may keep getting problematic as they begin to grow older [24].

According to the estimates of the Centers for Disease Control and Prevention across the USA, in 2016, about one in 54 children has been identified with ASD [3]. Furthermore, in the year 2019, the world health organization (WHO) mentioned that over

the last 50 years, the reported number of children with autism has been growing [25]. It also stated that some people with autism are capable of living independently, while some of them experience severe disabilities and require life-long care and support.

## **2.2 Eye-Tracking and Applications**

The process of capturing the eye movements is known as eye-tracking. In eye-tracking, the gaze information of a person looking at a particular object or perhaps freely observing the surrounding environment is recorded. An eye tracker is a machine that aims to record the gaze data. Eye trackers work based on detecting near-infrared light reflections, which is directed toward the pupil (the center of the eyes) and the outer-most optical element of the eye (the cornea). The mentioned types of reflections are tracked using an infrared camera, also known as pupil center corneal reflection [26].

An eye tracker is capable of recording eye fixations and saccades. A fixation record occurs when a participant stares at a particular location for a certain amount of time, while a saccade represents the transition between two fixations. Researchers use these features in different fields of research, such as visual marketing [27] and neuroscience [28]. Though, [29] stated that perhaps the most well-known application of eye-trackers is to study human visual attention in psychology. [30] mentioned that the fixation duration in eye-tracking data can be used as a parameter for detecting how much attention a particular point has attracted.

Applying eye-tracking for understanding human behavior has been studied for many decades. However, researchers usually recruited neurotically developed (TD) people [31]. An early study that utilized eye-tracking technology engaged three people to explore their gaze patterns while watching black and white photos of the human face [32]. For example, in one task, the participants were asked to match a previously seen face with a new photo. The study shows that each subject has different features of interest. Some of them had more fixations on the eyes, whereas some fixated their eyes more on the lips and nose, which are categorized as core features. As another example, a study shows that semantic consistency can change the gaze patterns of

an individual [33]. In other words, it can affect the saccades and fixation attributes in terms of duration and numbers. Other example of using eye-tracking is the abnormality analysis of eye movements in people with schizophrenia<sup>1</sup>. The authors of [35] stated that they employed 17 adults with attention deficit hyperactivity disorder (ADHD)<sup>2</sup>, 49 adults with schizophrenia and 37 adults without any specific cognitive condition to study their frequency of anticipatory and eye transitions. The results show that these variables are considerably different in the subjects with schizophrenia, whereas, in healthy and ADHD subjects, the recorded data appears to be normal. In the past two decades, using eye trackers to investigate autism has attracted the attention of researchers.

Autism can affect individual eye movements even in the early stages. For example, [37] examined a group of toddlers to study and understand whether they follow the gaze of others or not. The outcomes suggest that in naturalistic occasions, ASD toddlers did not show a similar gaze pattern as TD toddlers did. Furthermore, [30] investigated joint attention, which is the capacity to coordinate visual attention with another person and then shift the gaze toward a shared object or event. The authors stated that joint attention is known to be insufficient in many people with autism.

In summary, eye-tracking is used to study human behavior, and this technology can be used to investigate ASD since the visual attention of people with autism may have differences compared to TD people and eye-tracking appear to be capable of exhibiting these differences.

### **2.3 Eye-Tracking for Studying Autism**

The number of studies that use eye-tracking for studying autism began to rise in early 2000. One of these studies analyzed the eye movement of five adults with autism and five TD participants [38]. The task included recognizing the emotion of people in images. This study shows that the participants with autism spent less time focusing on eyes, nose and lips (the core features) while the control group tended to observe

---

<sup>1</sup> Schizophrenia is a chronic mental disorder. A person with schizophrenia may have difficulties concentrating. The other symptoms can include delusions or hallucinations [34].

<sup>2</sup> ADHD is a mental condition. An individual with ADHD may have differences in brain activities and the ability to concentrate and sit on a spot continuously [36].

the core features.

In another study, [17] used eye-tracking data to analyze the visual behavior of subjects with Williams Syndrome (WS)<sup>3</sup> and autism. People with autism seemingly avoid engaging in social interactions, while WS subjects are known to be hyper-social. In this study, the data was collected from 18 participants with WS and 26 people with autism. There are two different tasks, and one of them is looking at photos with faces embedded in the image and looking at a scrambled image. While they are looking at these images, an eye tracker recorded the gaze data. The results show that WS participants had long fixations on the face both in the embedded version and the scrambled image. However, subjects with autism tried to avoid the facial features in both photos; they had short fixations and long transitions and preferred to analyze other attributes in the image.

Eye-tracking has been usually employed to investigate the gaze patterns of people with autism while they observe social or facial scenes as well as naturalistic objects [16, 17, 18]. TD people have more fixations on core features of a naturalistic view, such as a face or a landscape. However, people with autism behave opposite and tend to have less or no fixations on core features such as mouth [40]. [41] stated that ASD people tend to fixate their eyes away from the center of the image and have longer transitions (saccadic records). However, [42] investigates the gaze patterns of 19 controls and ASD subjects by asking them to watch natural scenes freely. The results suggest that people with ASD concentrated more on the center of the image, had fewer fixations on objects in the image, and they seemed to have atypical preferences for specific purposes that appear in the image. For example, by looking at one image, which includes a tree, an adult elephant, and an infant elephant, the control-group participants mainly looked at the elephants' faces. However, the subjects with autism instead looked at a minor part of the tree in the background. This information is obtained by plotting the gaze records with a heatmap. A heatmap is a kind of visualization that illustrates which regions of an image attracted more attention than the other areas in the picture.

Some studies support using eye-tracking data collected on the web to detect people

---

<sup>3</sup> WS is a genetic condition, and it can affect different parts of the body. For example, people with WS may experience challenges in learning, face cardiovascular issues or have unique personal characteristics [39].

with autism. To be more specific, in one of the early studies based on the same data used in [43], authors analyze the scanpath of TD and ASD participants, and they stated that people with autism are less successful in finding information provided on a web page. Furthermore, they mentioned that the scanpath of ASD users is more variant compared to the TD group [20]. [44] used the same dataset and reported that ASD group put more cognitive effort to complete a set of specific tasks with the same accuracy compared to the TD participants. They concluded this since it took ASD subjects a higher amount of time to finish the tasks. [45] mentioned that ASD participants had more fixations on irrelevant visual elements, and they also tended to have longer scanpath compared to the TD users. This study also adds that the ASD subjects made fixations with a shorter duration on visual elements and longer transitions between the web elements compared to the TD participants. It demonstrates that ASD subjects employ different strategies for searching the web. [2] used another eye-tracking dataset which is also used by [19] and reported that while the ASD participants were completing a synthesis task, they looked at more irrelevant elements when they were interacting with visually complex pages or with web pages that contained elements with a poor level of distinguishability.

Overall, eye-tracking data can provide intuitive information about human behavior. Hence, researchers have begun using it to investigate autism. The studies are mostly centered around analyzing the gaze data while the participants observe photographic scenes. These studies suggest that people with autism pay less attention to the core features of an image and have longer saccades. Moreover, there are a limited number of studies that provided insightful information regarding the interaction of ASD people with web and reported solid evidence about the differences compared to TD subjects.

## **2.4 Autism Detection With Eye-Tracking**

As we earlier discussed, eye-tracking technology empowers the researchers to study human behavior and cognitive disabilities, such as autism. We also discussed that eye-tracking analysis shows that ASD people tend to look at a photo, video or a web page differently compared to TD participants. These differences can be used to train su-

ervised learning algorithms for autism detection. Table 2.1 demonstrates a summary of the literature which uses supervised learning algorithms and eye-tracking analysis to address autism detection. We analyzed the previously conducted researches based on the machine learning models and their ability, as well as the data used to address this binary classification problem, evaluation approach, the age range of recruited participants, and the amount of data they use to train their models.

Table 2.1 Summary of our reviewed studies that use eye-tracking data for ASD detection. Evaluation metrics are reported in percent. NA stands for not applicable. In the sample size column, in each row, the first and the second values indicate the number of participants with ASD and TD, respectively.

Ref	Sample Size	Age Range	Approach	Model	Evaluation	Accuracy	AUC	Sensitivity	Specificity
[46]	30-29	NA	Image	ANN	10F-CV	88.9	79.2	88.9	69.4
[47]	32-32	76.32 months	Image	ANN	10F-CV	78.13	82.91	NA	NA
[48]	30-29	on avg: 8 years	Video	ANN	10F-CV	NA	92	NA	NA
[49]	17-15	8 to 10	Video	LSTM	1FV	NA	NA	75	100
[50]	14-14	5 to 12 years	Image	CNN+LSTM	1FV	74.22	NA	NA	NA
[51]	20-19	Unclear	Image	ResNet+LSTM	LOOCV	99	1	1	98
[52]	20-19	Adults	Image	VGG-16	LOOCV	92	92	93	92
[41]	14-14	5 to 12 years	Image	CNN+RF	100 CV	NA	75	NA	NA
[53]	14-14	5 to 12 years	Image	CNN (ResNet)	1FV	65.41	69	NA	NA
[54]	14-14	5 to 12 years	Image	YOLO	1FV	59.3	59.5	NA	50.56
[55]	14-14	5 to 12 years	Image	RM3ASD	1FV	59	-	50	-
[19]	19-19	32 to 41 years	Web Page	LR	100F-CV	73	NA	NA	NA
[1]	15-15	33 to 37 years	Web Page	LR	100F-CV	75	NA	NA	NA
[56]	22-22	Adults	Video	Ensemble	10F-CV	76	NA	NA	NA
[57]	23-35	8 to 34 years	Facial Image	RF	LOOCV	86.2	93.5	91.3	82.9
[58]	29-29	4 to 11 years	Facial Image	SVM	LOOCV	88.51	89.63	93.1	86.2
[59]	49-48	3 to 6 years	Facial Image	SVM	1FV	85.44	93	NA	NA
[60]	37-37	4 to 6 years	Video	SVM	5F-CV	85.1	NA	86.5	83.8
[61]	53-136	4 to 8 years	Video	SVM	20F-CV	93.7	NA	NA	NA
[45]	15-15	33 to 37 years	Web Page	STA	100F-CV	63	NA	NA	NA

The majority of studies target children for autism detection since ASD diagnosis can be very beneficial in the early stages. [62] mentioned that evidently, early diagnosis and treatment can significantly improve the quality of life of individuals with ASD and their carers and families. On the other hand, adults with autism could be left undiagnosed, and there are limited studies that focus on investigating high functioning autism in adults, such as as [1, 19, 56, 52]. There are also studies in which the age of the participants varies from eight to 34 years (e.g., [57]) or two to 60 years (e.g.,[63]). The rest of the studies target toddlers (e.g., [59, 60, 58]) infants (e.g.,

[47]) and children (e.g., [54, 49]).

Researchers typically employ supervised learning algorithms for autism detection. The data given to supervised learning functions are labeled according to the supervision of humans or machines. In a supervised learning task, the goal is to train a model capable of mapping an input to a correct output, based on a set of observed examples. A supervised learning algorithm learns from the training data, and consequently, it provides an inferred function that can be employed for mapping new data records [64]. A supervised learning model can be based on statistical models or neural networks. Deep learning models are more suitable to be used when the data includes videos and images due to the greater volume of the data they provide. For example, [49, 50, 51] used long-short term memory (LSTM) [65] in their study. [52] used a convolutional neural network (CNN) for the same purpose. This study uses a visual geometry group (VGG-16) model, which is used for object recognition. Deep networks are known to require high computational power and are usually time-consuming to be trained. On the other hand, some of the authors use statistical learning models. For example [59, 58, 60] use support vector machines (SVM) which is first introduced by [66], and [1] uses logistic regression (LR) [67]. Some studies such as [57, 41] use random forest (RF) algorithm [68].

The volume of the collected data ranges from 28 (14 ASD - 14 TD) [69] to 97 (49 ASD - 48 TD) [59]. However, the majority of the literature have a small number of participants in their studies. Furthermore, some of the studies do not provide demographic information about the dataset. For instance, [51] did not mention the age range and the gender of the participants recruited for their investigations. Lastly, the usage of imbalanced data can be seen in the literature. For example, [57] recruits 35 TD control subjects and 23 ASD participants, which shows the data used for classification is imbalanced. It is crucial to have the same number of participants in each class to be able to evaluate the performance of the classifier properly.

Researchers employ cross-validation to compensate for the small number of participants. For example, [1] recruited 15 ASD and 15 TD participants (30 in total), and the authors used a cross-validation method. When the participant starts interacting with the web page, many data points were generated for each person. However, [1, 19]



implement a real case scenario to evaluate their approach. They used the data from 10 participants in the training set, and 5 participants in the test set. They replicated the process 100 times, and each time, the data associated with a participant is either in the test or the training set. Although cross-validation provides better information regarding the robustness of an approach, [54, 53, 70, 50, 55] used a 1-fold validation (1FV) and did not use cross-validation. For example, [54, 55] trained the model with 300 data samples and evaluated it with 100 records. Similarly, [53] used 80% of the available data for training and 20% for the validation of their proposed learning model. While studies such as [49] stated that the data used for training and testing is 75% and 25%, respectively, [59, 61] provide insufficient information regarding the ratio of the train and the test sets. [51, 52, 57] used leave-one-out cross-validation (LOOCV) method. In this approach, the number of folds is equal to the data samples in the dataset. Hence, in each iteration, one record will be hidden, the learning function is trained with the rest of the data, and the hidden record is used for the evaluation. Finally, the authors report the statistical measures, such as the average of the recorded performance.

Researchers use different metrics to measure the performance of learning models. However, accuracy, area under the curve (AUC), sensitivity and specificity are the most common units reported in the literature. In order to describe these metrics, we first need to define the following terms:

- True Positive (**TP**): is the number of times that an ASD subject is classified correctly.
- True Negative (**TN**): is the number of times that a TD subject is classified correctly.
- False Positive (**FP**): is the number of times that an ASD subject is misclassified.
- True Positive (**FN**): is the number of times that a TD subject is misclassified.

Given TP, TN, FP and FN, we can define the following metrics:

- **Accuracy**: is the ratio of correctly classified samples per the entire records in

the test set.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (21)$$

- **AUC:** stands for the area under the ROC (receiver operating characteristics). ROC is a graphical plot demonstrating the performance of a binary classifier for its discriminative parameters. For example, in a deep neural network, the number of hidden layers could be considered a discriminative parameter, and the performance of the model might differ based on that. Therefore, for each outcome, a specific curve could be plotted. AUC is plotted based on the specificity on the x-axis and sensitivity on the y-axis of the ROC graph.
- **Specificity:** measures the ability of the classifier in identifying negative cases. It is also known as the true negative rate (TNR). Mathematically, specificity can be defined by the following formula:

$$Specificity = TNR = \frac{TN}{TN + FP} \quad (22)$$

- **Sensitivity:** measures the ability of the classifier in identifying positive cases, and it is also known as Recall or the true positive rate (TPR). The following formula is used to measure Sensitivity:

$$Sensitivity = recall = TPR = \frac{TP}{TP + FN} \quad (23)$$

- **FPR:** stands for false positive rate and is the rate of participants that were mistakenly classified as ASD. FPR can be computed using the following fraction:

$$FPR = \frac{FP}{FP + TN} \quad (24)$$

- **FNR:** stands for false negative rate and is the rate of participants that were mistakenly classified as TD. FNR can be computed using the following fraction:

$$FNR = \frac{FN}{FN + TP} \quad (25)$$

- **Precision:** is the ratio of correctly predicted positive observations to the total predicted positive observations.

$$FNR = \frac{TP}{TP + FP} \quad (26)$$

- **F1:** Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

$$F1 = \frac{2 * (Recall * Precision)}{Recall + Precision} \quad (27)$$

Reporting accuracy or AUC alone may not demonstrate the actual ability of a classifier. In healthcare applications of supervised learning methods, it is crucial to know the rate of false classifications. AUC is incapable of providing intuitive information regarding the rate of misclassifications. Hence, it is not a reliable metric for illustrating the influence of the false classification on the overall performance of the method [71].

[59] only reported AUC and accuracy based on the chosen features, but the rate of false classification is not clear. In this case, the confusion matrix can be very helpful. A confusion matrix for a binary classifier is a 2\*2 matrix which demonstrates the rate of correctly-classified and miss-classified records. [1] reported the confusion matrix and with a 100F-CV, on average, 29.4% of TD subjects are misclassified as ASD and 27.8% of ASD participants are misclassified as TD. This rate of negative classification may be considered high for a healthcare application. [43] uses scanpath trend analysis (STA) algorithm with the same dataset and reported lower accuracies ranged from 55.1% to 64.8% for each web page separately, and the authors do not report the performance of the STA model for the combination of all web pages. Although, by considering the given accuracy per each web page, the accuracy on average is approximately 60%, unlike other machine learning algorithms, the STA approach could not be applied to all the pages simultaneously. However, for some of the web pages [43] shows better performance compared to [1]. Furthermore, [43] does not report the confusion matrix, but unlike [1], [43] reported other metrics such as precision, recall and the F-score which includes sufficient information regarding the ability of the clas-

sifier. [72] provided insightful information about evaluation metrics such as precision and F-score.

The techniques and materials used to collect data from participants mostly focus on videos or photos of naturalistic scenes to the participants. For example, in [73], gaze information is recorded alongside the demographic variables such as age and gender of the participants. In this study, each subject is asked to look at videos and images. As another example, [56] used videos that show four white dots to a participant, and if a dot starts to move, a yellow arrow starts following it. The participants were supposed to observe the changes in the dots' location, and their gaze information was recorded. Likewise, [48] used the video of a presenter, and the participants were examined based on the eye contact they made and their level of focus on the video. [49, 48, 60] also used video for ASD classification. While [59, 57] used only images with facial contents, [53, 74] used images that contain naturalistic scenes including animals, indoor/outdoor objects and human figure. The only studies that used web pages for detecting high functioning autism in adults are [1, 19, 43]. The eye-tracking data used in these studies are collected while the users interacted with web pages.

In summary, we reviewed and discussed the previous studies that use eye-tracking data for autism detection. There are various approaches, each of which uses a different perspective with dissimilar data types to address ASD detection. The diversity of the collected data has empowered the researchers to approach ASD detection using cutting edge learning models. However, there are a limited number of studies that focus on investigating autism detection in adults. Moreover, there is a limited number of studies centered around exploring how people with autism interact with web pages.

## **2.5 Autism Detection with Other Data Types**

Table 2.2 shows a summary of some studies using the behavioral data excluding eye-tracking for autism detection, such as functional magnetic resonance imaging (fMRI)<sup>4</sup>

---

<sup>4</sup> fMRI data shows which regions are activated in the brain by measuring small changes that occur in the brain [15].

and electroencephalogram (EEG)<sup>5</sup>. The information provided in this section tries to explain the advantages and disadvantages of using eye-tracking versus other types of datasets for autism detection.

Table 2.2 A summary of some of studies that use other techniques than eye-tracking for autism detection.

Ref	Data Type	Sample Size (ASD-TD)	Age	Model	Evaluation	Accuracy %
[75]	fMRI	1028-1141	5 to 64	3D-CNN	5F-CV	64
[76]	fMRI	505-530	7 to 64	AE	10F-CV	67.5
[77]	EEG	10-5	5-17	RF	Unclear	93.33
[78]	Kinematic	16-14	21 - 39	SVM	LOOCV	86.7
[79]	Audio	30 - 51	Children	SVM	Unclear	66

[79] uses audio samples of both TD and ASD subjects to detect autism. The audios contain different word utterances. The authors generated ten random groups of word utterances, trained an SVM model with nine samples, and tested it on the one remaining data sample. However, it is unclear whether the authors included the data from one participant both in the training set and the test set. The authors reported the rate of misclassification in this study (FN: 19% to 45%, FP: 27% to 21%) that may be considered high for such applications. Audio data is considered unstructured, albeit easier for human-beings to interpret; it could be more challenging to analyze such data types and extract potentially useful patterns than eye-tracking data, which is tabular and structured.

fMRI or resting-state fMRI (rs-fMRI)<sup>6</sup> data have also been applied for ASD detection. The majority of these studies use ABIDE I [15] and ABIDE II [80] or a combination of these two datasets. ABIDE is a website for fMRI data exchange, and it involves 17 international sites that share previously gathered fMRI/rs-fMRI with the scientific community. [81] uses ABIDE I dataset for ASD detection. They implement a single layer perceptron on the top of an auto-encoder and reported the average accuracy of 67.5%. [15, 80] provided with a substantial amount of data which makes researchers employ deep networks to address ASD detection [75, 81, 82]. However, some studies employ classical learning algorithms using the same type of data [83, 84, 85].

<sup>5</sup> Electroencephalography is a monitoring approach to capture the electrical activity of the brain [14]

<sup>6</sup> For rs-fMRI the brain should be in a resting state [80].

Either way, training a model based on fMRI data is relatively more time consuming and requires more memory storage than tabular datasets, such as eye-tracking. Furthermore, collecting fMRI data could be more expensive compared to collecting eye-tracking data. ABIDE I and ABIDE II datasets are collected with several institutions' collaboration, which means collecting fMRI data requires consuming more resources compared to eye-tracking.

[78] uses Kinematic parameters to detect autism. In this study, the participants were asked to observe a sequence of hand movements and then try to imitate them. Based on twenty kinematic parameters, such as duration, and vertical amplitude, authors tried to classify ASD and TD participants. For each parameter, eight value is computed, and the average and standard deviation of these parameters are used. Therefore, 40 parameters, including 20 average values and 20 standard deviation values are calculated. This approach may be considered expensive in terms of computation. Because, in addition to calculating 40 parameters for 30 participants separately, authors perform statistical data analysis, and train an SVM model as well. Moreover, the evaluation approach of this study is leave-one-out cross-validation, which requires more computation to be performed, compared to N-fold cross-validation.

In another study that used EEG, [77] reported that they obtained 93% accuracy. However, they use the data of 15 participants, including 5 ASD and 10 TD. Firstly, the amount of data is considerably low. Secondly, the classes are highly imbalanced in this scale. In other words, concerning the fact that some of the data should be used for training and some testing, having in total 5 ASD participants may not be adequate to train a random forest. Furthermore, in this study, the authors did not state the training set ratio to the test set. Finally, it is also unclear whether the data from a participant in the training set is included in the test set or not.

## **2.6 Lessons Learned**

We list the gaps we found in the literature as followings:

- A limited number of studies are centered around web accessibility for people with autism and autism detection using eye-tracking gathered on the web. The

number of web users is increasing, and the need to interact with web pages is rising to satisfy the daily needs of societies, such as online shopping and ordering food and education. Therefore, it is crucial to understand how autistic people interact with the web because autism prevalence is increasing sharply.

- Many studies provide insufficient information about the performance of their proposed method. If we consider ASD detection as a healthcare application of machine learning, evaluating the models with different performance measurement tools is essential.
- The number of data samples is low in eye-tracking approaches, yet some researchers did not use the cross-validation method.
- In the studies that we reviewed, only [1, 19, 45, 56] entirely focus on autism detection in adults, and most of the articles concentrate on autism detection in children.
- Only [1, 19, 45] use eye-tracking data associated with web pages for autism detection, and using web pages would be advantageous because it provides a natural-like setting. Concerning the fact that autism prevalence is increasing, it is essential to take ASD people's ability to use web technologies into account.
- Many of the studies have unbalanced classes for ASD detection. Therefore, the reported model accuracies of such analyses may not be reliable. For fair comparison, we need an equal number of participants in both groups to judge the ability of a classifiers. Furthermore, alternative approaches can be explored to address the imbalanced data issue.
- fMRI data requires relatively much more memory to be stored. Gathering fMRI data is expensive, and it requires heavy preprocessing computations. Eye-tracking data is tabular, structured and requires less memory capacity to be saved. Furthermore, collecting eye-tracking data and eye tracker equipment is much cheaper compared to fMRI facilities. The same goes for collecting EEG data. In the literature, we learned that the authors perform heavy computations to generate new features for EEG and audio datasets. Whereas, in this study, we use raw dataset and the operations we perform is algorithmically

much cheaper. At last but not least, modern eye trackers are portable and can be transferred from one place to another easily while this is not the case for EEG and fMRI equipment.



## CHAPTER 3

### METHODOLOGY

This chapter explains the materials, apparatus, and procedure used by [20] to collect the data adopted for this study. We also present the method we used to develop the overall approach in this thesis. To this end, we begin with defining the dataset, following by data preparation. Then we describe the model development procedure and the method we used to evaluate our proposed model. Lastly, we describe the feature selection method we used, how it affects our model development process, and revalidate the new model.

#### 3.1 The First Dataset

The eye-tracking data used in this thesis study was initially collected to examine whether web users with ASD encounter barriers on the web or not [20]<sup>1</sup>. This section explains the type of subjects employed to gather the gaze records, the materials and apparatus used to collect the eye tracking data, and the procedure used to gather the dataset.

##### 3.1.1 Participants

The ASD participants were recruited through a charity organization located in Birmingham, the UK and the student enabling center at the University of Wolverhampton. The typically developing (TD) participants were chosen through snowball sampling. ASD people were officially diagnosed using the Autism Diagnostic Obser-

---

<sup>1</sup> Full ethical approval was obtained by the University of Wolverhampton Faculty of Arts Ethics Committee before data gathering [1].

vation Schedule (ADOS)<sup>2</sup>. In contrast, according to the Autism Quotient (AQ)<sup>3</sup> test results, none of the TD participants manifested a high density of autism-related attributes. 18 ASD and 18 TD participants were selected. ASD subjects included six females and 12 males, the statistical mean ( $\mu$ ) age in years for them was 37 with a standard deviation (SD) of 9.4, and the mean number of years spent in formal education was  $\mu=16$  with  $SD=3.33$ . TD participants included eight females and ten males, the mean age for them was  $\mu=33.66$  with  $SD=8.6$ , and the mean number of years spent in formal education was  $\mu=18.35$  with  $SD=2.47$ . Due to head movements and lack of proper calibration, the data of three ASD participants were discarded. Hence, three participants were randomly chosen, and their data were eliminated from the TD group to have the same number of subjects in each group. The remaining participants were 15 ASD (nine males and six females) and 15 TD (eight males and six females).

### 3.1.2 Materials and Apparatus

The screenshots of six web pages were used. The web pages varied in terms of visual complexity (low, medium, high), measured by ViCRAM [88]. The web pages included Apple (Low), Babylon (Low), AVG (Medium), Yahoo (Medium), Godaddy (High) and BBC (High). The web pages were randomly selected from ALEXA.com<sup>4</sup>, which lists the most visited websites, by ensuring that there is a balance among low, medium and high visual complexities. The participants were supposed to interact with these web pages while a Gazepoint GP3 video-based eye tracker<sup>5</sup> (60Hz sampling rate and accuracy of 0.5-1 degree of visual angle) was recording their gaze data. The screenshots of the web pages were illustrated on a 19 LCD monitor. The distance between eye tracker and participants was measured via a sensor integrated within the Gazepoint software, and it was approximately 65cm.

---

<sup>2</sup> ADOS is an evaluation of social interaction, communication, imagination, and play designed for use in the diagnostic examination of people referred for a possible ASD diagnostic criteria [86].

<sup>3</sup> AQ is a tool to assess the degree to which an adult with average intelligence has the characteristics associated with the ASD [87].

<sup>4</sup> <https://www.alexa.com/topsites>

<sup>5</sup> <https://www.gazept.com/>

### **3.1.3 Tasks**

The web pages were shown to both ASD and TD participants. Each group member was asked to complete two different tasks; a browsing task and a searching task that are both associated with the Marchionini search activities method, a common approach used in this field [89]. According to [89], searching and browsing task both fall under the investigative group. The participants could spend up to two minutes per page in the browsing task to look for any information that attracts their attention. They were free to read textual information and observe visual elements. Furthermore, within two minutes, they moved to the next web page whenever they felt ready. In the searching task, each participant was given 30 seconds per web page, and they were supposed to look for specific information. For each web page, they were asked two questions, and these questions are represented in Table 3.1. Notably, the questions were asked by the researcher verbally, and the participants were expected to read aloud their responses. While the users were performing the tasks, the eye tracker recorded their eye movements.

### **3.1.4 Procedure**

Participants became familiar with the purpose and procedure of the experiment and then signed a consent form. Then, the TD participants were asked to fill the Autism Quotient test for determining whether they have autistic traits or not. The demographic information such as diagnosis, age and gender were collected from the participants, and a nine-point calibration of the eye tracker was performed. After a successful calibration, each participant viewed the screenshots of six web pages. The participants were asked to perform the browsing and searching tasks.

## **3.2 The Second Dataset**

The second dataset was collected to examine the effects of distinguishability of web page elements and their visual complexity [2] on the behaviour of web users with autism. [2] is a comparative eye-tracking study with 19 TD and 19 ASD subjects on

Table 3.1 The questions that the participants were asked for the searching task in the first dataset [1].

Page	Question
Apple	(a) Can you locate the link that allows to watch the TV ads relating to iPad mini? (b) Can you locate a link labelled iPad on the main menu?
Babylon	(a) Can you locate the link that you can download the free version of Babylon? (b) Can you find and read the names of other products of Babylon
AVG	(a) Can you locate the link which you can download the free trial of AVG Internet Security 2013? (b) Can you locate the link which allows you to download AVG Antivirus Free 2013?
Yahoo	(a) Can you read the titles of the main headlines which have smaller images? (b) Can you read the first item under the News title?
GoDaddy	(a) Can you find a telephone number for technical support and read it? (b) Can you locate the text box where you can search for a new domain?
BBC	(a) Can you read the first item of Sport News? (b) Can you locate the table that shows market data under the Business title?

eight web pages with varying visual complexity and distinguishability, with synthesis and browsing tasks. The outcomes of this study demonstrate that ASD subjects have a higher number of fixations and make more transitions with synthesis tasks or the web pages with low level of distinguishability and high complexity.

### 3.2.1 Participants

In [2] authors recruited 19 ASD and 19 TD subjects using the same methods and through a charity organization. In the ASD group, there were eight females and 11 males, and in the TD group, there were 13 females and six males. The statistical average age for the ASD group was  $\mu=41.05$  with the  $STD=14.04$ , and the mean age of the TD participants was  $\mu = 32.15$  with the  $STD=9.93$ . Among the ASD participants, 11 had a higher education degree, and six had a UK equivalent of a high-school degree. The rest of the ASD participants did not report information about their level of education. 18 TD participants provided data regarding their education level, 15 of them had a higher education degree, and the rest had a UK equivalent of a high-school degree.

### **3.2.2 Materials and Apparatus**

The eye tracker used to collect the second dataset is the same as the one used for the first dataset. They categorized these web pages into four classes based on their complexity and distinguishability. Whatsapp and WordPress had low visual complexity and high distinguishability (Class 1), Netflix and Outlook had low complexity and low distinguishability (Class 2), Amazon and YouTube had high visual complexity and high distinguishability (Class 3), and finally, Adobe and BBC had high visual complexity and low distinguishability (Class 4). ViCRAM [90] was used to determine a visual complexity score (VCS) for each web page, and the distinguishability was computed based on the white spaces between the elements. The distance between the participants and the eye tracker was the same as the first dataset.

### **3.2.3 Tasks**

The participants completed two tasks on the web pages called browsing and synthesis, both associated with the Marchionini search activities model [89]. In the browsing tasks, participants viewed the web pages for 30 seconds freely. However, in the synthesis tasks, specific questions were asked. The synthesis questions were designed accordingly, and the participants had to combine multiple facts from different elements to produce new information in a maximum of 120 seconds to answer the questions. Table 3.2 shows the questions researchers asked participants for the synthesis task.

### **3.2.4 Procedure**

The procedure for collecting the second dataset was the same as the one used to gather the first dataset. See section 3.1.4.

## **3.3 Areas of Interests (AOI)**

The eye-tracking machine captures various parameters, such as the fixation position and timestamp associated with each record. A scheme of the initial raw dataset is

Table 3.2 The web pages used in the second dataset with their level of distinguishability (D) and complexity (C), and the synthesis questions [2].

Page	C	D	Question
WhatsApp	Low	High	(a) Under frequently asked questions, what topic features in both the iPhone and Android columns?
			(b) Under frequently asked questions, which sections feature topics relating to notifications?
WordPress	Low	High	(a) Which of the WordPress plans offers community support instead of email support?
			(b) What is the cheapest plan you can get that offers Email and Live Chat support?
Netflix	Low	Low	(a) Which is the cheapest plan that allows you to watch movies on your laptop, TV and tablet?
			(b) How much more would you have to pay compared to the basic plan if you wanted to have Ultra HD?
Outlook	Low	Low	(a) According to the text, is Twitter a partner app of Outlook?
			(b) The names of which apps are both mentioned in the text and presented as logos below the text?
Amazon	High	High	(a) Which item has the largest price discount measured in percentage?
			(b) Which product has been rated by the largest number of users?
YouTube	High	High	(a) Under the American football category, which videos have been posted within the last three months?
			(b) Under the NBA topic, which video has the largest number of views?
Adobe	High	Low	(a) Which is the product that is targeted to UX designers and for which students can get a discount?
			(b) How many types of clouds are offered by Adobe within this page?
BBC	High	Low	(a) According to the schedule, which Grand Prix takes place first: the Australia or the Bahrain?
			(b) Which of the following sports does not have its own tab on BBC Sport: Golf, Cricket, or Volleyball?

demonstrated in Table 3.3. In this study, we use fixation coordinates to determine which area of interest (AOI) the participants choose to observe. An AOI in this context is a visual element or a region on the web page. Similar to [1], we use both page-specific and grid AOI approach.

### 3.3.1 Page-Specific AOIs

The page-specific AOIs were initially declared in [91], and they were generated using the Vison-Based Page Segmentation (VIPS) algorithm<sup>6</sup>. The document object model (DOM) is a language-independent and cross-platform interface that is the foundation of web pages and has a tree-shape structure. DOM structure consist of nodes that represent web elements with a link to other nodes that represent relationships. VIPS requires to use DOM. In other words, the DOM structure provides the necessary information for VIPS to process including the details of all the elements used to develop a web page.

VIPS labels nodes in DOM and generates a hierarchical tree structure. The VIPS algorithm then finds visual blocks by first extracting them, separating them, and con-

<sup>6</sup> VIPS algorithm uses source code of web pages and their visual illustrations based on different granularity levels to generate segmentations [92].

structuring the content structure. The major part of the VIPS algorithm is block extraction. Each extracted block holds a value so-called as the degree of cohesion (DoC). When the DoC value is bigger than a predefined value for the block, VIPS stop segmenting the corresponding node.

Next, VIPS detects the separator between every two blocks, and a value will be assigned to each separator, which initiates the connection of two blocks. Then, a tree of blocks (different from DOM, but the root of the tree is as same as the root of DOM) will be generated. These steps will be repeated for each node, and the process continues until there is no other node to examine. The adjacent nodes can be combined and shape a new tree, which is more straightforward than DOM and can represent the generated web segments. Figure 3.1 demonstrates the apple web page segmented using the VIPS algorithm.

### **3.3.2 Grid AOIs**

Similar to [1], we use 2\*2, 3\*3 and 4\*4 grid segmentation for the grid AOI approach. 2\*2, 3\*3 and 4\*4 grid segmentation will result in having 4, 9 and 16 segments, respectively. For instance, through the 3\*3 segmentation approach, the web page will be divided into nine equal segments, and each segment will be considered as an AOI. Figure 3.2 illustrates the apple web page, which is segmented using a 2\*2 segmentation approach.

## **3.4 The Overall Method**

This section explains the data preparation, statistical tests to identify the most contributing features, binary classification model, and the evaluation process. Figure 3.3 shows the process model of the overall method.

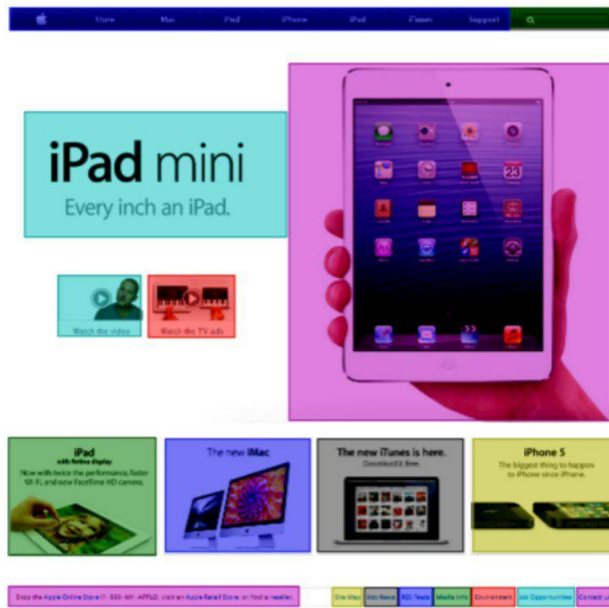


Figure 3.1 Page specific AOIs. The figure is taken from [1].



Figure 3.2 Grid AOIs. The figure is taken from [1].



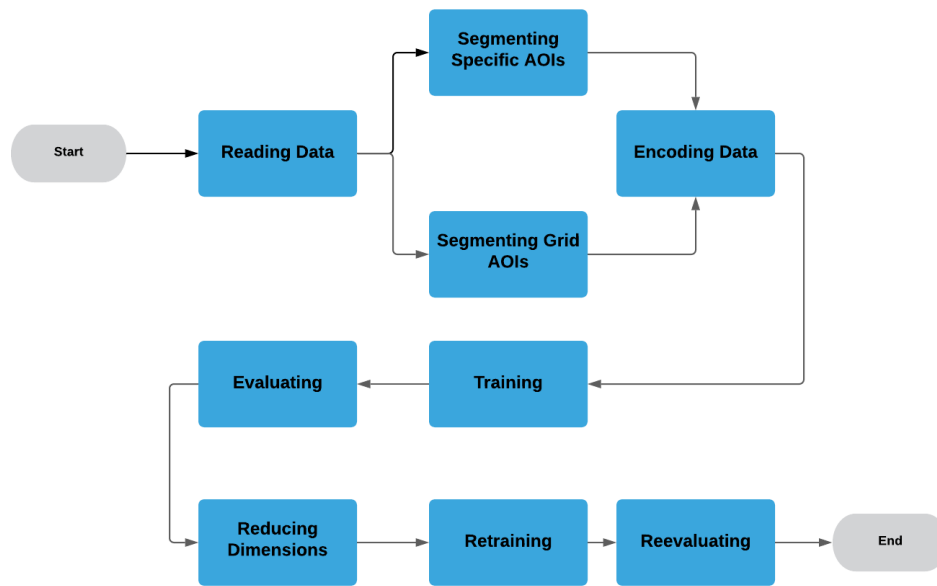


Figure 3.3 This figure shows the process model of the overall method.

### 3.4.1 Data Preparation

Data preparation is necessary to manipulate the data in an interpretable and understandable manner for supervised learning algorithms.

The eye-tracking data we use in this study include the coordinates of fixations and their duration. [1] counts the number of times a particular AOI (web segment) is visited and the total amount of duration that a participant spent looking at a segment. However, this approach would lead to the following issues:

- Performing more computation to prepare a new dataset based on the raw dataset.
- The data does not represent a sequence. However, eye-tracking data represents scanpath, which has a sequential structure. In other words, the generated dataset loses its sequential attributes.

We took a step back and considered another path for data preparation by performing less computation and keeping the sequential attribute of the dataset intact. We did so by using the raw data instead of generating a new dataset. We first extracted the fixation coordinate and their timestamps demonstrated in Table 3.3. Then, we chose

to label the data records with the ID of the segment. The possible options are page-specific segmentation and grids segmentation. Figure 3.4 illustrates a more detailed process model of the page-specific segmentation procedure.

For the page-specific segmentation, we considered whether a particular segment is related to the question (for the searching task). Furthermore, we examined if the segment is central (located at the center of the web page) or peripheral (located on the sides of the web page). Moreover, given a searching task, some segments would be relevant to the task while others would be irrelevant, and we also used this as a feature in our dataset. Therefore, with the page-specific approach, the search dataset is preprocessed differently. Note that there will be no central/peripheral and relevant/irrelevant attribute in the browsing dataset as there is no specific question for this task. The motivation behind generating central and relevant features is to investigate whether they help the accuracy of the classifier increases. Table 3.4 shows a scheme of the dataset after detecting the segments for the searching task.

According to Figure 3.5, with the grid segmentation approach, fewer features will be generated, and the complexity of this approach is lower as well. It is worth noting that the second dataset is used to evaluate the machine learning model regarding the 3\*3 segmentation because we obtained the best results using the first dataset with this approach. Therefore, we did not perform the page-specific data preparation for this dataset. We developed the model based on the first dataset, and we used the second dataset to evaluate the approach we developed with the first dataset. Hence, 3\*3 grid segmentation is used to prepare the second dataset, and it should be taken into account that there is no searching task in the second dataset. Instead, there is a synthesis task that is entirely different from the searching task. Another motivation to evaluate our model using the second dataset is to understand the classifiers' effects on distinct tasks. As the results using grid segmentation showed the highest performance, we also considered to try 2\*2 and 4\*4 segmentation to compare the obtained results with [19].

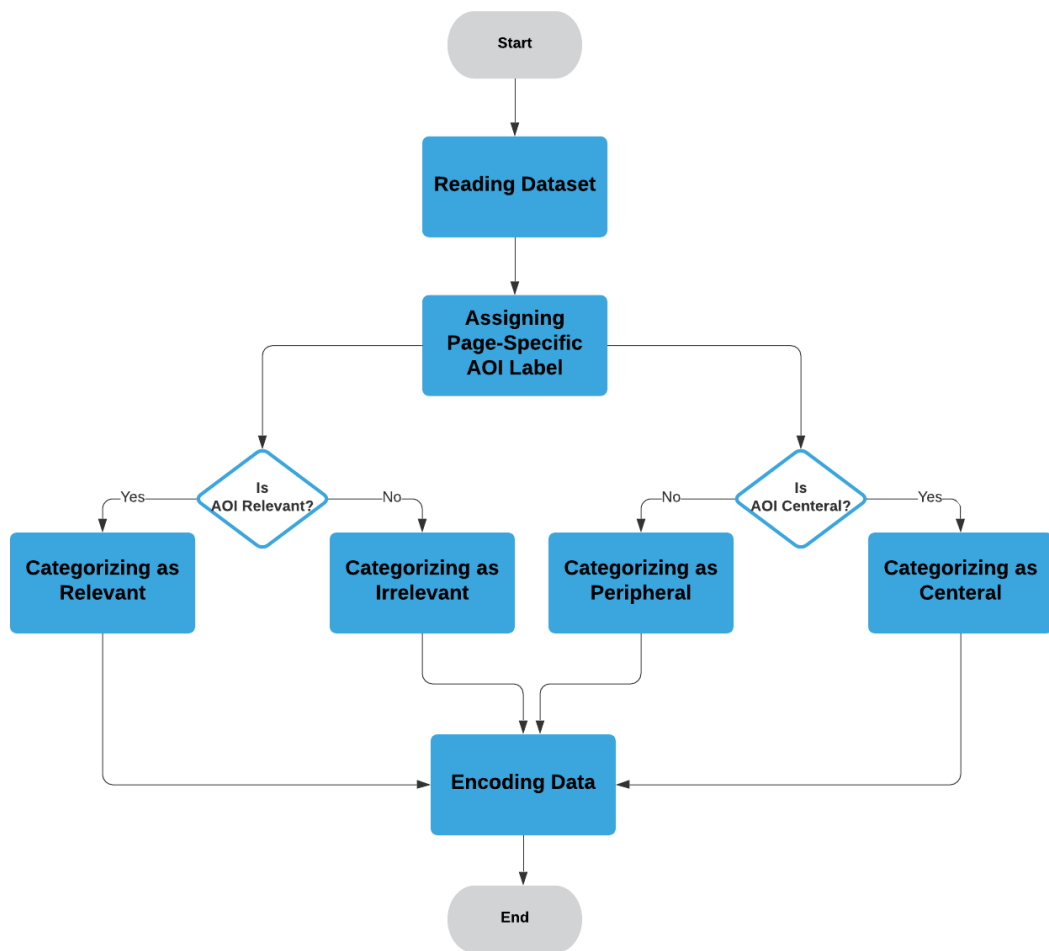


Figure 3.4 This figure shows the flow-chart diagram of the labeling approach for the page-specific AOI segmentation.



Figure 3.5 This figure shows the process model of the labeling approach for the grid AOI segmentation.

Table 3.3 Schematic of the initial dataset that this study uses.

Media Name	X	Y	Time
Apple	0.1298	0.6356	41
Apple	0.6325	0.7452	135
AVG	0.1452	0.3648	265
AVG	0.1347	0.0312	32
BBC	0.3564	0.7896	635
BBC	0.7895	0.1235	1253

To prepare the data, we use a binary one-hot encoding approach [93]. One-hot encoding is a method often used for demonstrating the state of a variable. For example, when we consider the web page column, it can have several values, such as BBC, AVG and Yahoo. With a binary one-hot encoding approach, each value becomes a separated column in the dataset. Then, if the web page value in the original dataset is, for example, BBC, in the one-hot-encoded dataset, in column BBC, we will have 1, and the rest of the columns associated with the value of the web page will be set to 0. The same approach will be used for segments. We do not replace the categorical names with integer values because, in our case, a supervised learning algorithm may learn an incremental pattern. However, when each web page becomes a separated column, it introduces a new dimension to the problem, but the learning model will not be misled by integer values representing categories. Table 3.5 demonstrates the scheme of the encoded version of Table 3.4.

Table 3.4 This table presents the schematic of the search dataset after preprocessing with a page-specific approach. Note that 0 represents False, and 1 represents True. If the value of Central is 0 for a record, it means that the segment was peripheral.

Media Name	Segment	Central	Relevant	Time
Apple	E	1	1	41
Apple	E	1	1	135
AVG	H	0	1	2
AVG	A	0	0	32
BBC	C	0	0	12
BBC	D	1	1	1253

Table 3.5 This table shows the schematic of the search dataset after preprocessing and encoded with a one-hot approach.

BBC	AVG	Apple	A	C	D	E	H	Central	Relevant	Time
0	0	1	0	0	0	1	0	1	1	41
0	0	1	0	0	0	1	0	1	1	135
0	1	0	0	0	0	0	1	0	1	2
0	1	0	1	0	0	0	0	0	0	32
1	0	0	0	1	0	0	0	0	0	12
1	0	0	0	0	1	0	0	1	1	1253

### 3.4.2 Feature Selection

We need to perform statistical comparisons to identify the features that help increase the accuracy of the classifiers. To this end, we need to recognize which variables are significantly different between groups.

We should first clarify that we perform the statistical analysis only for when we use 3\*3 grid segmentation. The result of statistical analysis for the page-specific segmentation is already reported in [45]. We used T-test<sup>7</sup> and Mann-Whitney U-test (U-test)<sup>8</sup> to explore how significant the difference is for a particular feature. The Mann-Whitney U test is often used when the data does not follow a specific distribution. If two groups (ASD and TD) are significantly behaving differently regarding the number of times visiting a segment or the amount of time they spent on each segment, these two tests would statically prove this fact.

### 3.4.3 Model Development

The programming language used to implement the solution is *Python Programming Language*. The libraries used are *Numpy* for working with arrays and perform-

---

<sup>7</sup> An independent-group t-test can be carried out for a comparison of means between two independent groups, with a paired t-test for paired data. As the t-test is a parametric test, samples should meet certain preconditions, such as normality, equal variances, and independence [94]

<sup>8</sup> This test can be used to investigate whether two independent samples were selected from populations having the same distribution [95]

ing matrix-related operations [96], *Pandas* to read and manipulate the eye-tracking data [97], *Seaborn* for data visualization [98] and *Scikit-Learn* to employ various machine learning models.

First, we perform the classification without any feature selection. Then, based on the results we obtained in statistical comparison, we remove those features that do not seem significantly different from one another, and we keep the rest. After identifying the most contributing features, we perform the classification again. The following is the list of the classifiers we used with a brief explanation about how they function:

- Logistic Regression (LR) is used to model the probability that a class or an event occurs as an estimation [99]. LR can be used to measure the relationship between categorical dependent parameters [100].
- K Nearest Neighbour (KNN) is first introduced by [101] and is an instance-based learning method that considers the K closest instances to a new data record to classify it [102].
- Support Vector Machine (SVM) is first introduced by [66] and uses a non-probabilistic binary linear classification to discriminate to predict the label of the class for a new data sample.
- Decision Tree (DT) is a tool for decision support purposes with a tree-based structure. This model can work based on information entropy gain or Gini index. Entropy in Information Theory is initially introduced by [103] and is widely used to compute the uncertainty inherent in the possible outcomes of a variable. Equation 31 is used to calculate entropy:

$$Entropy = - \sum_{i=1}^{i=n} P(x_i) \text{Log}(P(x_i)) \quad (31)$$

where  $P(x_i)$  is the probability that class  $x_i$  appears in a node and  $i$  is the number of the class.

Gini index is also used to compute inequality in the frequency of distributions [104]. Equation 32 is used to calculate the gini index:

$$Gini = 1 - \sum_{i=1}^{i=n} P^2(x_i) \quad (32)$$

where  $P(x_i)$  is the probability that class  $x_i$  appears in a node and  $i$  is the number of the class.

Entropy can be used to compute the information gain. The information gain (Kullback-Leibler divergence) is a statistical tool to measure the difference between two distribution demonstrated in probabilities [105]. Equation 33 is used to calculate the information gain:

$$IG(S, \gamma) = E(S) - E(S|\gamma) \quad (33)$$

where  $IG(S, \gamma)$  is the information gain,  $E(S)$  is the entropy of  $S$ , and  $E(S|\gamma)$  is the entropy of  $S$ , given  $\gamma$ .  $S$  represents a set of training samples and  $\gamma$  is an attribute.

A DT model works based on computing the information gain. In other words, a DT grows based on finding attributes with the highest information gain. The feature with maximum information gain becomes the parent node, and the rest of the dataset will be split based on that. This procedure continues until the nodes contain information about the classes of the dataset and cannot be split any further. According to [106], DT is used successfully in diverse areas, such as medical diagnosis.

- Random Forests (RF) uses the ensemble technique to overcome overfitting issues in a decision tree by growing multiple trees. Each tree performs a separated classification, and finally, the statistical model will be reported [107]. In general, ensemble methods are used to obtain better predictive performance for a machine learning model [108].
- Extra Trees (ET): is a relatively newer version of tree-based supervised learning methods. This model randomizes the cutting points as well as the attributes chosen to split the data [109]. In noisy datasets, this method provides all the attributes of a dataset to be chosen. This way, the tree grows uniformly.
- Adaptive Boosting (AD) which is also known as AdaBoost, is a machine learning meta-algorithm<sup>9</sup> which uses the boosting techniques to overcome overfit-

---

<sup>9</sup> A meta-algorithm is a higher-level procedure to provide an adequate solution for a sophisticated optimization problem [110].

ting [111]. In machine learning, boosting is an ensemble technique to combine several weak learners into a single strong learner [112]. The primary goal of using boosting techniques is to reduce the bias and the variance of a predictive model [113].

- Gradient Boosting (GB) is introduced by [114], and it uses the boosting technique to enhance the generalization of a set of weak learners.
- Gaussian Process (GP) is a collection of random variables with joint Gaussian distributions [115]. GP can be used as a predictive model for both regression and classification problems.
- Gaussian Naive Bayes (GNB) is a form of the Naive Bayes (NB) method. NB is a probabilistic classifier that uses Bayes' theorem [116]. GNB uses a regular (or Gaussian) kernel to compute the probability for each class. This model can be used When the dataset has continuous data, and the attributes follow a Gaussian distribution [117].
- Bernoulli Naive Bayes (BNB) is used for the classification of Bernoulli event models, and is usually used for text classification [118].
- Linear Discriminant Analysis (LDA) is an approach used for dimensionality reduction based on finding a linear combination of features. This method is developed based on Fisher's linear discriminant method [119, 120].
- Quadratic Discriminant Analysis (QDA) is similar to LDA. and considers that the variables of classes are typically distribute [121]. However, unlike LDA, QDA does not assume that the covariance of each class is identical.

#### **3.4.4 Model Evaluation**

The main issue is to distinguish ASD participants from TD users. [1] used logistic regression and leave five out cross-validation methods to train and evaluate their work, respectively. In this research, 13 different classifiers that were described earlier are trained and evaluated using the same leave five out cross-validation methodology.



For each dataset, we have two different tasks that we explained earlier. The training and evaluation procedure is same for both tasks. We have two classes of participants named ASD and TD. After data preparation and feature identification, we feed each classifier with eye-tracking data. To generate a random training set and a random testing set, for each classifier, we initiate a list containing a set of integer values from 1 to 15 (the number of remaining participants in each group is 15) for the first dataset. The program randomly chooses a number between 1 to 15, adds it to a new list that we call the training list, and then removes the chosen number from the main list. At each iteration, the integers in the training list are identical. We continue the previous step until we have 10 identical numbers on our training list. What remains in the main list is five identical integer values. We use the training list to choose 10 participants and train the classifiers, and we use the five remaining values in the main list to test the model. Note that we perform this for both ASD and TD groups. Consequently, we will have 20 participants for training each classifier and 10 participants for evaluation. After we train each classifier, we use equations 21, 26, 23 and 27 to compute and evaluate the performance of each classifier. We repeat this step for 100 times, and finally, report the statistical average for each parameter.

For the second dataset, we perform the same procedure, but the number of participants in the training and testing sets is different. For the second dataset, in which we have 19 participants in each category, we considered using the data from 14 participants to train the classifiers and used 10 participants similar to the first dataset to evaluate the model. For both datasets, each misclassification deducts 10% of the model accuracy. For instance, if one person in the testing set is misclassified, the accuracy of the model is reported to be 90%, if two participants are misclassified, the accuracy would be 80%, etc.

## CHAPTER 4

### RESULTS

In this chapter, the focus is on presenting the results obtained by applying the methodology explained in Chapter 3. This chapter is divided into three sections; Exploratory Analysis, Statistical Analysis and Classification Performance. As explained in the methodology section, 3\*3 grid segmentation is used, and prepared the data in a different way compared to [1], and therefore, we consider the segments as important features. Furthermore, unlike [1], this study does not interpret the time by computing how much each participant spent on each AOI. Therefore, the timestamp is used as a raw feature. In the exploratory analysis chapter, we illustrate data visualization to highlight the differences between people with the Autism Spectrum Disorder (ASD) as well as typically developing (TD) people. This study also uses some statistical methods to highlight the differences between ASD and TD. In this chapter, we used two datasets; the first dataset consists of a searching task and a freely browsing task, and the second dataset includes a synthesis task and a time-restricted browsing task, which is different from the browsing task in the first dataset. The first and the second dataset are used to train machine learning models which were explained in Chapter 3.

#### 4.1 Exploratory Analysis

This section uses data visualization to highlight the differences between people with the ASD and TD groups. We show two sets of figures for both the first and the second dataset. Figures 4.1, 4.2, 4.3 and 4.4 use bar charts to highlight the differences between ASD and TD people in terms of the total number of fixation, while, Figures 4.5, 4.6, 4.7 and 4.8 use box blot to show the range of total fixation duration per

each segment. Box plots can be useful to show the min, max and the statistical average of a parameter. As we explore in this chapter, fixation duration provides more statistical significance. Hence, box plots are used to provide more detailed visualization regarding the fixation duration. In addition to the value interval associated with a parameter, box plots show the average value.

Figure 4.1 illustrates the differences between the total number of fixations per each segment for the ASD and the same parameter TD groups. According to Figure 4.1, ASD group made more fixations in each segment. Both ASD and TD subjects made the maximum number of fixations on segment *E* and after that *B*. These two segments represent the middle of the web page. Both groups made fixations on segment *A* and *D* that represent the left side of the web page. In these four segments, ASD participants make more fixations compared to the TD participants. The ASD group made a low number of fixations in segments *F*, *G* and *H*, and TD participants had no fixations in these segments. Both group had no fixations in segments *C* and *I*. This is somewhat the same case for the browsing task of the first dataset. However, in the browsing task, the visualization in Figure 4.2 shows more fixations in segment *H*.

In the second dataset, Figure 4.3 and Figure 4.4 illustrate the same trends and patterns we encountered in the first dataset. For example, segments *A*, *B*, *D* and *E* attracted most of the participants' attention. However, there are small differences, such as visiting segment *F* by TD participants more and having approximately the same number of visits by two groups in segment *B* in the browsing task of the second dataset.

Figure 4.5 demonstrates the minimum and maximum amount of time each group spent on each segment exclusively. Looking at Figure 4.5 reveals more information that cannot be interpreted by looking at Figure 4.1. For instance, TD participants spent a small amount of time on segment *H*, representing the middle segment at the bottom of the web page. ASD participants spent significantly more time on segment *H*. ASD participants also spent the maximum amount of time looking at segment *G*, which cannot be interpreted from Figure 4.1. Similarly, ASD participants spent time on segments *I*, *C* and *F*, while ASD participants spent no time on them. In segments *A*, *B*, *D* and *E*, similar to all the other segments, ASD participants spent more time. The visualization for the fixation duration with regards to the browsing task in the

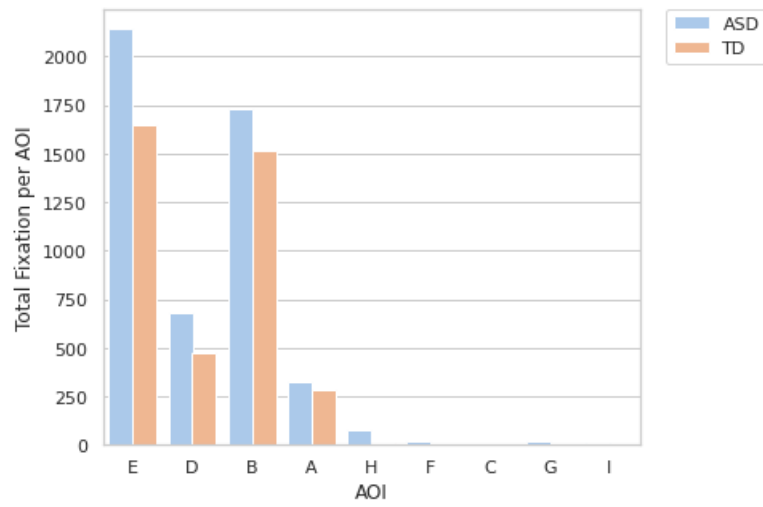


Figure 4.1 The total number of fixations on the segments of all the pages for the ASD and TD groups with regards to the searching task.

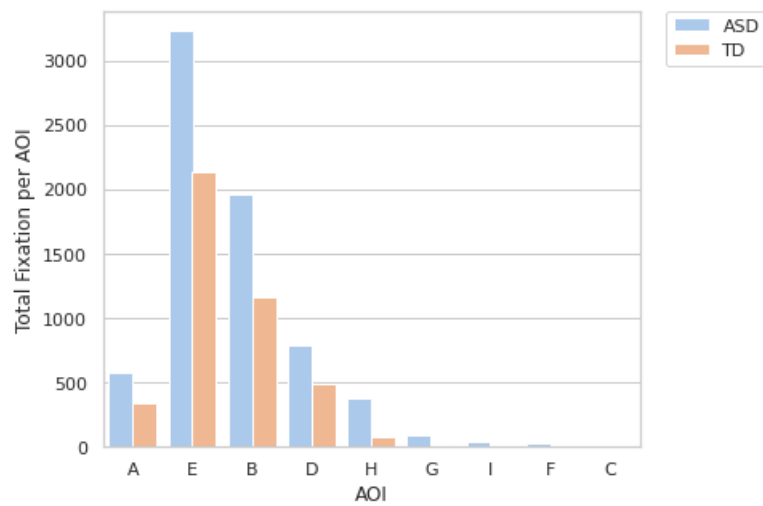


Figure 4.2 The total number of fixations on the segments of all the pages for the ASD and TD groups with regards to the browsing task in the first dataset.

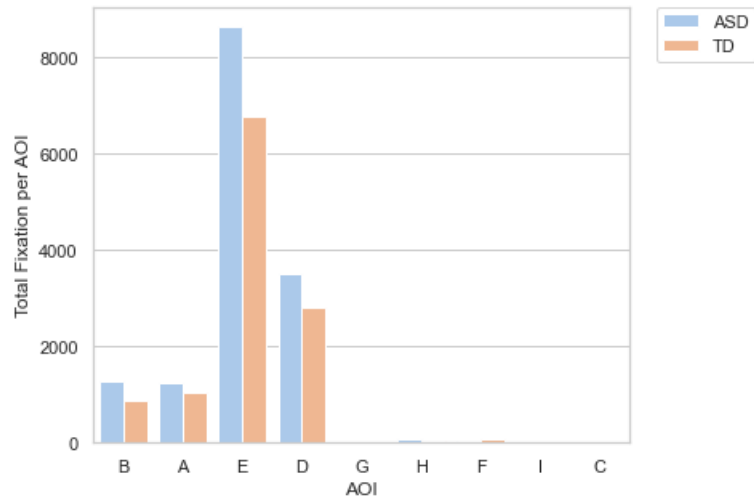


Figure 4.3 The total number of fixations on the segments of all the pages for the ASD and TD groups with regards to the synthesis task.

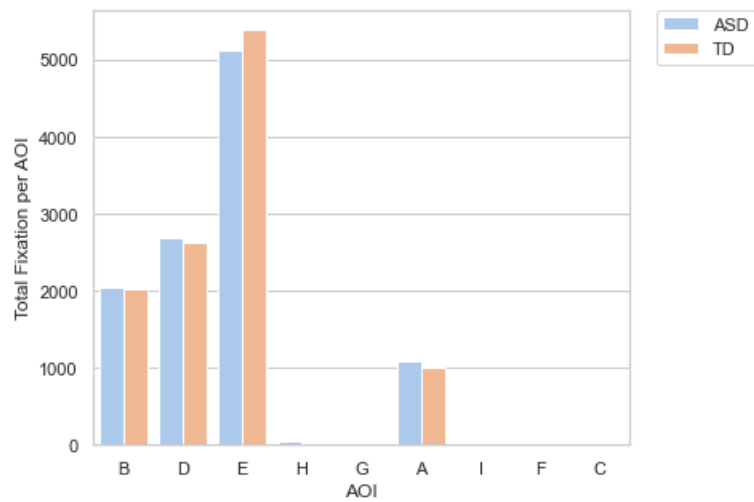


Figure 4.4 The total number of fixations on the segments of all the pages for the ASD and TD groups with regards to the browsing task in the second dataset.

first dataset shows almost the same trend in Figure 4.6. However, there are some differences, such as a larger time interval spend to observe each segment. Moreover, unlike the searching tasks, participants spent time on segment G as well. In the second dataset, as Figure 4.7 and Figure 4.8 show, the trends are the same about the majority of the time spent on segments. However, in the browsing task of the second dataset, according to Figure 4.8, in some of the segments such as B, TD participants spent more time compared to ASD participants.

The page-specific AOI segmentation [45] provides extensive analysis to highlight the differences between the ASD and the TD group for the first dataset. For the second dataset, [2] takes a comparative approach to highlight the differences between the ASD and the TD group for synthesis and a different browsing task. The findings of these two articles are discussed in Chapter 2.

## 4.2 Statistical Analysis

Using the T-test and Mann-Whitney U-test, Table 4.1 and Table 4.2 report the features with significant differences in the different tasks of both datasets with respect to the number of fixations and fixation duration, respectively. Comparing the first and the second dataset, the first dataset has more significant features, as the visualization also shows. Most of the significantly different features belong to the searching task, and participants showed more difference in terms of *fixation duration* than the *fixation count*. ASD participants seemed to perform the task slower, and therefore, spend more time on addressing the task compared to TD participants. This is the same case for the number of fixation duration, but the results show that difference for fixation count is not as significant as the difference in fixation duration.

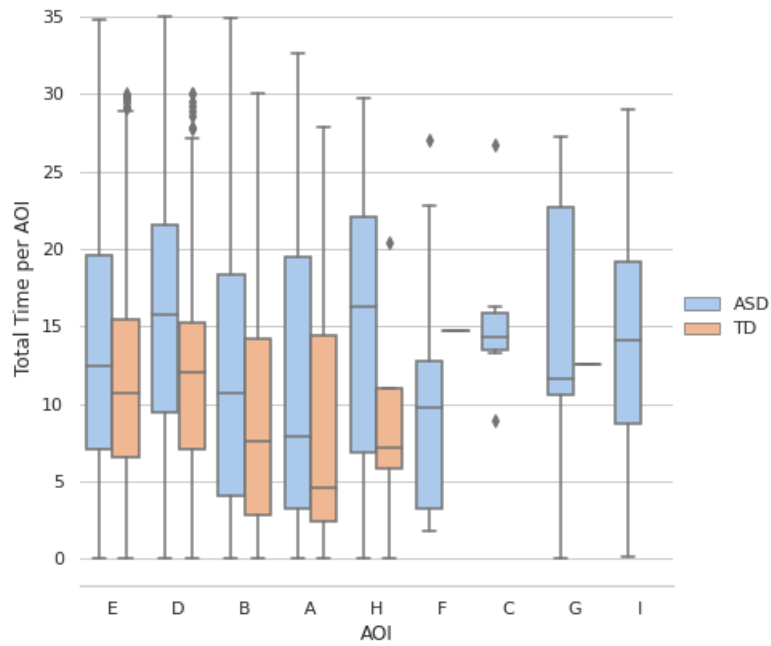


Figure 4.5 The minimum and maximum amount of time spent on the segments of all pages by ASD and TD participants in the searching task.

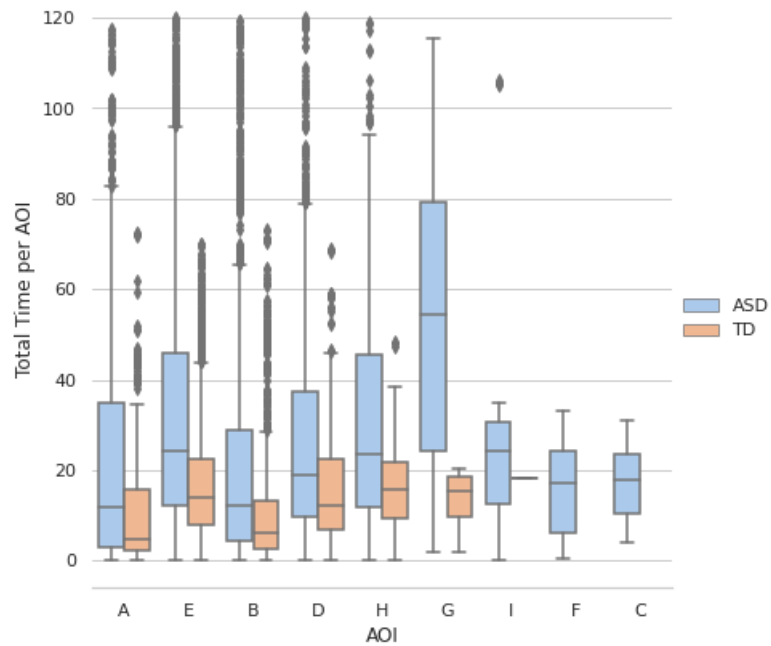


Figure 4.6 The minimum and maximum amount of time spent on the segments of all pages by ASD and TD participants in the browsing task of the first dataset.

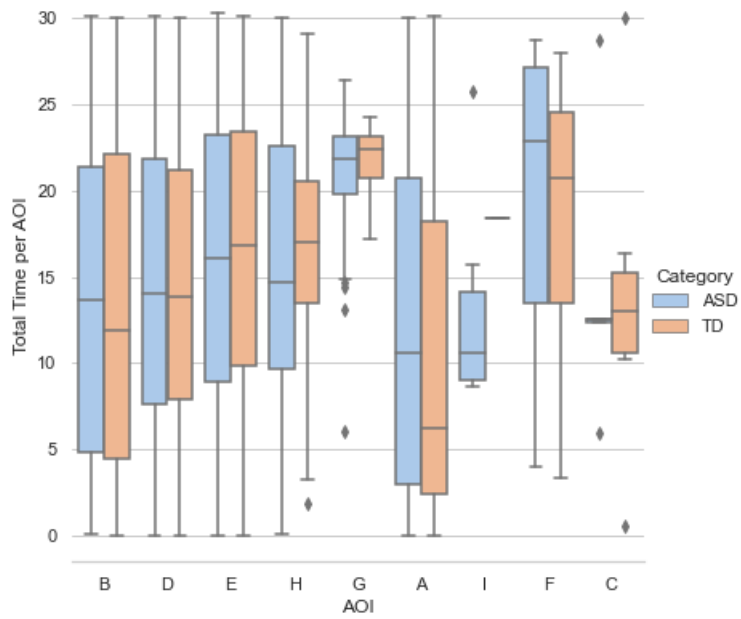


Figure 4.7 The minimum and maximum amount of time spent on the segments of all pages by ASD and TD participants in the synthesis task.

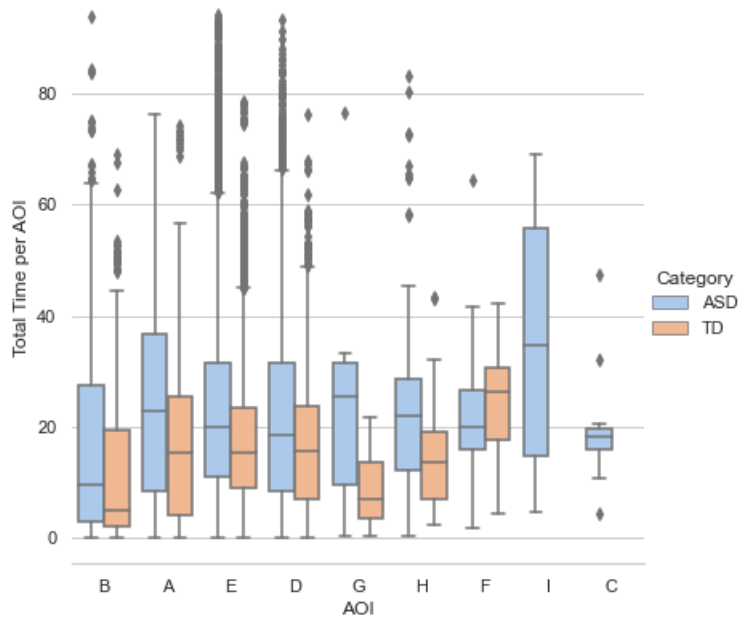


Figure 4.8 The minimum and maximum amount of time spent on the segments of all pages by ASD and TD participants in the browsing task of the second dataset.



Table 4.1 Significant features with respect to the total number of fixations for all web pages.

Media	Segment	t-val(p-val)	u-val(p-val)	ASD-avg(sd)	TD-avg(sd)
Searching Task					
Apple	B	2.16(0.04)	NA	16.86(7.83)	2.4(1.2)
Apple	E	2.29(0.02)	NA	25.6(6.83)	19.53(7.17)
Babylon	B	2.5(0.01)	NA	22.2(10.4)	13.53(7.72)
AVG	E	2.16(0.03)	NA	23.26(9.03)	16.86(6.36)
GoDaddy	E	160.5(0.04)	NA	12.53(7.36)	7.73(6.42)
BBC	B	NA	172.5(0.01)	9.33(3.92)	6.2(2.28)
Browsing Task - First Dataset					
Apple	B	NA	166(0.02)	16.7(14.5)	6.4(6.7)
AVG	A	NA	177.5(0.007)	16.7(14.5)	6.4(6.7)
GoDaddy	B	NA	171(0.01)	25(17.9)	13.4(7.9)
Yahoo	A	NA	162(0.03)	10.4(12.37)	3.8(3.1)
Synthesis Task					
Adobe	D	104(0.02)	NA	17(8.8)	11.6(5.4)
Adobe	H	NA	264.5(0.01)	14.3(7.5)	8.4(4.7)
Outlook	E	NA	278(0.004)	26.7(14.1)	16.1(3.8)
Outlook	F	NA	265.1(0.01)	16.36(7.4)	11(4.3)
Outlook	G	NA	258(0.02)	15.2(8.2)	9.6(3.8)
Whatsapp	H	NA	255.5(0.02)	32.5(16.2)	22.5(12.1)
Browsing Task - Second Dataset					
Whatsapp	C	NA	2.1(0.03)	2.1(2.4)	4(2.7)

Table 4.2 Significant features with respect to the total fixation duration for all web pages.

Media	Segment	t-val(p-val)	u-val(p-val)	ASD-avg(sd)	TD-avg(sd)
Searching Task					
Apple	B	2.29(0.008)	NA	248.24(149.71)	122.53(46.7)
Apple	D	NA	181(0.004)	154.8(135)	46.22(31.2)
Apple	E	2.76(0.01)	NA	303.6(155.77)	178.03(67.67)
Babylon	B	NA	179(0.006)	274.4(207.8)	85.8(65.8)
Babylon	D	2.14(0.04)	NA	99.6(75.19)	53.85(27.09)
Babylon	D	NA	164(0.03)	404(183)	267(105)
AVG	E	2.47(0.01)	NA	368(205)	16.86(6.36)
GoDaddy	E	NA	122.9(0.02)	60.2(82.1)	7.73(6.42)
BBC	B	NA	200(0.0003)	74.3(63)	15.6(9)
Browsing Task - First Dataset					
Apple	B	NA	172(0.01)	278(358)	55(90.7)
Apple	E	NA	163(0.03)	776(1468)	118(135)
AVG	A	NA	173(0.01)	113(272)	11(13)
AVG	B	NA	161(0.04)	373(492)	123(119)
GoDaddy	B	NA	174(0.01)	626(1258)	114(107)
Synthesis Task					
Adobe	D	254(0.03)	NA	445.5(315.8)	235.7(225.2)
Adobe	E	NA	261(0.01)	331.9(191)	186.3(156.8)
Adobe	H	NA	265(0.01)	303.5(265)	132.7(103.7)
Outlook	E	NA	290(0.001)	575.3(665.2)	193.9(94.7)
Outlook	F	NA	191(0.001)	445.6(472.3)	164.2(106.7)
Outlook	G	NA	277(0.005)	395.6(404.8)	159.7(93.8)
Outlook	H	NA	277(0.005)	627.3(365.6)	357.6(160.6)
Outlook	I	NA	279(0.004)	285.3(188.2)	163(89.4)
Whatsapp	H	NA	274(0.006)	1025.8(469)	631.5(442.4)
Youtube	H	NA	270(0.009)	295.2(255.2)	118.8(77.6)
Browsing Task - Second Dataset					
Whatsapp	C	NA	103(0.02)	34.1(61.9)	77.2(81.9)

### 4.3 Classification

We performed various experiments that we explain in this section, and the results for each experiment are represented in a separate table in the Appendix. Table 4.3 highlights the classifiers with the best accuracy for each experiment. Note that the media name is a feature in all experiments. Therefore, to keep the table more readable, we do not mention media name as a feature in Table 4.3.

Table 4.3 List of classifiers with the best accuracy in each experiment. The abbreviations used include XY: gaze point location, C: CNT is an attribute of the eye tracking dataset. Whenever a packet is successfully transferred (including a fixation record), CNT increases. Since it has 99% correlation with TIME, we considered CNT as the timestamp in our early experiments. S: page-specific AOI segments. T: actual timestamp, a feature of the raw eye tracking dataset which represents the timestamp. 22G: 2\*2 Grid segmentation. 33G: 3\*3 Grid segmentation. The best column also highlightsts the classifier with the highest performance accuracy in each experiment.

Experiment	Task	Features	Best	Accuracy	Precision	Recall	F1	TP	TN	FP	FN
1	Browse	XY	GNB	53.5	31.66	8	12.44	0.4	4.95	0.05	4.6
2	Browse	XYC	AB	61.3	59.11	87.99	69.15	4.4	1.72	3.28	0.6
3	Search	XY	GNB	53	56.86	56.4	62.93	3.04	2.16	2.83	1.96
4	Search	XYC	ET	76.3	81.99	69.6	73.66	3.48	4.15	0.85	1.52
5	Browse	S	QDA	55.2	6.6	13	8.8	0.65	4.37	0.63	4.35
6	Browse	SC	LR	62	60.8	89.39	70.15	4.47	1.73	3.27	0.53
7	Search	S	BNB	50.7	50.31	91.19	63.92	4.56	0.51	4.49	0.44
8	Search	SC	ET	80	84.45	77.79	79.15	3.89	4.11	0.89	1.11
9	Browse	ST	LR	63.7	61	94.4	72.66	4.72	1.65	3.35	0.24
10	Search	ST	ET	84	92.9	75.79	81.91	3.79	4.61	0.39	1.21
11	Search	ST	MLP	65.8	66.74	79	68.21	3.95	2.63	2.37	1.05
12	Browse	ST	MLP	60.1	58.41	83.2	66.5	4.16	1.85	3.15	0.84
14	Browse	ST	BNB	62	59.13	83.4	67.59	4.17	2.03	2.97	0.83
15	Search	ST	ET	77.5	77.11	83.59	78.89	4.18	3.57	1.43	0.82
16	Browse	22GT	LR	61.49	57.92	98	72.2	4.9	1.25	3.75	0.1
17	Search	22GT	DT	89.69	87.58	94.8	90.49	4.74	4.23	0.77	0.26
<b>18</b>	<b>Search</b>	<b>33GT</b>	<b>DT</b>	<b>90.5</b>	<b>91.47</b>	<b>91.39</b>	<b>90.7</b>	<b>4.57</b>	<b>4.48</b>	<b>0.52</b>	<b>0.43</b>
<b>26</b>	<b>Search</b>	<b>33GT</b>	<b>DT</b>	<b>91.6</b>	<b>91.8</b>	<b>92.9</b>	<b>91.8</b>	<b>4.65</b>	<b>4.51</b>	<b>0.49</b>	<b>0.35</b>

Firstly, we trained 13 classifiers explained in Chapter 3 using the browsing task of

the first dataset, and according to the obtained results, the classifiers did not learn anything from these features. We also included CNT to observe the effects of having a timestamp in the dataset. As earlier mentioned, CNT increases as the user starts to work on the task and generates gaze records. Compared to the first experiment, the model's accuracy has increased relatively for almost all the classifiers. We can conclude that CNT is potentially a useful feature, and therefore, we keep using it throughout experiments. Full classification results for these two experiments are demonstrated in Table A.1 and Table A.2. We repeated the same experiments for the searching task, so the features that we used in this experiment are the same. Similar to the first experiment, the lack of CNT leads to lower accuracy, and when CNT was included in the dataset the accuracy increased. The results show that the search dataset provides potentially more useful patterns to the classifiers. Detailed results can be seen in Table A.3 and Table A.4.

Previously we used gaze coordinates, and in the next experiments, given gaze coordinates, we prepared the data using page-specific AOIs. Additionally, we could determine whether a specific segment is central or peripheral. Note that the relevant feature is not suitable for using the browsing task dataset to train the classifiers, because a relevant attribute indicates if a particular segment is related to a searching task. We used the browsing task dataset and we did not include CNT. Hence, we trained the classifiers with media names, segments and central features, and the results show that if we exclude CNT, even page-specific AOI segments will not help the classifiers learn. Next, we aimed to observe the differences in improved accuracy after including CNT as a feature. As expected, CNT helps classifiers improve their performance. The highest accuracy of 61.4% belongs to LR. Table A.5 and Table A.6 represent the detailed classification results for this experiment.

We repeated the previous experiment using the searching task dataset. Similarly, when we exclude CNT, the model cannot learn from the rest of the features. However, using the searching task dataset, the accuracy is higher. In the next step we use the searching task dataset with CNT included. Hence, the attributes used in this experiment are Media Name, Segments, CNT, Central and Relevant. The results of this experiment demonstrate that DT, RF, and ET are performing significantly better. Considering TP, TN, FP and FN, these models classify over 8 participants in the test

set correctly and miss less than 2 participants in the overall 100 folds. More detailed results of these two experiments are reported in Table A.7 and Table A.8.

We are aware that a lack of timestamp (CNT) leads to a lower model accuracy, and CNT is positively correlated with TIME. Therefore, in the next experiments, we substituted CNT with TIME. We first used the browsing task data with the mentioned features excluding Relevant. LR outperformed all the other classifiers when we use the browsing task dataset by 63.7% accuracy. Then, we use the searching task dataset with the same features plus the Relevant attribute. ET outperforms all the other classifiers in all the experiments up to this point. Table A.9 and Table A.10 consist of more details regarding these two experiments. Although TIME and CNT correlate 99%, according to the Pearson Correlation Coefficient (PCC), TIME provides more useful information for the learning models. It is worth noting that previously we observed that the TN rate is so low, which means the classifiers were predicting most of the data records in the test set as ASD. However, in this experiment, the models show a better performance in classifying TD participants.

In the next step, we tried to develop a suitable deep artificial neural network architecture to train it using the dataset we prepared. We tried to explore how many hidden layers our model needs and how many neurons we should use in each layer to enhance the performance. The variations we used are listed in Table A.11. We use a fully connected multi-layer perceptron (MLP) model and train it with the features and data to achieve the best accuracy in the previous experiments. Therefore, we used the searching task dataset with features including Segment, Relevant, Central, TIME and media name. According to Table A.12, the accuracy does not change much. However, according to the 10th trial, we get 65.8% accuracy, which outperforms many of the statistical classifiers in the experiments we conducted so far. This accuracy is achieved by using the searching task dataset. In the next experiment, we use the browsing task dataset with the same architecture as used in the 10th trial. Results are reported in Table A.13. As we expected, even with the architecture with optimized parameters, it led to a lower accuracy than when we use the searching task data.

We calculate the difference value between two consecutive rows for the TIME variable as a feature engineering approach. We assumed that there might be differences

between the transitions of ASD and TD participants. We call this new feature “Difference”. We generated this feature, added it to the dataset and trained all 13 classifiers. As a result, all the classifiers’ accuracy decreased, and the highest precision using the searching task data belongs to ET with approximately 75% accuracy. However, without this new extracted feature, we achieved 84% accuracy. Therefore, we conclude that this feature misleads the classifiers.

In the next experiment, similar to [1], we try to use specific pages for classification. We only use the data associated with AVG and Apple web pages in the browsing task dataset. According to Table A.14, excluding specific pages from the dataset cannot help classifiers increase their prediction accuracy with regards to our data preparation approach. On the contrary, this approach helped [1] to improve its model accuracy. Similar to [1], we also use the searching task dataset for Apple and Babylone web pages only, and the results are reported in Table A.15. The accuracy is less than what we achieve by including all the web pages.

Up to this point, we tried to use page-specific AOI segments, including the headers, sidebars, icons at the top, etc. However, in the next step, we tried to specify generic AOIs with a grid segmentation approach. We used 2\*2 segmentation, and therefore, we have four segments with this approach. In the previous experiments, we had 25 segments categorized from A to Y. Nevertheless, with this perspective, we had only four segments. We have two purposes for conducting this experiment. The first one was to reduce the dimensionality of the data, and the second one was to increase the model’s generalization. In other words, instead of precisely looking into the number of fixations in each major web element (headers, sidebars, etc.), we divided it into four major segments to see if the classifiers can gain a more efficient performance. The idea was to prevent reducing the number of data records in the dataset. For this experiment, we used the browsing task dataset, and the results are reported in Table A.16.

All the classifiers performed poorly. GP could not converge, and therefore, did not provide any results. We repeated the previous experiment by using the searching task dataset. Results are represented in Table A.17. In this experiment, DT and ET performed significantly better. However, for DT, the rate of false classification

is not balanced. For DT, the rate of FP and FN is 0.77 and 0.26, respectively, and it shows that more data records were misclassified as ASD. Hence, with 2\*2 grid segmentation, DT outperforms other classifiers and is more successful in identifying people without autism. This is in contrast with [1], in which the authors report a higher rate for FN.

In the next step, we try 3\*3 page segmentation with the searching task dataset. Therefore, the total number of features will change from 11 (2\*2 grid segmentation) to 16. DT outperforms all the other classifiers, so far, this experiment led to the best results. We used the browsing task dataset to repeat the experiment. According to the results we obtained up to this point, DT, ET and RF demonstrate the best performances in descending order. Table A.18 and Table A.19 represent the detailed information for these experiments.

We performed a sanity check to verify whether the results we obtained in the previous experiments are due to an overfitted model or not:

1. We concatenate all the TD and ASD searching task dataset gathered from each participant exclusively and then removed the target column from the generated dataset.
2. We generate random target values and concatenated this column to the generated dataset.
3. We feed DT, ET, and RF classifiers with the new data.

80% of the dataset was used to train the dataset, and the rest was used for the evaluation. The classifications results are reported in Table A.20, and they show that the models are just predicting random classes. In binary classification, this is normal to have an accuracy of around 50% in this case since there is a probability of 0.5 for each category to show up as the predicted value. We also performed the sanity check for the browsing task dataset A.21, the obtained results are similar to when we used the searching task dataset.

After running T-test and Mann-Whitney U-test, we realized that page Yahoo and segments A, C, F, G, H, and I are not significantly different to the classification's learning

model with regards to the searching task. Therefore, we decided to remove them for the sake of dimensionality reduction. The results reported in Table 4.4 show that the decision tree classifier performs even better and obtains 91.6% model accuracy. Table 4.4 also shows the result of classifications with all segments included. As the results show, in all cases except ET in the searching task, all the results improved after feature selection. We also use another dataset which is first used in [2]. The obtained results are lower than when we use the 3\*3 segmentation approach.

Furthermore, the low results were not unexpected since our statistical analysis demonstrates fewer significantly different 3\*3 segments. However, the results show that feature selection improves almost all the dataset results and all four tasks. Some of these improvements, such as the improvement in the second dataset's browsing task, is considerable, even though the overall accuracy is not high. The best accuracy obtained is 91.6% using a decision tree and leave one out cross-validation. The best result was obtained when we used the searching task, prepared by a 3\*3 grid segmentation approach and removed Yahoo, A, C, F, G, H and I from the dataset. We also achieved 76.3% with the browsing task of the first dataset after feature selection.

Table 4.4 The classification results using the searching task and 3\*3 grid segmentation before and after feature selection. The numerical value out of the parentheses shows the result after feature selection, and the value reported within parentheses shows the result before feature selection.



Model	Accuracy	Precision	Recall	F1	TP	TN	FP	FN
Searching Task								
DT	91.6(90.5)	91.8(91.4)	92.9(91.3)	91.8(90.7)	4.65(4.5)	4.5(4.4)	0.4(0.5)	0.3(0.4)
RF	83.4(81.1)	79.7(77.7)	94.8(93)	85.6(83.6)	4.7(4.6)	3.6(3.4)	1.4(1.5)	0.2(0.3)
ET	86.9(88.1)	85.4(88)	91.1(92.7)	88(88.7)	4.64(4.5)	4(4.2)	0.9(0.7)	0.3(0.4)
Browsing Task, First Dataset								
DT	76.3(68.9)	81.3(73.4)	73.1(68.2)	74.9(67.8)	3.66(3.4)	3.97(3.4)	1.03(1.5)	1.34(1.5)
RF	65.3(60)	65.1(57.7)	82.4(84.8)	70.5(67.6)	4.1(4.2)	2.4(1.7)	2.59(3.2)	0.88(0.7)
ET	73.8(65.1)	74.9(65.8)	78.3(76.5)	74.7(68.4)	3.92(3.8)	3.46(2.6)	1.54(2.3)	1.08(1.1)
Synthesis Task								
DT	56.7(54.8)	54.5(53.3)	88.2(93.6)	66.9(67.6)	4.41(4.68)	1.26(0.81)	3.74(4.19)	0.59(0.32)
RF	53.9(53.4)	52.2(52.2)	98.4(0.96)	68.1(67.5)	4.92(4.8)	0.47(0.55)	4.53(4.45)	0.08(0.2)
ET	57.8(54.8)	54.9(52.9)	95.2(95.8)	69.4(68)	4.76(4.79)	1.02(0.69)	3.98(4.3)	0.24(0.21)
Browsing Task - Second Dataset								
DT	52.3(45.8)	32(37.4)	7(27.2)	0.12(29.6)	2.52(1.36)	3.17(3.23)	1.83(1.77)	2.48(3.64)
RF	56.9(40.6)	58(27.7)	50.4(26.8)	49.8(26.3)	2.52(1.34)	3.17(2.73)	1.83(2.27)	2.48(3.66)
ET	50.4(44.5)	36.5(35.8)	12.2(27.1)	17.4(28.7)	0.61(1.36)	4.43(3.09)	0.57(1.91)	4.39(3.64)

As the last experiment we trained the best classifiers with the eye tracking data associated with each web page separately. The results are reported in Table A.22 and Table A.23 show the results for the searching task dataset and the browsing task dataset, respectively. The results suggest that the BBC and Apple web pages provide the highest accuracy for in the searching task and browsing task dataset, respectively.

By looking at Table 4.4 we can see that the second dataset did not help the classifiers learn to predict high functioning autism. Similar to the first dataset, the browsing task in the second dataset leads to lower accuracy. We also used the data gathered for each web page separately to train the classifiers. The results shown in Table A.24 demonstrate that Outlook web page for the synthesis task provides the classifiers with more useful patterns for classification. For the second dataset’s browsing task, Whatsapp and Youtube web pages are better compared to others. However, for both tasks, the results are lower than what we obtained using individual web pages in the first dataset. Since [19] conducted experiments using 2\*2 and 4\*4 grid segmentation, we decided to test it using the approach developed in this thesis. However, as expected, results do not demonstrate any improvement in the classification results. The results are reported in Table A.25.

In summary, we used data visualization and statistical analysis in Chapter 3 to highlight the differences between the ASD and the TD participants while interacting with

different web pages. After that we used various classifiers and observed their performance while we trained them with eye tracking data prepared using different approaches. One approach was preparing the data using page-specific AOI segments, and the other one was based on grid segmentation. Even though we obtained considerable results without feature selection, after running statistical analysis and identifying the most contributing features of the dataset, we could enhance the performance of DT, ET and RF. Although the classifiers did not demonstrate high accuracy using the second dataset, after feature selection, the results improved, as demonstrated in Chapter 4.

## CHAPTER 5

### DISCUSSION AND CONCLUSION

#### 5.1 Discussion

This study uses machine learning models to detect high functioning autism using eye-tracking data. The data was previously collected while participants perform specific tasks using webpages. We first reviewed the literature about autism and tried to understand what may cause this problem, the symptoms, and what tools are used to study Autism Spectrum Disorder (ASD) detection. In the past two decades, there have been studies that aim to investigate ASD using eye-tracking. Researchers primarily used videos or photos containing facial or naturalistic features such as outdoor objects to study how people with ASD observe such figures. Some of these studies use the gathered data to train machine learning algorithms for autism detection. However, before our study, only [1, 20, 19, 45] use eye-tracking data gathered during web interactions to detect autism. This can be considered as a significant gap in the field, given that web users with ASD may spend a considerable portion of their time interacting with the web. Therefore, eye-tracking data during web interaction can be handy to investigate ASD and help other researchers investigate the accessibility issues.

One point that can be discussed is the pros and cons of using eye-trackers instead of other equipment such as EEG or fMRI or even clinical techniques. Referring to [122], we reviewed different types of behavioral data used for ASD detection. For example, the Diagnostic Observation Schedule (ADOS) introduced in the year 2000 is a test to diagnose children with autism. In this test, the examiner and the person under assessment have social interactions, and the examiner observes the subjects' behavior and reactions and then assigns them to specific categories that can be considered ASD

symptoms. However [123] states that in their studies 66.5% of a group of 161 subjects received one or more diagnoses different from ASD by psychiatrists when they used standardized tools such as ADOS, ADR-I, etc. When we consider the sharp increase of ASD prevalence and the possibility of results going misdiagnosed, it might be necessary to investigate autism using other data collection methods, including fMRI, EEG or eye-tracking.

Eye-trackers may be relatively less expensive than fMRI or EEG equipment; the data they provide requires less storage than fMRI. Consequently, it takes less processing time and units to extract information from eye-tracking data. After preprocessing the data, the machine learning algorithms used for eye-tracking data are mostly statistical models. However, according to the literature, fMRI data provides a considerable volume of data, and it enables researchers to use deep learning networks for ASD detection (e.g., [75]). Various studies investigate the level of attention in people with ASD using an eye-tracking approach and report that there may be significantly different attention levels in people with ASD than the TD people.

Eye movements are also investigated to help study attention, as eye movement can show if an individual is paying attention to a subject or an event. However, fMRI tries to find the regions of activation (ROI) of the brain by measuring the level of blood flow in ROIs. Hence, fMRI data may not provide much information regarding the subjects' attention, especially if collected in a resting state (resting fMRI). The reason is that the subject in a resting state is not asked to perform any particular task. Therefore, the activation regions will not be analyzed concerning a given task requiring attention and cognitive effort. However, eye-tracking data is usually used by giving the subjects a task as an experiment to analyze their approach towards completing a task, which includes analyzing the level of attention.

ROIs are analyzed by processing voxel, which contains embedded time series, which provides researchers with numerical values that can be used to make comparisons and look for distinguishable patterns. In other words, fMRI data may require more computational operations to become ready for training machine learning models with them. Therefore, looking for differences in eye-tracking data may be easier to extract and use by statistical learning models.

Various tasks, data preparation, and features led to obtaining various model performances. In this thesis the highest accuracy we obtained was 91.6% using the search task with selected features. According to the data visualization and statistical analysis, there are significantly different patterns in the visited AOIs and their time spent. One explanation for obtaining this accuracy, which is considerably higher than what was obtained in [1] can be the data preparation technique with a different algorithm. In our method, except for the time timestamp, all other features are represented using one-hot encoding, which means they either hold the value 1 or 0. For this representation, we create new columns for each categorical data, which adds to the data dimensions. However, this representation technique allows learning algorithms such as decision tree to rely on computing entropy for each column. In [1], and [19], the raw data is processed by considering how many times each AOI is visited and how much time each participant spent on each AOI, making the features of the processed data continuous. Using a Decision tree algorithm with this data preparation technique would not be suitable, as the categories have unique continuous values, making it difficult for tree-based algorithms to learn from them. We could obtain relatively higher accuracy with the same approach for the browse task of the first dataset. With feature selection, training a decision tree algorithm using the first dataset's browse task, we obtained 76.3. The accuracy we obtained for the browsing task is higher than the best accuracy obtained in [1] for the search task, given the task, which is approximately 75%, given the fact in both our study and [1], the search was found more suitable to train learning algorithms.

In [1] the raw eye-tracking data is used to generate a dataset based on various page-specific areas of interest (AOI), and each gaze record is checked to find out whether it is positioned on an AOI or not. Moreover, the amount of time spent on each AOI is computed for each fixation record and each participant. Using a different approach, [43] generates the scan paths for each participant, including sequences, each of which is labeled with a page-specific AOI ID following the amount of time being spent on that region of the web page. There could be two major issues with these approaches. Firstly, these methods only consider the fixation records located within one of the page-specific AOIs and do not consider those fixations that were still on the screen but not inside the generated AOIs. This approach leads to dropping a con-

siderable number of gaze records, making the learning models biased. Even though the grid segmentation is also investigated in [1], the best model performance is obtained when the authors used a page-specific method for segmenting the web pages. Secondly, performing such operations requires computational processes, which can increase the solution's complexity. The current study used the raw dataset to train classifiers.

The web pages were initially selected and used by [20] to investigate if web users with autism experience barriers during web interactions. The web pages were among the most visited websites according to Alexa.com and varying in level of complexity based on ViCRAM [90]. When considering the effects of the web pages, our statistical analysis suggests that the web pages can affect machine learning algorithms' performance. This might be due to different levels of complexity or the variety of content on the webpages. For instance, the BBC web page has a high complexity level, including facial and naturalistic features on distinct web page regions. Therefore, the BBC webpage might have been more successful in capturing the different patterns in visually processing the web pages. When only trained with the data associated with one page, machine learning algorithms' performance shows that their performance while training with the BBC web page is relatively better. The statistical analysis also demonstrates that features related to the BBC web page are more significant than the other webpages, which is in line with the obtained classifier accuracy. However, we have not investigated this issue further since there is a limited number of web pages in this study, which can be considered a point that requires improvement in further investigations.

In the current study, both grid segmentation and page-specific segmentation are investigated. The accuracy results obtained when we prepared the data with page-specific segmentation are relatively lower than when we use grid segmentation. One possible explanation could be that we have to delete some of the data points because they are not fixated on any specific AOI, which reduces the amount of data used to train machine learning algorithms. Furthermore, by grid segmentation, we reduce the number of features in our dataset which helps the algorithms perform while encountering a lower level of complexity. When we consider the browse task and the search task, grid segmentation leads to better learning model performance. The browse task gen-

erally leads to less accuracy because participants freely looked at the web page in the browse task, which provided less distinguishable patterns for the learning models. For supporting this point of discussion we can refer to [45], which investigates finding empirical evidence regarding the distraction of adults with high functioning autism while interacting with the web. The results of this study suggest that people with ASD may not be able to focus on a particular task on the web, given that they get distracted by different web elements. The accuracy values varied for each web page. The amount of differences between the TD and the ASD groups is the reason for achieving different accuracy. For example, for the browse task, as reported in Chapter 4, there are fewer features that show significant differences according to our statistical analysis. Table 5.1 shows the difference between the volume of the data in different tasks in comparison to [1].

Table 5.1 The difference between the number of data points of the first dataset used in [1] compared to the current study.

	<b>Specific AOI</b>		<b>Generic AOI</b>	
	Browse	Search	Browse	Search
Yaneva. et.al	3360	3360	720	720
Current Study	6184	4534	8970	6468

On the other hand, there are relatively more features that our feature selection method considers as significantly different for the search task. For the second dataset, the number of significantly different features is more diminutive in both tasks than in the first dataset. One possible reason could be the effect of the task on eye movements. The tasks in the first dataset are different from the tasks in the second dataset. In the search task, which is in the first dataset, the participants were supposed to look for a piece of information in a particular region on each webpage, while in the synthesis task, which belongs to the second dataset, they were required to look for information in one part and then use that information to look for other information on other parts of the webpage. Referring to the figures provided in Chapter 4 (Figures 4.3, 4.7, 4.4 and 4.8), when we look at the data visualization of time spent on each generic AOI and the number of times each AOI is visited for the second dataset, in both tasks (synthesis and browsing) the differences are less significant, and our statistical analysis also

suggests the same (Table 4.1 and Table 4.2). However, as earlier mentioned, the tasks' effects can be investigated further by repeating the experiments and gathering a new dataset by new groups of ASD and TD subjects. It is essential to investigate the effects of different web tasks for ASD detection by conducting various experiments and collecting more gaze records from more web users using a variety of tasks.

We also used statistical tests to understand which features are providing learning models with potentially useful information. We then removed those features identified as *significantly indifferent* according to the tests. This helped reduce the dimensionality of the data, which increased the efficiency of the model and the classification performance at the same time. For instance, after feature selection, for the search task, the decision tree obtained a 1.1% improvement in the predictions, and the browse task of the first dataset obtained 6.4% improvement. We also observed improvements after feature selection for the second dataset (Table 4.4). In the mentioned studies that use the same dataset for autism detection, the prepared data's dimensionality is higher, but there are trials for selecting features limited to using particular webpages. However, in the current study, we investigate the possibility of selecting features across the webpages and the AOIs concerning the number of visits and the time spent systematically.

In [1] only the performance of a logistic regression classifier is reported, and it is only stated that this model provides the best results, and the authors did not share more details addressing why this model has the highest performance and what is the performance of other models. On the other hand, we investigated the performance of 13 classifiers given the prepared data, and among them, Decision Tree (DT), Random Forests (RF), and Various Trees (ET) showed the best performance. The reason possibly is the approach we used to prepare the data. As earlier mentioned, we keep the timestamp as a feature that gives a sequential attribute to the dataset, and the rest of the features are represented using the one-hot encoding method. This approach allows DT to track each feature's occurrence and use the Gini index to split efficiently. Gini index is used to compute the probability of a feature being wrongly classified when it is randomly picked. According to the literature, tree-based algorithms are widely used for autism detection. However, only a limited number of studies used them alongside eye-tracking data, and to the best of our knowledge, the current study



is the only research which uses tree based algorithm and eye-tracking data gathered during web interactions to detect high functioning autism.

In this study we encountered several constrains. One of our limitations was having access to the gaze records of only 30 participants in total for the first dataset, which we used to develop our methodology. We were aware that deep learning models and even statistical learning models require a sufficient amount of data, and inadequate data leads to unreliable conclusions. As an issue that can be investigated further, researchers may consider gathering data from many participants, and scientists may consider developing other methodologies for gathering more data faster. For instance, more web pages could be used as a stimulus to gather gaze records.

Finally, people with ASD may experience a different level of severity, which may have various symptoms. In the current study, we only investigate the gaze record gathered from adults with high functioning autism. As future work, this can be extended to investigate the gaze records of subjects with different levels of autism severity during web interactions. Furthermore, the targeted age range can change from adults to children and adolescents too.

## **5.2 Conclusion**

This study first introduces the problem, which is detecting high functioning autism using eye-tracking data analysis and machine learning models. The next step presents detailed literature review and shows that there is a limited number of studies on autism detection using eye trackers, and the number of studies that consider using webpages for this purpose is even more limited compared to other approaches. Therefore, we used webpages to address some of these gaps and could successfully improve the results.

This improvement consists of both computation efficiency and the performance of the learning model. We increased our model's efficiency in computation by performing dimensionality reduction and performing fewer operations for the data preparation. Moreover, the best model accuracy obtained in this study was 91.6%, and it outperforms [1] by over 15% improvement.

This study can be used to further investigate ASD detection using web pages and eye-tracking data by collecting data from more participants, investigating the effect of tasks on model accuracy, and trying to duplicate the results using the same task. The outcome of such investigations can help clinical psychologists speed up the process of diagnosis with higher precision using state-of-the-art technologies.

We also experienced limitations in this study. One restriction was the number of participants. Moreover, we need another dataset with the same tasks to see if we can duplicate the results and conduct more research regarding the effects of tasks on visual processing strategies of ASD people and machine learning algorithms' performance. However, gathering such a dataset is time-consuming which cannot fit within a Master's studies time frame.

In summary, despite the study's limitations, we managed to improve the results, used the eye-tracking data more efficiently, and highlighted some of the gaps that can be addressed in future studies.

## APPENDIX A

### APPENDIX

#### A.1 Experiment Settings

The parameters used in each algorithm such as the Decision Tree algorithm are the ones used by default in the Scikit-Learn library.

Table A.1 First experiment: The classification results using the browsing task dataset with gaze coordinates and media name features.

Model	Accuracy	Precision	Recall	F1	TP	TN	FP	FN
LR	49.8	49.88	99.6	66.47	4.98	0	5	0.02
KNN	49.09	49.44	96.6	65.32	4.83	0.08	4.92	0.17
SVM	50	49.93	98.8	66.18	4.94	0.06	4.94	0.06
DT	49.5	49.76	95.8	65.38	4.79	0.16	4.84	0.21
RF	47.7	47.26	83	59.29	4.15	0.62	4.38	0.85
ET	46.9	46.82	84.8	59.9	4.24	0.45	4.55	0.76
AB	49.9	49.88	99.4	66.41	4.97	0.02	4.98	0.03
GB	49.5	49.33	0.98	65.59	4.9	0.05	4.95	0.1
GP	50.09	49.62	99	66.1	4.95	0.06	4.94	0.05
<b>GNB</b>	<b>53.5</b>	<b>31.66</b>	<b>8</b>	<b>12.44</b>	<b>0.4</b>	<b>4.95</b>	<b>0.05</b>	<b>4.6</b>
BNB	49.8	49.75	98.19	65.91	4.91	0.07	4.93	0.09
LDA	50	50	99.8	66.61	4.99	0.01	4.93	0.09
QDA	52.5	36.61	40	34.29	2	3.25	1.75	3

Table A.2 Second experiment: The classification results using the browsing task dataset with gaze coordinates, media name and CNT features.

Model	Accuracy	Precision	Recall	F1	TP	TN	FP	FN
LR	58.6	56.1	97.4	70.4	4.87	0.99	4.01	0.13
KNN	59.8	56.55	98.2	71.3	4.91	1.07	3.93	0.09
SVM	51.2	50.68	100	67.24	5	0.12	4.88	0
DT	57.9	56.17	83.59	66.17	4.18	1.61	3.39	0.82
RF	57.4	60.75	58.2	55.94	2.91	2.83	2.17	2.09
ET	55.2	58.26	53.6	52.88	2.68	2.84	2.16	2.32
<b>AB</b>	<b>61.2</b>	<b>59.11</b>	<b>87.99</b>	<b>69.15</b>	<b>4.4</b>	<b>1.72</b>	<b>3.28</b>	<b>0.6</b>
GB	59.1	57.99	92.2	69.14	4.61	1.3	3.7	0.39
GP	58.1	54.93	98.8	70.4	4.94	0.87	4.13	0.06
GNB	56.6	57	13.39	21.38	0.67	4.99	0.01	4.33
BNB	49.9	49.8	97.6	65.75	4.88	0.11	4.89	0.12
LDA	55.5	54.3	97.4	68.9	4.87	0.68	4.32	0.13
QDA	55.49	56.89	23.6	29.01	1.81	4.37	0.63	3.82

Table A.3 Third experiment: The classification results using the searching task dataset with gaze coordinates and media name.

Model	Accuracy	Precision	Recall	F1	TP	TN	FP	FN
LR	50	50	100	66.6	5	0	5	0
KNN	47	48.2	86.6	61.6	4.33	0.37	4.63	0.67
SVM	50	50	100	66.6	5	0	5	0
DT	48.3	48.7	86.2	61.98	4.31	0.52	4.48	0.69
RF	48.2	48.5	57.2	50.5	2.8	1.98	3.01	2.2
ET	49.2	50.8	60.8	52.39	2.86	1.94	3.07	2.14
AB	50.59	50.5	97.2	66.04	4.81	0.24	4.75	0.19
GB	49.69	49.78	97.8	65.89	4.89	0.08	4.92	0.11
GP	50	50	100	60.6	5	0	5	0
<b>GNB</b>	<b>53</b>	<b>56.86</b>	<b>56.4</b>	<b>62.93</b>	<b>3.04</b>	<b>2.16</b>	<b>2.83</b>	<b>1.96</b>
BNB	50	50	100	66.6	5	0	5	0
LDA	50	50	100	66.6	5	0	5	0
QDA	51.4	54.2	73.8	57.57	2.27	2.74	2.25	2.73

Table A.4 Fourth experiment: The classification results using the searching task dataset with gaze coordinates, media name and CNT features.

Model	Accuracy	Precision	Recall	F1	TP	TN	FP	FN
LR	50.7	50.38	100	67	5	0.07	4.93	0
KNN	48.7	56.1	96.19	70.32	4.81	1.06	3.94	0.19
SVM	50.9	50.62	99.6	67.04	4.98	0.11	4.89	0.02
DT	62.2	58.86	88.19	69.98	4.41	1.81	3.19	0.59
RF	74.69	78.93	68.99	72.23	3.45	4.02	0.98	1.55
<b>ET</b>	<b>76.3</b>	<b>81.99</b>	<b>69.6</b>	<b>73.66</b>	<b>3.48</b>	<b>4.15</b>	<b>0.85</b>	<b>1.52</b>
AB	65.4	66.15	78.59	68.74	3.93	2.61	2.39	1.07
GB	66.79	65.63	84.99	72.25	4.25	2.43	2.57	0.75
GP	57.7	55.73	97.2	70.15	4.86	0.91	4.09	0.14
GNB	60.79	61.49	53.8	49	2.69	3.39	1.61	2.31
BNB	50	50	100	66.6	5	0	5	0
LDA	50.7	50.4	99.8	66.95	4.99	0.08	4.92	0.01
QDA	55.89	47.52	43.8	38.86	2.19	3.4	1.6	2.81

Table A.5 Fifth experiment: The classification results using the browsing task dataset with page-specific AOI segments, central and media name.

Model	Accuracy	Precision	Recall	F1	TP	TN	FP	FN
LR	49.5	49.57	98.6	65.93	4.93	0.02	4.98	0.07
KNN	50	50	100	66.6	5	0	5	0
SVM	49.7	49.5	97.8	65.53	4.98	0.08	4.92	0.02
DT	50	50	100	66.6	5	0	5	0
RF	50	49.94	99.4	66.45	4.97	0.03	4.97	0.03
ET	49.59	49.75	0.99	66.2	4.95	0.01	4.99	0.05
AB	49.6	49.33	98	65.57	4.9	0.06	4.94	0.1
GB	49.5	49.33	0.98	65.59	4.9	0.05	4.95	0.1
GP	49.8	49.83	94.4	66.36	4.97	0.01	4.99	0.03
GNB	50.2	4.9	6.4	5.18	0.32	4.7	0.3	4.68
BNB	50	49.01	95.6	64.68	4.78	0.22	4.78	0.22
LDA	50.2	50.1	100	66.76	5	0.02	4.98	0
QDA	50.8	6.6	13	8.8	0.65	4.37	0.63	4.35

Table A.6 Sixth experiment: The classification results using the browsing task dataset with page-specific AOI segments, central, CNT and media name.

Model	Accuracy	Precision	Recall	F1	TP	TN	FP	FN
<b>LR</b>	<b>62</b>	<b>60.81</b>	<b>89.39</b>	<b>70.15</b>	<b>4.47</b>	<b>1.73</b>	<b>3.27</b>	<b>1.73</b>
KNN	57.7	56.18	92.19	68.78	4.61	1.16	3.84	1.16
SVM	54	52.44	99	68.39	4.95	0.45	4.55	0.45
DT	49.3	50.05	74.6	58.89	3.73	1.2	3.8	1.2
RF	50.19	51.26	68.39	56.74	3.42	1.6	3.4	1.6
ET	53.29	54.04	68.79	58.47	3.44	1.89	3.11	1.89
AB	59.3	57.75	88.6	67.89	4.43	1.5	3.5	1.5
GB	57	56.74	87.2	66.63	4.36	1.34	3.66	1.34
GP	57.8	56.43	88.59	67.65	4.43	1.35	3.65	1.35
GNB	53.69	35.25	11.59	16.51	0.66	4.75	0.25	4.37
BNB	48.6	47.86	90.79	62.11	4.54	0.32	4.68	0.32
LDA	61.5	60.11	92.8	70.74	4.64	1.51	3.49	1.51
QDA	50.59	7.5	4.1	4	0.27	4.85	0.15	4.85

Table A.7 Seventh experiment: The classification results using the searching task dataset with page-specific AOI segments, media name, central and relevant.

Model	Accuracy	Precision	Recall	F1	TP	TN	FP	FN
LR	50	49.98	97.4	65.4	4.87	0.13	4.87	0.13
KNN	50	50	100	66.6	5	0	5	0
SVM	50	50.04	99.4	66.5	4.97	0.03	4.97	0.03
DT	49	49.14	93.14	63.43	4.67	0.23	4.77	0.33
RF	49.2	48.63	95	64.19	4.75	0.17	4.83	0.25
ET	49.1	49.52	92.59	63.75	4.63	0.28	4.72	0.37
AB	50.3	50.19	99.8	66.76	4.99	0.04	4.96	0.01
GB	50.1	50.18	98.8	66.35	4.94	0.07	4.93	0.06
GP	49.6	49.67	96.6	65.44	4.83	0.13	4.87	0.17
GNB	50	43.52	86.2	57.76	4.31	0.69	4.31	0.69
BNB	50.7	50.31	91.19	63.92	4.56	0.51	4.49	0.44
LDA	49.7	49.83	99.2	66.32	4.96	0.01	4.99	0.04
QDA	49.8	48.44	95	64.6	4.75	0.23	4.77	0.25

Table A.8 Eighth experiment: The classification results using the searching task dataset with page-specific AOI segments, media name, central, relevant and CNT.

Model	Accuracy	Precision	Recall	F1	TP	TN	FP	FN
LR	54.7	53.82	94.4	67.7	4.72	0.75	4.25	0.28
KNN	57	55.19	91.99	68.32	4.6	1.1	3.9	0.4
SVM	54.69	54.5	96.8	68.62	4.84	0.63	4.37	0.16
DT	79.09	79.1	84.19	80	4.21	3.7	1.3	0.79
RF	77.4	79.68	79.99	77.68	4	3.74	1.26	1
<b>ET</b>	<b>80</b>	<b>84.45</b>	<b>77.79</b>	<b>79.15</b>	<b>3.89</b>	<b>4.11</b>	<b>0.89</b>	<b>1.11</b>
AB	62	62.86	81.39	68.52	4.07	2.13	2.87	0.93
GB	62.3	70.03	74	64.92	3.7	2.53	2.47	1.3
GP	59.8	58.1	90	69.49	4.5	1.48	3.52	0.5
GNB	56.4	52.16	63.59	51.51	3.18	2.46	2.54	1.82
BNB	50.5	49.83	91.2	63.3	4.56	0.46	4.51	0.44
LDA	52	51.55	94.19	66.14	4.71	0.49	4.51	0.29
QDA	50.1	44	86.2	57.66	4.31	0.7	4.3	0.69

Table A.9 Ninth experiment: The classification results using the browsing task dataset with page-specific AOI segments, media name, central and timestamp.

Model	Accuracy	Precision	Recall	F1	TP	TN	FP	FN
<b>LR</b>	<b>63.7</b>	<b>61</b>	<b>94.4</b>	<b>72.66</b>	<b>4.72</b>	<b>1.65</b>	<b>3.35</b>	<b>0.28</b>
KNN	51.6	51.82	85.2	63.53	4.26	0.9	4.1	0.74
SVM	57.2	56.21	90.99	67.25	4.55	1.17	3.83	0.45
DT	58.4	59.59	69.79	62.21	3.49	2.35	2.65	1.51
RF	53	53.34	63.59	56.66	3.18	2.12	2.88	1.82
ET	57	59.31	61.6	57.77	3.08	2.62	2.38	1.92
AB	61.2	61.21	85.8	68.34	4.29	1.83	3.17	0.71
GB	57.2	57.31	89	66.81	4.45	1.27	3.73	0.55
GP	53	52.36	89	64.74	4.45	0.85	4.15	0.55
GNB	51.8	19.88	9.3	10.45	0.47	4.71	0.29	4.53
BNB	50.2	49.83	91.2	63.3	4.56	0.46	4.51	0.44
LDA	61.4	59.3	93.2	70.87	4.66	1.48	3.52	0.34
QDA	51	12.6	81.99	57.66	7.87	0.41	4.69	0.31

Table A.10 Tenth experiment: The classification results using the searching task dataset with page-specific AOI segments, media name, central, relevant and TIME.

Model	Accuracy	Precision	Recall	F1	TP	TN	FP	FN
LR	52.9	51.97	96.2	67.1	4.81	0.48	4.52	0.48
KNN	72.99	71.15	85.6	76.23	4.28	3.02	1.98	0.82
SVM	59.7	63.29	64.99	57.08	3.25	2.72	2.28	1.75
DT	81.4	86.06	78.4	80.72	3.92	4.22	0.78	1.08
RF	78.8	82.86	75.79	77.75	3.79	4.09	0.91	4.09
<b>ET</b>	<b>84</b>	<b>92.9</b>	<b>75.79</b>	<b>81.91</b>	<b>3.79</b>	<b>4.61</b>	<b>0.39</b>	<b>1.21</b>
AB	62.7	64.73	80.39	68.45	4.02	2.25	2.75	2.75
GB	64.5	72.03	70.6	65.22	3.53	2.92	2.08	1.47
GP	67.39	67.87	81.8	71.55	4.09	2.65	2.35	0.91
GNB	50.2	45.75	88.2	59.73	4.41	0.61	4.39	0.59
BNB	50.6	50.28	94.6	65.12	4.73	0.63	4.37	0.27
LDA	63	53.69	94.8	67.39	4.74	0.63	4.37	0.26
QDA	50.1	44	86.2	57.66	4.31	0.7	4.3	0.69

Table A.11 The list of the parameters with different values used in different MLP trials.

TrialID	HiddenLayer#	Neuron#	Activation	Solver
1	10	35	relu	lbfgs
2	100	35	relu	lbfgs
3	10	50	relu	lbfgs
4	10	15	relu	lbfgs
5	10	35	relu	sgd
6	10	35	relu	adam
7	100	35	relu	adam
8	100	50	relu	adam
9	100	15	relu	adam
10	500	15	relu	adam
11	500	10	relu	adam



Table A.12 Eleventh experiment: using fully connected neural networks and the searching task dataset with page-specific AOI segment features.

Trial	Accuracy	Precision	Recall	F1	TP	TN	FP	FN
1	61.49	64.31	75.59	64.12	3.78	2.37	2.63	1.22
2	61.7	62.56	81.4	67.87	4.07	2.1	2.9	0.93
3	59.3	62.18	78.2	64.34	3.91	2.02	2.98	1.09
4	61.7	65.65	76	65.19	3.92	4.22	0.78	1.08
5	54.69	54.1	91.59	66.59	4.58	0.89	4.11	0.42
6	64.4	68.68	78.39	68.72	3.92	2.52	2.48	1.08
7	63.09	68	75.99	65.35	3.8	2.51	2.49	1.2
8	62.5	66.3	79	67.21	3.95	2.3	2.7	1.05
9	65.79	69.17	78.19	68.56	3.91	2.67	2.33	1.09
<b>10</b>	<b>65.8</b>	<b>66.74</b>	<b>79</b>	<b>68.21</b>	<b>3.95</b>	<b>2.63</b>	<b>2.37</b>	<b>1.05</b>
11	65.69	68.82	73.6	65.55	3.68	2.89	2.11	1.32

Table A.13 Twelfth experiment: using fully connected neural networks and the browsing task dataset with page-specific AOI segment features.

Trial	Accuracy	Precision	Recall	F1	TP	TN	FP	FN
1	60.1	58.41	83.2	66.5	4.16	1.85	3.15	0.84

Table A.14 Thirteenth experiment: classification results using the browsing task while only the gathered data on AVG and Apple web pages are used.

Model	Accuracy	Precision	Recall	F1	TP	TN	FP	FN
LR	59.39	56.94	90.59	69.1	4.53	1.41	3.59	0.47
KNN	61.3	58.38	84.8	68.02	4.24	1.89	3.11	0.76
SVM	54.7	53.25	97	68.33	4.85	0.62	4.38	0.15
DT	53.69	52.34	68.79	58.34	3.44	1.93	3.07	1.56
RF	51.29	51.23	71.99	58.97	3.6	1.53	3.47	1.4
ET	57.3	56.02	67.59	59.81	3.38	2.35	2.65	1.62
AB	59.3	57.2	91.19	68.77	4.56	1.37	3.63	0.44
GB	56.7	55.2	90.79	67.43	4.54	1.13	3.87	0.46
GP	54.89	53.04	95.2	67.89	4.76	0.73	4.27	0.24
GNB	52	0.19	0.04	6.57	0.2	5.0	0.0	4.8
<b>BNB</b>	<b>62</b>	<b>59.13</b>	<b>83.4</b>	<b>67.59</b>	<b>4.17</b>	<b>2.03</b>	<b>2.97</b>	<b>0.83</b>
LDA	59.8	56.73	97.4	71.15	4.87	1.11	3.89	0.13
QDA	50.2	0.02	0.4	0.66	0.02	5.0	0.0	4.98

Table A.15 Fourteen experiment: classification results using the searching task while only the gathered data on Apple and Babylone web pages are used.

Model	Accuracy	Precision	Recall	F1	TP	TN	FP	FN
LR	61.3	58.05	94.79	71.41	4.74	1.39	3.61	0.26
KNN	76.8	79.93	76.6	76.33	3.83	3.85	1.15	1.17
SVM	63.99	64.1	88	71.64	4.4	2.0	3.0	0.6
DT	77.19	74.66	87.59	79.72	4.38	3.34	1.66	0.62
RF	76.7	77.28	81.6	77.98	4.08	3.59	1.41	0.92
<b>ET</b>	<b>77.5</b>	<b>77.11</b>	<b>83.59</b>	<b>78.89</b>	<b>4.18</b>	<b>3.57</b>	<b>1.43</b>	<b>0.82</b>
AB	69.1	71.70	76.59	76.59	70.83	3.83	3.08	1.17
GB	73.9	72.05	84.79	77.91	4.24	3.34	1.66	0.76
GP	73.9	72.05	84.99	76.71	4.25	3.14	1.86	0.75
GNB	63	55.75	40.6	42.23	2.03	4.27	0.73	2.97
BNB	49.5	49.86	84	60.78	4.2	0.75	4.25	0.8
LDA	61.9	58.86	91.39	70.86	4.57	1.62	3.38	0.43
QDA	50	8	16	10.66	0.8	4.2	0.8	4.2

Table A.16 Fifteenth experiment: The classification results using the browsing task dataset with 2\*2 grid segmentation, media name and timestamp.

Model	Accuracy	Precision	Recall	F1	TP	TN	FP	FN
<b>LR</b>	<b>61.49</b>	<b>57.92</b>	<b>98</b>	<b>72.2</b>	<b>4.9</b>	<b>1.25</b>	<b>3.75</b>	<b>0.1</b>
KNN	55.89	53.8	94.8	68.27	4.74	0.85	4.15	0.26
SVM	51.4	50.87	98.6	67.02	4.93	0.21	4.79	0.07
DT	61	61.09	0.74	64.67	3.7	2.4	2.6	1.3
RF	54.89	53.92	82.19	64.19	4.11	1.38	3.62	0.89
ET	59.1	57.39	80.39	65.65	4.02	1.89	3.11	0.98
AB	57.4	54.49	98.8	70.08	4.94	0.8	4.2	0.06
GB	54.69	52.91	99.4	68.89	4.97	0.5	4.5	0.03
GP	-	-	-	-	-	-	-	-
GNB	57	63.88	34	40.01	1.7	4	1	3.3
BNB	50	50	1	66.66	5	0	5	0
LDA	55.8	53.94	99.8	69.72	4.99	0.59	4.41	0.01
QDA	55.5	55.22	48	45.08	2.4	3.15	1.85	2.6

Table A.17 Sixteenth experiment: The classification results using the searching task dataset with 2\*2 grid segmentation, media name and timestamp.

Model	Accuracy	Precision	Recall	F1	TP	TN	FP	FN
LR	49.5	49.72	98.8	66.14	4.94	0.01	4.99	0.06
KNN	66.09	61.99	92	73.35	4.6	2.01	2.99	0.4
SVM	63.7	59.97	96.6	73.25	4.83	1.54	3.64	0.17
<b>DT</b>	<b>89.69</b>	<b>87.58</b>	<b>94.8</b>	<b>90.49</b>	<b>4.74</b>	<b>4.23</b>	<b>0.77</b>	<b>0.26</b>
RF	80.5	77.64	91.6	82.93	4.58	3.47	1.53	0.42
ET	88.5	86.29	94.8	89.62	4.74	4.11	0.89	0.26
AB	63.2	61.59	92.59	72.51	4.63	1.69	3.31	0.37
GB	65.8	64.37	93.8	74.07	4.96	1.89	3.11	0.31
GP	70.9	66.49	95.4	77.24	4.77	2.32	2.68	0.23
GNB	77.4	85.6	72.6	74.03	3.63	4.11	0.89	1.37
BNB	50.5	50.68	99.2	66.67	4.96	0.09	4.91	0.04
LDA	49.7	49.83	98.6	66.19	4.93	0.04	4.96	0.07
QDA	60.4	58.14	71.8	59.22	3.59	2.45	2.55	1.41

Table A.18 Seventeenth experiment: The classification results using the searching task dataset with 3\*3 grid segmentation, media name and timestamp.

Model	Accuracy	Precision	Recall	F1	TP	TN	FP	FN
LR	49.7	49.83	98.6	66.18	4.93	0.04	4.96	0.07
KNN	63.5	60.84	90.39	71.65	4.52	1.83	3.17	0.48
SVM	58.5	59.37	93.4	69.81	4.67	1.18	3.82	0.33
<b>DT</b>	<b>90.5</b>	<b>91.47</b>	<b>91.39</b>	<b>90.7</b>	<b>4.57</b>	<b>4.48</b>	<b>0.52</b>	<b>0.43</b>
RF	81.1	77.75	93	83.67	4.65	3.46	1.54	0.35
ET	88.1	88.01	91.19	88.71	4.56	4.25	0.75	0.44
AB	53.1	52.14	96.6	67.43	4.83	0.48	4.52	0.17
GB	56.9	55.95	95.6	69.37	4.78	0.91	4.09	0.22
GP	70.9	66.49	95.4	77.24	4.77	2.32	2.68	0.23
GNB	50	0	0	0	0	5	0	5
BNB	50	50	100	66.66	5	0	5	0
LDA	49.69	49.84	98	66.04	4.9	0.07	4.93	1
QDA	51.2	32.37	51.6	36.57	2.58	2.54	2.46	1.42

Table A.19 Eighteenth experiment: The classification results using the browsing task dataset with 3\*3 grid segmentation, media name and timestamp.

Model	Accuracy	Precision	Recall	F1	TP	TN	FP	FN
LR	62.1	59.5	91.39	71	4.57	1.64	3.36	0.43
KNN	57	55.21	94.4	68.85	4.72	0.98	4.02	0.28
SVM	53	52.09	97	67.37	4.85	0.45	4.55	0.15
<b>DT</b>	<b>68.99</b>	<b>73.4</b>	<b>68.2</b>	<b>67.8</b>	<b>3.4</b>	<b>3.4</b>	<b>1.5</b>	<b>1.5</b>
RF	60	57.7	84.8	67.6	4.2	1.7	3.2	0.7
ET	65.1	65.8	76.5	68.4	3.8	2.6	2.3	1.1
AB	53.1	52.14	96.6	67.43	4.83	0.48	4.52	0.17
GB	56.9	55.95	95.6	69.37	4.78	0.91	4.09	0.22
GP	70.9	66.49	95.4	77.24	4.77	2.32	2.68	0.23
GNB	50	0	0	0	0	5	0	5
BNB	50	50	100	66.66	5	0	5	0
LDA	49.69	49.84	98	66.04	4.9	0.07	4.93	1
QDA	51.2	32.37	51.6	36.57	2.58	2.54	2.46	1.42

Table A.20 Nineteenth experiment: The sanity check for the obtained results using the searching task dataset.

Model	Accuracy	Precision	Recall	F1
DT	49.97	62.15	56.74	59.32
RF	52.26	64.06	58.53	61.17
ET	53.88	64.53	59.96	62.16

Table A.21 Twentieth experiment: The sanity check for the obtained results using the browsing task dataset.

Model	Accuracy	Precision	Recall	F1
DT	53.37	63.34	62.45	62.89
RF	54.48	64.32	64.54	64.43
ET	55.18	65.47	62.52	63.96

Table A.22 Twenty first experiment: The accuracy of the three best classifiers using the searching task data with respect to one page at a time. The values inside parentheses is the F1 score.

	Apple	Babylon	AVG	GoDaddy	Yahoo	BBC
<b>DT</b>	72.2(73.9)	74.8(75.2)	72.4(67.9)	72.3(72.5)	63.9(57.5)	<b>82.6* (80.6)</b>
<b>RF</b>	67(69.2)	69.4(73.7)	71(69.4)	67.8(69.1)	59.9(55.2)	79.2(78.5)
<b>ET</b>	71.4(72.7)	72.9(74.9)	74.7(70.3)	72.5(71.5)	66.7(60.6)	79.29(76.9)

Table A.23 Twenty second experiment: The accuracy of the three best classifiers using the browsing task data with respect to one page at a time. The values inside parentheses is the F1 score.

	Apple	AVG	Babylon	GoDaddy	Yahoo	BBC
DT	62(62.2)	45(42)	58.3(54.9)	66(67.6)	53.2(56.9)	50(52)
RF	59(64.6)	44.6(48.6)	58.1(57.9)	65.3(62.9)	53.6(58.1)	48(53.7)
ET	<b>65(67.3)*</b>	47(46.9)	62(60.3)	66.3(67.4)	54.2(54.8)	50.7(53.1)

Table A.24 Model accuracy percentage (%) of Decision Tree (DT), Random Forests (RF) and Extra Trees (ET) algorithm using individual web pages of the second dataset for both tasks. The names of web pages in the second study are Adobe (AD), Amazon (AZ), BBC (BC), Netflix (NF), Outlook (OL), Whatsapp (WA), Youtube (YT) and Wordpress (WP).

	Synthesis								Browse							
	AD	AZ	BC	NF	OL	WA	YT	WP	AD	AZ	BC	NF	OL	WA	YT	WP
<b>DT</b>	51	58	58.7	55.2	<b>67.9</b>	<b>61.5</b>	<b>58.4</b>	42.6	49.8	35.2	52.7	55.2	49.3	<b>57.5</b>	<b>58.7</b>	43.7
<b>RF</b>	52.1	56.1	56.4	56.3	<b>63.8</b>	<b>59.7</b>	<b>59.9</b>	42.2	49.9	34.2	51.6	54.9	53.2	<b>54.1</b>	<b>56.7</b>	39.9
<b>ET</b>	53.2	58.2	57.1	57.5	<b>68.8</b>	<b>61.4</b>	<b>61.5</b>	41.3	47.4	37	54.2	53.5	51.4	<b>57.7</b>	<b>57.5</b>	40.5

Table A.25 Model accuracy percentage (%) of Decision Tree (DT), Random Forests (RF), Extra Trees (ET) algorithm using 2\*2 and 4\*4 grid segmentation. The values inside the parenthesis show the F1 score. The results are reported for both the synthesis and the browsing task.

	Synthesis			Browse	
	4*4	2*2	4*4-OL,AZ	4*4	2*2
<b>DT</b>	58.1(64.6)	60.6(67.67)	59.6(68.5)	46.9(31.5)	46.3(29.68)
<b>RF</b>	56.8(66.5)	58.4(68.3)	56.5(66.1)	39.8(27.1)	40.7(28.7)
<b>ET</b>	58.6(67.6)	58.2(66.5)	56.6(65.8)	42.8(30.8)	43.6(28.2)

TEZ İZİN FORMU / THESIS PERMISSION FORM

PROGRAM / PROGRAM

Sürdürülebilir Çevre ve Enerji Sistemleri / Sustainable Environment and Energy Systems

Siyaset Bilimi ve Uluslararası İlişkiler / Political Science and International Relations

İngilizce Öğretmenliği / English Language Teaching

Elektrik Elektronik Mühendisliği / Electrical and Electronics Engineering

Bilgisayar Mühendisliği / Computer Engineering

Makina Mühendisliği / Mechanical Engineering

YAZARIN / AUTHOR

Soyadı / Surname : Khalaji

Adı / Name : Erfan

Programı / Program : Computer Engineering

TEZİN ADI / TITLE OF THE THESIS (İngilizce / English) : .....

A MACHINE LEARNING APPROACH FOR DETECTING  
HIGH-FUNCTIONING AUTISM USING WEB-BASED EYE-TRACKING DATA

TEZİN TÜRÜ / DEGREE: Yüksek Lisans / Master  Doktora / PhD

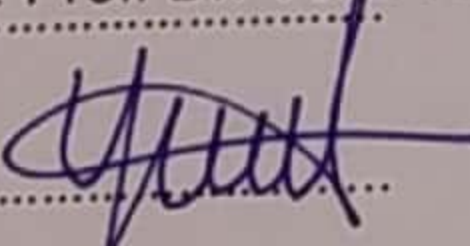
1. Tezin tamamı dünya çapında erişime açılacaktır. / Release the entire work immediately for access worldwide.

2. Tez iki yıl süreyle erişime kapalı olacaktır. / Secure the entire work for patent and/or proprietary purposes for a period of two years. \*

3. Tez altı ay süreyle erişime kapalı olacaktır. / Secure the entire work for period of six months. \*

Yazarın imzası / Author Signature:  Tarih / Date: 14.07.2021

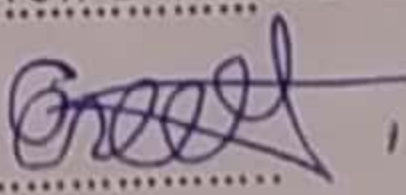
Tez Danışmanı / Thesis Advisor Full Name: Assoc. Prof. Dr. Yeliz Yesilada Yilmaz

Tez Danışmanı İmzası / Thesis Advisor Signature: 

Eş Danışmanı / Co-Advisor Full Name: Dr. Sukru Eraslan

Eş Danışmanı İmzası / Co-Advisor Signature: 

Program Koordinatörü / Program Coordinator Full Name: Assoc. Prof. Dr. Enver Ever

Program Koordinatörü İmzası / Program Coordinator Signature: 

## REFERENCES

- [1] V. Yaneva, L. A. Ha, S. Eraslan, Y. Yesilada, and R. Mitkov, “Detecting autism based on eye-tracking data from web searching tasks,” in *Proceedings of the Internet of Accessible Things*, pp. 1–10, 2018.
- [2] S. Eraslan, Y. Yesilada, V. Yaneva, *et al.*, ““keep it simple!”: an eye-tracking study for exploring complexity and distinguishability of web pages for people with autism,” *Universal Access in the Information Society*, pp. 1–16, 2020.
- [3] M. J. Maenner, K. A. Shaw, J. Baio, *et al.*, “Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, united states, 2016,” *MMWR Surveillance Summaries*, vol. 69, no. 4, p. 1, 2020.
- [4] J. S. Nicholas, J. M. Charles, L. A. Carpenter, L. B. King, W. Jenner, and E. G. Spratt, “Prevalence and characteristics of children with autism-spectrum disorders,” *Annals of epidemiology*, vol. 18, no. 2, pp. 130–136, 2008.
- [5] N. I. of Neurological Disorders and Strokes, “Autism spectrum disorder fact sheet.” [https://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/Fact-Sheets/Autism-Spectrum-Disorder-Fact-Sheet#3082\\_5](https://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/Fact-Sheets/Autism-Spectrum-Disorder-Fact-Sheet#3082_5), 2020.
- [6] A. Ronald and R. A. Hoekstra, “Autism spectrum disorders and autistic traits: a decade of new twin studies,” *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, vol. 156, no. 3, pp. 255–274, 2011.
- [7] C. for Disease Control and Prevention, “Screening and diagnosis of autism spectrum disorder for healthcare providers.” <https://www.cdc.gov/ncbddd/autism/hcp-screening.html>, 2020.
- [8] O. Tadevosyan-Leyfer, M. Dowd, R. Mankoski, B. Winklosky, S. Putnam, L. McGrath, H. Tager-Flusberg, and S. E. Folstein, “A principal components



analysis of the autism diagnostic interview-revised,” *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 42, no. 7, pp. 864–872, 2003.

- [9] C. Lord, S. Risi, L. Lambrecht, E. H. Cook, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter, “The autism diagnostic observation schedule—generic: A standard measure of social and communication deficits associated with the spectrum of autism,” *Journal of autism and developmental disorders*, vol. 30, no. 3, pp. 205–223, 2000.
- [10] M. E. Van Bourgondien, L. M. Marcus, and E. Schopler, “Comparison of dsm-iii-r and childhood autism rating scale diagnoses of autism,” *Journal of Autism and Developmental Disorders*, vol. 22, no. 4, pp. 493–506, 1992.
- [11] J. E. Gilliam, *Gilliam autism rating scale: Examiner’s manual*. Pro-ed, 1995.
- [12] A. P. Association *et al.*, *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.
- [13] C. Luciano, R. Keller, P. Politi, E. Aguglia, F. Magnano, L. Burti, F. Muraro, A. Aresi, S. Damiani, and D. Berardi, “Misdiagnosis of high function autism spectrum disorders in adults: an italian case series,” *Autism Open Access*, vol. 4, no. 131, p. 2, 2014.
- [14] E. Niedermeyer and F. L. da Silva, *Electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, 2005.
- [15] A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, *et al.*, “The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism,” *Molecular psychiatry*, vol. 19, no. 6, pp. 659–667, 2014.
- [16] S. Fletcher-Watson, S. R. Leekam, V. Benson, M. Frank, and J. Findlay, “Eye-movements reveal attention to social information in autism spectrum disorder,” *Neuropsychologia*, vol. 47, no. 1, pp. 248–257, 2009.
- [17] D. M. Riby and P. J. Hancock, “Do faces capture the attention of individuals with williams syndrome or autism? evidence from tracking eye movements,”

*Journal of autism and developmental disorders*, vol. 39, no. 3, pp. 421–431, 2009.

- [18] T. J. Heaton and M. Freeth, “Reduced visual exploration when viewing photographic scenes in individuals with autism spectrum disorder,” *Journal of abnormal psychology*, vol. 125, no. 3, p. 399, 2016.
- [19] V. Yaneva, S. Eraslan, Y. Yesilada, R. Mitkov, *et al.*, “Detecting high-functioning autism in adults using eye tracking and machine learning,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2020.
- [20] S. Eraslan, V. Yaneva, Y. Yesilada, and S. Harper, “Do web users with autism experience barriers when searching for information within web pages?,” in *Proceedings of the 14th Web for All Conference on The Future of Accessible Work*, pp. 1–4, 2017.
- [21] C. Lord, T. S. Brugha, T. Charman, J. Cusack, G. Dumas, T. Frazier, E. J. Jones, R. M. Jones, A. Pickles, M. W. State, *et al.*, “Autism spectrum disorder,” *Nature reviews Disease primers*, vol. 6, no. 1, pp. 1–23, 2020.
- [22] A. Mullin, A. Gokhale, A. Moreno-De-Luca, S. Sanyal, J. Waddington, and V. Faundez, “Neurodevelopmental disorders: mechanisms and boundary definitions from genomes, interactomes and proteomes,” *Translational psychiatry*, vol. 3, no. 12, pp. e329–e329, 2013.
- [23] R. J. Landa, “Diagnosis of autism spectrum disorders in the first 3 years of life,” *Nature Clinical Practice Neurology*, vol. 4, no. 3, pp. 138–147, 2008.
- [24] G. A. Stefanatos, “Regression in autistic spectrum disorders,” *Neuropsychology review*, vol. 18, no. 4, pp. 305–319, 2008.
- [25] T. W. H. Organization, *Autism spectrum disorders*, 2019 (accessed August 25, 2020).
- [26] I. Mitsugami, N. Ukita, and M. Kidode, “Robot navigation by eye pointing,” in *International Conference on Entertainment Computing*, pp. 256–267, Springer, 2005.

- [27] M. Wedel and R. Pieters, "A review of eye-tracking research in marketing," *Review of marketing research*, vol. 4, no. 2008, pp. 123–147, 2008.
- [28] D. E. Hannula, R. R. Althoff, D. E. Warren, L. Riggs, N. J. Cohen, and J. D. Ryan, "Worth a glance: using eye movements to investigate the cognitive neuroscience of memory," *Frontiers in human neuroscience*, vol. 4, p. 166, 2010.
- [29] A. T. Duchowski, "A breadth-first survey of eye-tracking applications," *Behavior Research Methods, Instruments, & Computers*, vol. 34, no. 4, pp. 455–470, 2002.
- [30] M. A. Just and P. A. Carpenter, "A theory of reading: From eye fixations to comprehension.," *Psychological review*, vol. 87, no. 4, p. 329, 1980.
- [31] Z. Boraston and S.-J. Blakemore, "The application of eye-tracking technology in the study of autism," *The Journal of physiology*, vol. 581, no. 3, pp. 893–898, 2007.
- [32] G. J. Walker-Smith, A. G. Gale, and J. M. Findlay, "Eye movement strategies involved in face perception," *Perception*, vol. 6, no. 3, pp. 313–326, 1977.
- [33] J. M. Henderson, P. A. Weeks Jr, and A. Hollingworth, "The effects of semantic consistency on eye movements during complex scene viewing.," *Journal of experimental psychology: Human perception and performance*, vol. 25, no. 1, p. 210, 1999.
- [34] I. I. Gottesman and J. Shields, *Schizophrenia*. CUP Archive, 1982.
- [35] R. G. Ross, A. Olinicy, J. G. Harris, B. Sullivan, and A. Radant, "Smooth pursuit eye movements in schizophrenia and attentional dysfunction: adults with schizophrenia, adhd, and a normal comparison group," *Biological psychiatry*, vol. 48, no. 3, pp. 197–203, 2000.
- [36] G. Weiss and L. T. Hechtman, *Hyperactive children grown up: ADHD in children, adolescents, and adults*. Guilford Press, 1993.
- [37] K. Chawarska, A. Klin, and F. Volkmar, "Automatic attention cueing through eye movement in 2-year-old children with autism," *Child development*, vol. 74, no. 4, pp. 1108–1122, 2003.

- [38] K. A. Pelphrey, N. J. Sasson, J. S. Reznick, G. Paul, B. D. Goldman, and J. Piven, “Visual scanning of faces in autism,” *Journal of autism and developmental disorders*, vol. 32, no. 4, pp. 249–261, 2002.
- [39] M. A. Martens, S. J. Wilson, and D. C. Reutens, “Research review: Williams syndrome: a critical review of the cognitive, behavioral, and neuroanatomical phenotype,” *Journal of Child Psychology and Psychiatry*, vol. 49, no. 6, pp. 576–608, 2008.
- [40] A. Klin, W. Jones, R. Schultz, F. Volkmar, and D. Cohen, “Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism,” *Archives of general psychiatry*, vol. 59, no. 9, pp. 809–816, 2002.
- [41] M. Startsev and M. Dorr, “Classifying autism spectrum disorder based on scanpaths and saliency,” in *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 633–636, IEEE, 2019.
- [42] S. Wang, M. Jiang, X. M. Duchesne, E. A. Laugeson, D. P. Kennedy, R. Adolphs, and Q. Zhao, “Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking,” *Neuron*, vol. 88, no. 3, pp. 604–616, 2015.
- [43] S. Eraslan, Y. Yesilada, V. Yaneva, and S. Harper, “Autism detection based on eye movement sequences on the web: a scanpath trend analysis approach,” in *Proceedings of the 17th International Web for All Conference*, pp. 1–10, 2020.
- [44] V. Yaneva, L. A. Ha, S. Eraslan, and Y. Yesilada, “Adults with high-functioning autism process web pages with similar accuracy but higher cognitive effort compared to controls,” in *Proceedings of the 16th Web For All 2019 Personalization-Personalizing the Web*, pp. 1–4, 2019.
- [45] S. Eraslan, V. Yaneva, Y. Yesilada, and S. Harper, “Web users with autism: eye tracking evidence for differences,” *Behaviour & Information Technology*, vol. 38, no. 7, pp. 678–700, 2019.
- [46] T. Akter, M. H. Ali, M. I. Khan, M. S. Satu, and M. A. Moni, “Machine learning model to predict autism investigating eye-tracking dataset,” in *2021 2nd*

*International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*, pp. 383–387, IEEE, 2021.

- [47] B. Li, E. Barney, C. Hudac, N. Nuechterlein, P. Ventola, L. Shapiro, and F. Shic, “Selection of eye-tracking stimuli for prediction by sparsely grouped input variables for neural networks: towards biomarker refinement for autism,” in *ACM Symposium on Eye Tracking Research and Applications*, pp. 1–8, 2020.
- [48] R. Carette, M. Elbattah, F. Cilia, G. Dequen, J.-L. Guérin, and J. Bosche, “Learning to predict autism spectrum disorder based on the visual patterns of eye-tracking scanpaths.,” in *HEALTHINF*, pp. 103–112, 2019.
- [49] R. Carette, F. Cilia, G. Dequen, J. Bosche, J.-L. Guerin, and L. Vandromme, “Automatic autism spectrum disorder detection thanks to eye-tracking and neural network-based approach,” in *International Conference on IoT Technologies for HealthCare*, pp. 75–81, Springer, 2017.
- [50] Y. Tao and M.-L. Shyu, “Sp-asdnet: Cnn-lstm based asd classification model using observer scanpaths,” in *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 641–646, IEEE, 2019.
- [51] S. Chen and Q. Zhao, “Attention-based autism spectrum disorder screening with privileged modality,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1181–1190, 2019.
- [52] M. Jiang and Q. Zhao, “Learning visual attention to identify people with autism spectrum disorder,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3267–3276, 2017.
- [53] C. Wu, S. Liaqat, S.-c. Cheung, C.-N. Chuah, and S. Ozonoff, “Predicting autism diagnosis using image with fixations and synthetic saccade patterns,” in *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 647–650, IEEE, 2019.
- [54] G. Arru, P. Mazumdar, and F. Battisti, “Exploiting visual behaviour for autism spectrum disorder identification,” in *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 637–640, IEEE, 2019.

- [55] P. Mazumdar, G. Arru, and F. Battisti, “Early detection of children with autism spectrum disorder based on visual exploration of images,” *Signal Processing: Image Communication*, vol. 94, p. 116184, 2021.
- [56] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, “Applying machine learning to kinematic and eye movement features of a movement imitation task to predict autism diagnosis,” *Scientific Reports*, vol. 10, no. 1, pp. 1–13, 2020.
- [57] M. Jiang, S. M. Francis, D. Srishyla, C. Conelea, Q. Zhao, and S. Jacob, “Classifying individuals with asd through facial emotion recognition and eye-tracking,” in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6063–6068, IEEE, 2019.
- [58] W. Liu, M. Li, and L. Yi, “Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework,” *Autism Research*, vol. 9, no. 8, pp. 888–898, 2016.
- [59] J. Kang, X. Han, J. Song, Z. Niu, and X. Li, “The identification of children with autism spectrum disorder by svm approach on eeg and eye-tracking data,” *Computers in Biology and Medicine*, p. 103722, 2020.
- [60] G. Wan, X. Kong, B. Sun, S. Yu, Y. Tu, J. Park, C. Lang, M. Koh, Z. Wei, Z. Feng, *et al.*, “Applying eye tracking to identify autism spectrum disorder in children,” *Journal of autism and developmental disorders*, vol. 49, no. 1, pp. 209–215, 2019.
- [61] J. Li, Y. Zhong, and G. Ouyang, “Identification of asd children based on video data,” in *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 367–372, IEEE, 2018.
- [62] J. H. Elder, C. M. Kreider, S. N. Brasher, and M. Ansell, “Clinical impact of early diagnosis of autism on the prognosis and parent–child relationships.,” *Psychology research and behavior management*, 2017.
- [63] S. Canavan, M. Chen, S. Chen, R. Valdez, M. Yaeger, H. Lin, and L. Yin, “Combining gaze and demographic feature descriptors for autism classification,” in *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 3750–3754, IEEE, 2017.

- [64] S. Russell and P. Norvig, “Artificial intelligence: a modern approach,” 2002.
- [65] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, “Lstm: A search space odyssey,” *IEEE transactions on neural networks and learning systems*, vol. 28, no. 10, pp. 2222–2232, 2016.
- [66] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [67] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein, *Logistic regression*. Springer, 2002.
- [68] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [69] H. Duan, G. Zhai, X. Min, Z. Che, Y. Fang, X. Yang, J. Gutiérrez, and P. L. Callet, “A dataset of eye movements for the children with autism spectrum disorder,” in *Proceedings of the 10th ACM Multimedia Systems Conference*, pp. 255–260, 2019.
- [70] W. Wei, Z. Liu, L. Huang, A. Nebout, and O. Le Meur, “Saliency prediction via multi-level features and deep supervision for children with autism spectrum disorder,” in *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 621–624, IEEE, 2019.
- [71] S. Halligan, D. G. Altman, and S. Mallett, “Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach,” *European radiology*, vol. 25, no. 4, pp. 932–939, 2015.
- [72] D. M. Powers, “Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation,” 2011.
- [73] D. Fabiano, S. Canavan, H. Agazzi, S. Hinduja, and D. Goldgof, “Gaze-based classification of autism spectrum disorder,” *Pattern Recognition Letters*, 2020.
- [74] A. Nebout, W. Wei, Z. Liu, L. Huang, and O. Le Meur, “Predicting saliency maps for asd people,” in *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 629–632, IEEE, 2019.

- [75] R. M. Thomas, S. Gallo, L. Cerliani, P. Zhutovsky, A. El-Gazzar, and G. van Wingen, “Classifying autism spectrum disorder using the temporal statistics of resting-state functional mri data with 3d convolutional neural networks,” *Frontiers in Psychiatry*, vol. 11, p. 440, 2020.
- [76] F. Saeed, T. Eslami, V. Mirjalili, A. Fong, and A. Laird, “Asd-diagnet: A hybrid learning approach for detection of autism spectrum disorder using fmri data,” *Frontiers in Neuroinformatics*, vol. 13, p. 70, 2019.
- [77] D. Haputhanthri, G. Brihadiswaran, S. Gunathilaka, D. Meedeniya, Y. Jayawardena, S. Jayarathna, and M. Jaime, “An eeg based channel optimized classification approach for autism spectrum disorder,” in *2019 Moratuwa Engineering Research Conference (MERCon)*, pp. 123–128, IEEE, 2019.
- [78] B. Li, A. Sharma, J. Meng, S. Purushwalkam, and E. Gowen, “Applying machine learning to identify autistic adults using imitation: An exploratory study,” *PloS one*, vol. 12, no. 8, p. e0182652, 2017.
- [79] Y. Nakai, T. Takiguchi, G. Matsui, N. Yamaoka, and S. Takada, “Detecting abnormal word utterances in children with autism spectrum disorders: machine-learning-based voice analysis versus speech therapists,” *Perceptual and motor skills*, vol. 124, no. 5, pp. 961–973, 2017.
- [80] A. Di Martino, D. O’connor, B. Chen, K. Alaerts, J. S. Anderson, M. Assaf, J. H. Balsters, L. Baxter, A. Beggiano, S. Bernaerts, *et al.*, “Enhancing studies of the connectome in autism using the autism brain imaging data exchange ii,” *Scientific data*, vol. 4, no. 1, pp. 1–15, 2017.
- [81] T. Eslami and F. Saeed, “Auto-asd-network: a technique based on deep learning and support vector machines for diagnosing autism spectrum disorder using fmri data,” in *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pp. 646–651, 2019.
- [82] A. S. Heinsfeld, A. R. Franco, R. C. Craddock, A. Buchweitz, and F. Meneguzzi, “Identification of autism spectrum disorder using deep learning and the abide dataset,” *NeuroImage: Clinical*, vol. 17, pp. 16–23, 2018.



- [83] Y. Liu, L. Xu, J. Li, J. Yu, and X. Yu, "Attentional connectivity-based prediction of autism using heterogeneous rs-fmri data from cc200 atlas," *Experimental Neurobiology*, vol. 29, no. 1, p. 27, 2020.
- [84] S. Rane, E. Jolly, A. Park, H. Jang, and C. Craddock, "Developing predictive imaging biomarkers using whole-brain classifiers: Application to the abide i dataset," *Research Ideas and Outcomes*, vol. 3, p. e12733, 2017.
- [85] C. P. Chen, C. L. Keown, A. Jahedi, A. Nair, M. E. Pflieger, B. A. Bailey, and R.-A. Müller, "Diagnostic classification of intrinsic functional connectivity highlights somatosensory, default mode, and visual regions in autism," *NeuroImage: Clinical*, vol. 8, pp. 238–245, 2015.
- [86] K. Gotham, S. Risi, A. Pickles, and C. Lord, "The autism diagnostic observation schedule: revised algorithms for improved diagnostic validity," *Journal of autism and developmental disorders*, vol. 37, no. 4, p. 613, 2007.
- [87] S. Baron-Cohen, S. Wheelwright, R. Skinner, J. Martin, and E. Clubley, "The autism-spectrum quotient (aq): Evidence from asperger syndrome/high-functioning autism, males and females, scientists and mathematicians," *Journal of autism and developmental disorders*, vol. 31, no. 1, pp. 5–17, 2001.
- [88] E. Michailidou, "Vicram: Visual complexity rankings and accessibility metrics," *Accessibility and Computing*, p. 24, 2010.
- [89] G. Marchionini, "Exploratory search: from finding to understanding," *Communications of the ACM*, vol. 49, no. 4, pp. 41–46, 2006.
- [90] M. J. Gingerich and C. Conati, "Constructing models of user and task characteristics from eye gaze data for user-adaptive information highlighting," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.
- [91] S. Eraslan and Y. Yesilada, "Patterns in eyetracking scanpaths and the affecting factors.," *J. Web Eng.*, vol. 14, no. 5&6, pp. 363–385, 2015.
- [92] M. E. Akpınar and Y. Yesilada, "Vision based page segmentation algorithm: Extended and perceived success," in *International Conference on Web Engineering*, pp. 238–252, Springer, 2013.

- [93] D. M. Harris, S. L. Harris, P. Prinz, and T. Crawford, “Digital design and computer architecture,” 2019.
- [94] T. K. Kim, “T test as a parametric statistic,” *Korean journal of anesthesiology*, vol. 68, no. 6, p. 540, 2015.
- [95] H. B. Mann and D. R. Whitney, “On a test of whether one of two random variables is stochastically larger than the other,” *The annals of mathematical statistics*, pp. 50–60, 1947.
- [96] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, “Array programming with NumPy,” *Nature*, vol. 585, pp. 357–362, Sept. 2020.
- [97] T. pandas development team, “pandas-dev/pandas: Pandas,” Feb. 2020.
- [98] M. L. Waskom, “seaborn: statistical data visualization,” *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021.
- [99] J. Tolles and W. J. Meurer, “Logistic regression: relating patient characteristics to outcomes,” *Jama*, vol. 316, no. 5, pp. 533–534, 2016.
- [100] G. Rodríguez, “Lecture notes on generalized linear models,” URL: <http://data.princeton.edu/wws509/notes/c4.pdf>, 2007.
- [101] N. S. Altman, “An introduction to kernel and nearest-neighbor nonparametric regression,” *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.
- [102] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.
- [103] C. E. Shannon, “A mathematical theory of communication,” *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.
- [104] C. Gini, “Variabilità e mutabilità,” *vamu*, 1912.

- [105] S. Kullback and R. A. Leibler, “On information and sufficiency,” *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.
- [106] S. R. Safavian and D. Landgrebe, “A survey of decision tree classifier methodology,” *IEEE transactions on systems, man, and cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.
- [107] T. K. Ho, “Random decision forests,” in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1, pp. 278–282, IEEE, 1995.
- [108] D. Opitz and R. Maclin, “Popular ensemble methods: An empirical study,” *Journal of artificial intelligence research*, vol. 11, pp. 169–198, 1999.
- [109] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.
- [110] L. Bianchi, M. Dorigo, L. M. Gambardella, and W. J. Gutjahr, “A survey on metaheuristics for stochastic combinatorial optimization,” *Natural Computing*, vol. 8, no. 2, pp. 239–287, 2009.
- [111] Y. Freund, R. Schapire, and N. Abe, “A short introduction to boosting,” *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771-780, p. 1612, 1999.
- [112] M. Kearns, “Thoughts on hypothesis boosting,” *Unpublished manuscript*, vol. 45, p. 105, 1988.
- [113] L. Breiman, “Bias, variance, and arcing classifiers,” tech. rep., Tech. Rep. 460, Statistics Department, University of California, Berkeley . . . , 1996.
- [114] L. Breiman, “Arcing the edge,” tech. rep., Technical Report 486, Statistics Department, University of California at . . . , 1997.
- [115] C. E. Rasmussen, “Gaussian processes in machine learning,” in *Summer School on Machine Learning*, pp. 63–71, Springer, 2003.
- [116] I. Rish *et al.*, “An empirical study of the naive bayes classifier,” in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, pp. 41–46, 2001.

- [117] G. H. John and P. Langley, “Estimating continuous distributions in bayesian classifiers,” *arXiv preprint arXiv:1302.4964*, 2013.
- [118] A. McCallum, K. Nigam, *et al.*, “A comparison of event models for naive bayes text classification,” in *AAAI-98 workshop on learning for text categorization*, vol. 752, pp. 41–48, Citeseer, 1998.
- [119] A. M. Martínez and A. C. Kak, “Pca versus lda,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 2, pp. 228–233, 2001.
- [120] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.
- [121] A. Tharwat, “Linear vs. quadratic discriminant analysis classifier: a tutorial,” *International Journal of Applied Pattern Recognition*, vol. 3, no. 2, pp. 145–180, 2016.
- [122] K. K. Hyde, M. N. Novack, N. LaHaye, C. Parlett-Pelleriti, R. Anden, D. R. Dixon, and E. Linstead, “Applications of supervised machine learning in autism spectrum disorder research: a review,” *Review Journal of Autism and Developmental Disorders*, vol. 6, no. 2, pp. 128–146, 2019.
- [123] L. Fusar-Poli, N. Brondino, P. Politi, and E. Aguglia, “Missed diagnoses and misdiagnoses of adults with autism spectrum disorder,” *European archives of psychiatry and clinical neuroscience*, pp. 1–12, 2020.