Machine Learning Classification of User Attributes Via Eye Movements

von Sahar Mahdie Klim Al-Zaidawi

Dissertation

zur Erlangung des Grades eines Doktors der
Ingenieurwissenschaften
- Dr.-Ing. -

Vorgelegt im Fachbereich 3 (Mathematik & Informatik)
der Universität Bremen
im March 2022

Datum des Promotionskolloquiums: 22.04.2022

Gutachter: Prof. Dr. Sebastian Maneth (Universität Bremen)
Prof. Dr. Mehul Bhatt (Örebro University)

Abstract

The advent of modern eye tracking devices has spawned a plethora of new research on eye movements. Applications of these research results include the prediction of diseases, of biometrics, of gender, or of cognitive developments in children. One particularly well studied topic is user identification. Another, less well studied one is gender prediction. In this thesis, a common framework to predict users and gender is proposed. Using this framework, we were able to improve the state-of-the-art accuracies for both user identification and gender prediction. Further, unlike previous studies, the proposed approach was tested with different datasets consisting of varying stimuli. We identify several factors that affect the identification accuracy. Our main improvements in identification accuracy are due to three factors, selecting optimal hyper-parameters of the segmentation algorithm, adding higher-order derivatives, and including blink information. For gender prediction, the thesis establishes several new insights. For instance, that gender prediction is possible for prepubescent children aged 9–10. Previous research had suggested that significant gender differences in eye movements can only be observed in adults. Various factors are identified which affect the accuracy of gender prediction; for example, the length of the gaze trajectory, possible fatigue of the participant (gender prediction works better in the presence of fatigue), and the choice of feature ranking algorithms.

Zusammenfassung

Deutsche Zusammenfassung Das Aufkommen moderner Eye-Tracking-Geräte hat eine Fülle neuer Forschungsarbeiten über Augenbewegungen hervorgebracht. Zu den Anwendungen dieser Forschungsergebnisse gehören die Vorhersage von Krankheiten, der Biometrie, des Geschlechts oder der kognitiven Entwicklung von Kindern. Ein besonders gut untersuchtes Thema ist die Identifizierung von Benutzern. Ein anderes, weniger gut untersuchtes Thema ist die Vorhersage des Geschlechts. In dieser Arbeit wird ein gemeinsames Modell für die Vorhersage von Nutzern und Geschlecht vorgeschlagen. Mithilfe dieses Rahmens konnten wir den Stand der Technik in Bezug auf Genauigkeiten sowohl bei der Benutzeridentifikation als auch bei der Vorhersage des Geschlechts verbessern. Außerdem wurde im Gegensatz zu früheren Studien der vorgeschlagene Ansatz mit verschiedenen Datensätzen getestet, die aus unterschiedlichen Stimuli bestchen. Wir konnten mehrere Faktoren identifizieren, die die Identifikationsgenauigkeit beeinflussen. Unsere Hauptverbesserungen in der Identifikationsgenauigkeit sind auf drei Faktoren zurückzuführen: die Auswahl der Hyperparameter des Segmentierungsalgorithmus, das Hinzufügen von Ableitungen höherer Ordnung und die Einbeziehung von Blinzelinformationen. Für die Geschlechtsvorhersage liefert die Arbeit mehrere neue Erkenntnisse. Zum Beispiel, dass die Geschlechtsvorhersage möglich ist für vorpubertäre Kinder im Alter von 9–10 Jahren. Frühere Forschungen hatten ergeben, dass signifikante geschlechtsspezifische Unterschiede in den Augenbewegungen nur bei Erwachsenen beobachtet werden können. Verschiedene Faktoren wurden identifiziert, die die Genauigkeit der Geschlechtsvorhersage beeinflussen, z. B. die Länge des Blickverlaufs, mögliche Ermüdung des Teilnehmers (die Vorhersage des Geschlechts funktioniert bei Müdigkeit besser) und die Wahl der Algorithmen zur Merkmalsbewertung.

Danksagung

The work presented in this PhD thesis is a result of drawing support from various sources. It is a matter of great pleasure to express my appreciation and gratitude to those who have been contributing to this work.

Firstly, I would like to thank my supervisor Prof. Dr. Sebastian Maneth the head of the database group at the University of Bremen to give me an opportunity to work on this interesting topic in his group. I am very grateful for his encouragement throughout the course of this thesis and his unconditional support. Also, I learned alot from him, for instance the art of scientific writing from him.

Secondly, I would like to take this opportunity to thank Christoph Schröder PhD student in the Institute for Computer Graphics and Virtual Reality at the University of Bremen I discussed my first ideas of biometrics-based on eye moment with him to which he was very welcoming and always supported me scientifically and invited me for many machine learning talks with his team. I also thank my colleague Martin H.U. Prinzler a PhD student in the Database group at the University of Bremen for his continuous guidance and constructive feedback during this thesis work. I consider myself fortunate to be a part of the Database group which provided me with wonderful colleagues that supported me, Dr. Peter Leupold and Martin Vu who took the time to proofread my dissertation and provide me with their useful feedback. Also, I would like to thank my master thesis student turned colleague Rishabh Haria who worked hard and contributed to part of my work indirectly.

Thirdly, I would like to express my gratitude and acknowledge the support of the DAAD PhD scholarship (award number 91645228) for funding my research. Additionally, I thank Prof. Dr. Mehul Bhatt for accepting me as a PhD student in Germany and providing me an opportunity to come here.

Fourthly, I would like to thank various brilliant friends from different Universities or Institutes who helped with providing me with their valuable feedback or proofreading my thesis. Like Chandni Sidhu who is working as a research scientist at MPI for Marine Microbiology, Adebayo Emmanuel Abejide who is working in the IT-Institute of Telecommunications at the University of Aveiro in Aveiro, Portugal, and Andrew Adewale Alola who is working as Assistant Professor at University of Vaasa in Finland. I also would always be thankful to my special friend Shivesh Kumar a senior researcher at DFKI for all the scientific discussions that we had and the time he spent with me. His feedback has always been crucial for improving my work. When I look back, I would not be where I am without his support, so once again my big thanks to him. Right from the beginning he has been of great help and was the person I could reach anytime and ask any questions regarding work or life in Germany. I have a genuine appreciation for his friendliness and patience.

Lastly, without the support of friends and family the PhD life won't be easy. I am grateful for all my friends in Bremen Anke, Rosa, Veronika, Chandani, Vipul, Roberta, Khadeeja, Ameena, and Mihaela for creating some beautiful memories during my PhD life. I would be eternally grateful to my family (especially my father) for their love and encouragement in shaping me into who I am today.

Contents

Chapter 1

Introduction

This chapter is the entry point of this thesis and serves five main purposes: 1) provide the preliminaries of the topic and shows its relevance, 2) explain the motivation and the main objectives of this thesis , 3) give a brief description of the methodology developed and highlight the main contributions., 4) describe the structure of the thesis., and 5) list the scientific publications on which this thesis is based.

1.1 Preliminaries

Eye movements can comprise rich and sensitive information about an individual, including biometric identity, gender, age, ethnicity, personality traits, drug consumption habits, moods and emotions, skills, preferences, cognitive processes, and physical and mental health conditions [Kröger et al., 2020]. Eye movements are an excellent predictor of human desires and focus. They are inextricably linked with human cognitive and perceptual processes. A brief summary of the human visual system and the types of eye movements is illustrated in the following.

1.1.1 Human Visual System

People see the world with their eyes which are located side by side. The same object in the world is seen by both eyes separately, each eye generating an independent signal representing its visual field. Then a unified image from these signals is produced by the brain [Iqbal, 2012].

According to [Wikipedia, 2022], there were already ancient Greek schools that provided a primitive explanation of how the human vision works. However, the first modern studies of visual perception are often credited to Hermann von Helmholtz. Hermann concluded that vision is a result of "unconscious inference", he coined that term after his study on the human eyes and his conclusion was that the eyes are un-

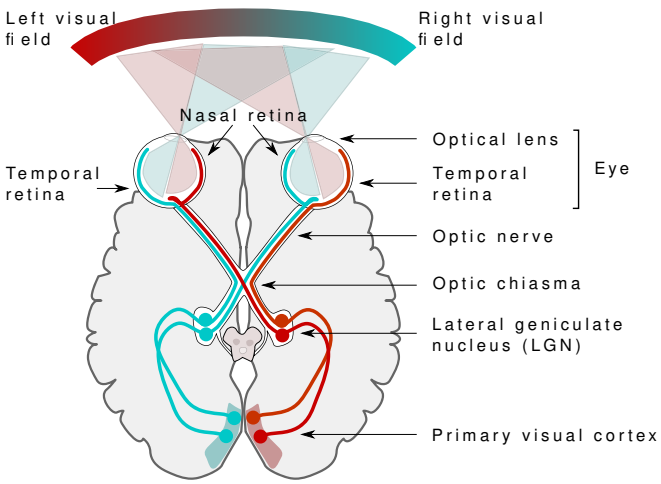


Figure 1.1: Human visual system [Nieto, 2015], CC 4.0

able to produce a high-quality image and the vision is impossible without sufficient information. Hermann suggested that the brain, based on previous experiences is making presumptions and conclusions from incomplete data [von Helmholtz, 1925]. These studies opened doors to several investigations in the field of vision systems. One of such was the assertion that relates about 50 percent of neurons in the brain to visual information processing [Jenkin, 2012]. This fact is not surprising since virtually every visual engagement is connected to all human activities. In simple words, we interpret colors, classify objects, detect and analyze motion, estimate distances, and create a three-dimensional reality from the two-dimensional pictures that fall on our retinas.

Both eyes and brain are part of the visual system. Light enters our eyes and contacts the retina, causing light receptors to transmit electrical signals via our optic nerve to the back of the brain, where the first stages of visual perception occur. The brain then sends progressively filtered signals until we can see and take action [Vanderah and Gould, 2020]. Figure 1.1 represents the schematic view of the human visual system.

1.1.2 Eye Movements

An eye movement is a motion in the eye's position in its orbit, and there are different types of these movements [Babin et al., 2006]. Eye movements enable diverse visual information to reach to the retina, the visual cortex, and the higher cortex's center responsible for language, vision, recognizing objects in space (visuospatial recognition), and awareness. Therefore, eye movements are essential for vision, attention, and

memory. Many scientific studies suggested that brain circuits that perform specific executive functions, such as spatial attention and spatial working memory, overlap with some brain circuits that control eye movements [Babin et al., 2006].

The first ophthalmologist who reported that the eyes movements are not continuous is Louis Émile Javal in 1879. He found that the eye movements make short and rapid movements (saccades) intermingled with short stops (fixations). The French term 'saccade' was suggested by Dodge, in 1916 to describe the rapid movements of the eyes that occur while reading. These rapid movements were observed before L.E. Javal's studies and were called jerks in the English language [Wade et al., 2003].

To provide comprehensive knowledge of the visual system, different eye movements are explained in the following. They can be extracted from the raw eye tracking data by different algorithms (e.g. identification velocity threshold algorithm) or are sometimes readily available by the eye tracker (e.g. Tobii eye-tracker).

- **Fixations:** The fixations are the stages when the eyes are focused on specific regions for short time.
- **Saccades:** Saccades are rapid eye movements between every two consecutive fixations that enable a visual environment to be scanned.
- **Smooth Pursuit:** These eye movements are significantly slower than saccades and are meant to maintain a moving stimulus on the fovea (area of the retina). Such movements are under voluntary control so that the observer can track a moving stimulus [Purves et al., 2001]. An example of smooth pursuit is following a ball while watching a football match.
- **Vergence:** Vergence is related to the capacity of the eyes to focus on different things in the 3D world. It entails the eyes moving in opposing directions (e.g., more left and more right) to point at the same spot in a visual scene [Banks et al., 2012]. The two forms of vergence are convergence (in which the eyes point closer together) and divergence (in which the eyes point away from each other). Convergence allows us to focus on nearby items, whereas divergence helps us concentrate on distant things. It is also aided by altering the lens and pupil shape.
- **Vestibulo Ocular movements:** Vestibulo ocular movements help steady the eyes concerning the outside environment, correcting for head movements [Evans et al., 2012]. These reflex reactions keep visual pictures from "slipping" on the retina's surface when the head turns. By fixating an item and moving the head from side to side, we may see the operation of vestibulo-ocular motions. The eyes adjust for head movement by moving the same distance but in the opposite direction, maintaining the picture of the object in

roughly the same spot on the retina. The vestibular system detects abrupt, temporary changes in head position and generates quick corrective eye movements [Faan et al., 2010, Leigh and Zee, 2015].

- **Optokinetic response movements / post-rotatory Nystagmus:** Post-rotatory nystagmus occurs when we spin around in one place too often. As a result, the eyes move in the opposite direction to compensate for the quick movement. This action happens also in the absence of light, indicating that the vestibular system is involved in controlling this movement. These Optokinetic response movements are types of nystagmus. However, they are not pathological (they are present in normal eye movement behavior) [Sperry, 1950].

1.1.3 Eye Tracking Data and its Application

Eye tracking is a method to record a persons' eye movements using an eye tracking device. An eye tracker is a sensor technology that can identify the presence of a person and monitor what they are looking at in real-time. The first non-invasive and precise eye tracking technique, using light reflected from the cornea was developed by Dodge and Cline in 1901. Later, vast technological advances in tracking eye movements during the 1960s and1970s are still seen in the most commercially available eye tracking systems today. In the 1980s, computers became powerful enough to do eye tracking in real-time which enabled the application of video-based eye trackers to human-computer interaction. A brief history of the eye tracking devices developed in the past years is provided in [Jacob and Karn, 2003]. The system turns eye movements into a data stream including pupil position, gaze vector for each eye, and gaze point. Essentially, the technology decodes eye movements and converts them into insights that may be utilized in various applications or as a secondary input modality. An eye tracking system typically consists of cameras, specific light sources, and processing capabilities. Algorithms use sophisticated image processing to convert camera feeds into data points. Using eye tracking, it can be determined where users are looking at a given moment in time, how long they are looking at any object, and the route their eyes take when looking at any visual stimuli. Figure 1.2 shows the basic eye tracking principle. Researchers can monitor the motions of a participant's eyes during various activities using eye tracking.

Today, a vast number of different eye tracking devices are available and have been used by researchers to produce many high-quality datasets. These datasets are analyzed in different contexts; for example: eye tracking data has been used in medical research for diagnosing diseases such as dyslexia [Benfatto et al., 2016], and other examples of disorder detection [Armstrong and Olatunji, 2012, Billeci et al., 2017], gaming [Lin et al., 2004, Alkan and Cagiltay, 2007, Lankes and Stoeckl, 2020,

Figure 1.2: A participant is using a PC that has EyeLink1000 eye tracker attached to it [EyeLink, 2022].

Isokoski et al., 2009, Majaranta and Bulling, 2014], automotive research to study the visual attention of drivers [Crundall and Underwood, 2011], gender prediction [Sargezeh et al., 2019, Moss et al., 2012, Zaidawi et al., 2020], and user identification [Kasprowski and Ober, 2004, George and Routray, 2016, Rigas and Komogortsev, 2017, Schröder et al., 2020].

Let us describe the eye tracking measures with reading stimuli examples. People generally believe that the eyes move smoothly over the text during reading, determining information word by word [Rayner, 1998]. Nevertheless, this belief is incorrect. Saccades are the reason why they may recognize words in such a short period. Saccades occur between times of stabilization of eye movements, known as fixations (see Section 1.1.2). The reader evaluates the visual information during fixations. The eyes typically tend to stay on new, complicated, or technically wrong phrases and various sections of a word. Words that have already been read and already been processed extensively. When reading aloud, readers may ignore up to 30% of the words on the first read-through, these are generally short words and, sometimes, entirely predictable words.

With the assumption that participants concentrate on the text within their foveal vision (the small frame of sight that endures in focus), rather than that information contained within their parafoveal or peripheral vision, eye trackers measure the gaze coordinates of the eye movement trajectory based on which fixation and saccade information can be extracted. Several common eye tracking measurements for example, fixations and saccades, are shown in Figure 1.3.

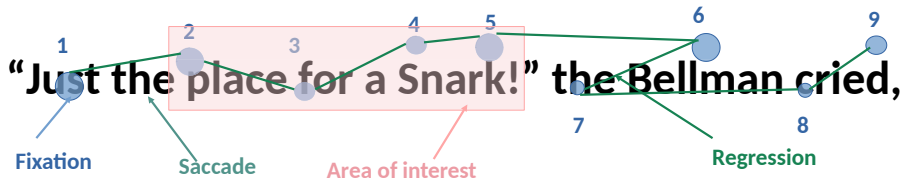


Figure 1.3: Eye-movement metrics, which are often used in eye tracking research

1.2 Motivation and Objectives

The general objective of this thesis to perform biometrics and gender prediction using eye tracking data. It provides a consistent approach that can improve the state of the art accuracy for these two tasks.

1.2.1 User Identification

Traditionally user identification has been performed in different ways for example using voice recognition, face recognition, hand recognition, signature, and fingerprints. This information can be used in different applications in the real world in the security field (for example criminal detection, identifying the person, and performing access control for different purposes e.g. financial statements or mobile phones access, at airports, or border entry recognition). However, these individual methods are prone to failure and misuse. Therefore, nowadays many researchers are investigating eye tracking data to achieve user identification [Babich, 2012] as a viable alternative.

In the 2015 BioEye competition [Rigas and Komogortsev, 2017], different systems for user identification based on eye movement were competing. The winning system is based on first segmenting the eye tracking trajectory into two types of eye movements fixations and saccades. The latter is achieved using the Identification Velocity Threshold (IVT) method (see e.g. [Salvucci and Goldberg, 2000], where this method is analyzed and some prior references are given). The IVT idea is based on the observation that low velocities characterize fixations, while high velocities characterize saccades. The IVT algorithm uses the *velocity threshold* (VT) parameter and labels segments of the given eye tracking trajectory as fixations, whenever the velocity within those segments is below the VT and the remaining segments as saccades. Later, this algorithm was extended by introducing the *minimum fixation duration* (MFD) parameter, so that segments that have been labeled as fixations but are shorter than the MFD are merged to saccades. After segmentation, a number of features (such as duration, velocity, acceleration etc.) are computed for each segment. These fixation and saccade based features are used to build two dedicated

machine learning based classifiers[1] (each for fixation and saccade) which outputs the prediction probabilities for each class (user ID in this case). The average of both classifiers is used as the final prediction probability and the class corresponding to maximum prediction probability is the identified user ID.

The winning system IVT version uses the above described two parameters which are set to certain default values. It was not investigated, whether other settings of these parameters would give higher identification accuracies. Also, in the winning system, the used features were until acceleration level (i.e. 2nd order derivative of gaze trajectory) and using only fixation and saccade segments. Therefore, it is interesting to check if tuning IVT parameters and increasing/decreasing the features or add higher-order derivatives will increase the user identification accuracy. Furthermore, it is known that the eye tracking data can contain invalid data which can be due to user-specific reasons such as blinking, loss of attention (micro-sleeping) or eye tracker faults (e.g. solo missed gaze points). However, the majority of the invalid data is caused by blinks. Therefore, it will also be interesting to investigate if adding new segments e.g. blinks in addition to fixations and saccades will increases the biometrics accuracy. Moreover, most of the previous studies of eye movement biometrics have tested their approach using *only small* numbers of participants, utilizing either *only one or two* datasets e.g. [George and Routray, 2016, Jäger et al., 2019, Krishna et al., 2019]. Furthermore, since most prior studies train and assess their models on datasets compiled over a short time span, the permanence of eye movements remains unexamined.

The first main goal of this thesis is to fill the above gaps by assessing these methods with different datasets containing a larger number of participants and some that were not utilized for eye movement biometrics. One of the aims of this thesis is to investigate the impact of including these factors: tuning IVT parameters, adding higher-order derivative features, segmenting blinks, adding a third classifier fed by blinking features on the biometrics accuracy, we would like to develop a general approach for user prediction that works consistently and robustly across different datasets of varying stimuli and which improves the state-of-the-art.

Overall, the user prediction approach developed in this thesis is tested on a total of four datasets. Two datasets from the 2015 Bioeye competition are used [Rigas and Komogortsev, 2017]. They consists of 153 participants who looked

[1]Classification is the process of predicting the class which categorizes similar data points. Classes are sometimes called as targets/labels or categories. Machine learning based classifiers for e.g. random decision trees, Naive Bayes, SVM etc. are some of the popular methods.

at two different stimuli: random moving dots and a poem. The third biometrics dataset is about a visual searching task [Li et al., 2018a] and comprises 58 participants. The final dataset is known as "gaze on faces" and contains 378 participants, this dataset is used to identify “scanning strategies” that are different for men than for women [Coutrot et al., 2016]. The entire data utilized in our study provides a high number of participants and a broad range of age groups and stimuli.

1.2.2 Gender Prediction

There are a wide range of possible applications of gender prediction, e.g., improving the user experience in online learning [Philbin et al., 1995, Garland and Martin, 2005, Annetta et al., 2007], gaming [Romrell, 2014], and recently, eye tracking technology has been combined with Virtual Reality (VR) headsets [Clay et al., 2019] which opens new opportunities in virtual marketing [Research, 2020]. Gender is one important factor in the virtual shopping experience. Present e-commerce websites may be soon replaced by immersive virtual shops to enhance the experience of online shopping. The physical shops have a fixed layout for everyone regardless of age and gender and have specific zones for children, women, men, etc. which they need to find themselves. Hence, predicting certain aspects of the user can help in optimizing the shopping experience and targeted marketing. For example, if the virtual shop can predict the gender of the user with eye tracker in the VR headset, the whole shopping experience can be optimized on the fly. This motivates us to study gender prediction based on eye tracking data and in order to do so we have to understand the relationship between gender and eye movement.

An earlier study [Miyahira et al., 2000b] suggests that there is *no significant* difference between eye movements of female versus male children. One of the objectives of this work is to increase or decrease the degree of belief in this hypothesis, using state-of-the-art machine learning methods. The first used eye tracking dataset for gender prediction contains non-dyslexic and dyslexic children. In addition to investigating if it is possible to predict gender in children, this dataset gives an additional goal to measure the impact of dyslexia on gender prediction.

It is important to emphasize that gender prediction using eye tracking data and Machine Learning (ML) methods are reported in only two previous studies. Both studies are on only *healthy adults* and they have been done using *images* as stimuli. In [Mercer Moss et al., 2012] an accuracy of **64.0** % is obtained using a Naïve Bayes classifier (26 females and 26 males). In [Sargezeh et al., 2019] an accuracy of **70.0** %

was obtained using a Support Vector Machine classifier (20 females and 25 males); they speculate that women have more exploratory gaze behavior than men.

One problem with the studies mentioned above is the *instability of their results*, they only averaging their results over 5 or 10 runs[2] which means that the accuracy has a large standard deviation (SD); e.g., for the results of [Sargezeh et al., 2019] an SD of **13.2**% is reported. Another limitation of these studies mentioned is *the number of participants* used. Also, these two prior works showed another problem of using a density feature that is computed for a grid size of at least sixteen by sixteen. Thus, their number of features is *at least* 256. The use of high the number of features increases the computational requirements of the classification algorithm.

The objective of this thesis is to show if the state-of-the-art accuracy for gender prediction can be improved. Also, to show if our approach is consistently working effectively with different datasets of varying stimuli, diverse participants (children, adults, or mixed group of non-dyslexic and dyslexic), less number of features, more numbers of runs. Therefore, in addition to the dyslexia dataset (185 participants), two other dataset were used: a visual searching task which consists 58 participans and the gaze on faces data (GOF) comprising of 378.

In [Sæther et al., 2009] pictures of faces are used as stimuli and it is observed that female gaze attention was more towards the eye region as compared to males. Accordingly, another objective of this thesis is to reconfirm these findings about face stimuli, by studying the gender differences in four regions defined in the face, one for each eye, one for the nose, and one for the mouth. Then study the impact of these "regions of interest" (ROI) towards gender prediction and investigate if we are still able to predict gender when restricting the eye movement trajectories only to the ROI. Finally, we would like to check if it is possible to predict gender using only very short recordings.

1.3 Contributions

The general method and the framework developed in this thesis for both user and gender prediction is shown in Figure 1.4. We start with pre-processing the participant's eye tracking dataset. Then we segment the data into fixations and saccades

[2]A run represents a single trial of a machine learning experiment. Runs are used to monitor the execution of a trial, and store output of the trial, and to analyze results of variations in experiment conditions (for e.g. the selecting different sets of train and test data). Higher number of runs usually indicate higher stability of the results.

using the IVT algorithm. After that we extract features from fixation and saccade segments and then train different machine learning classifiers using these features to perform the user or gender prediction tasks. In the following, the main contributions of the thesis are described.

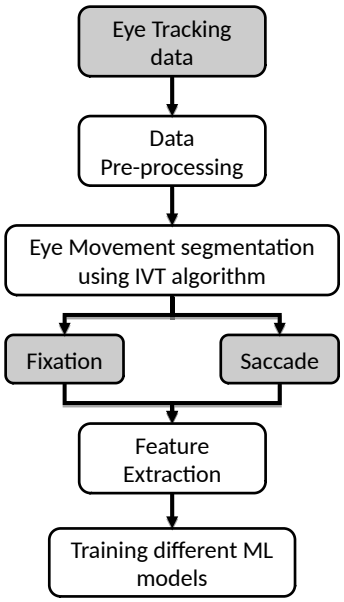


Figure 1.4: Workflow of Eye tracking experiments

1.3.1 User Identification

An extensive study of user identification via eye movements across different datasets was conducted. First, the winning pipeline in the 2015 BioEye competition [Rigas and Komogortsev, 2017], i.e. [George and Routray, 2016] was implemented[3] and extended in our initial collaborative work with Christoph Schröder and was tested on Bioeye RAN, TEX, and MIT datasets [Schröder et al., 2020]. Two extensions of the work in [George and Routray, 2016] were presented: the first extension uses more features with the previously proposed RBFN classifier which increased the classification accuracy by 5%; the second extension was the use of a different classifier namely Random Forest which also outperformed the classification accuracy in [George and Routray, 2016] by 1%. Further, in this work we investigated the degree to which classification performance depends on task-independence (weak or strong). It was found that the classification performance drops significantly when the classifier is trained on one task and tested on a different task. Also, we studied the effect of amount of train and test data on the classification accuracy. With just

[3] An exact re-implementation of the method by George and Routray and all our methods are publicly available as python module (`https://cgvr.cs.uni-bremen.de/research/smida_ml/`).

300 training samples, our method achieved accuracy of 94.7%.

Later, this work was extended in [Zaidawi et al., 2022] for the above mentioned datasets and two more datasets namely VST and GOF and the proposed approach was extensively analyzed with respect to various factors (e.g. age, gender) and extended by studying the effect of stimuli, tuning the IVT parameters, including higher order derivative features and blink information. The following are the proposed improvements:

- The IVT parameters were optimized which increased the user identification accuracy by 3 % for RAN, by 2 % for TEX, by 5 % for GOF, and by 9 % for VST.
- Adding higher derivative features increased the accuracy by 2 % for RAN, by 1 % for TEX and VST, and by 3 % for GOF.
- Adding blinking features increased the accuracy by 1 % for RAN and VST, by 0.5 % for TEX.
- Combining the above three methods of improvement (IVT parameters, higher order derivative features, blink classifier) increased the accuracy by 4 % for RAN, by 3 % for TEX, and by 9 % for VST.
- We find that user identification works better in the solo female group (accuracy of 88.85 %) than in the solo male group (accuracy of 77.37 %) with the GOF dataset.
- Similarly, user identification works better in an older age group (accuracy of 91.43 % with age group 41–72) than in a younger age group (accuracy of 85.96 % with age group 20–40) of participants.
- We analyze the effect of different tracking lengths on the performance (study Trajectory-Length using RAN, MIT, and VST data).

1.3.2 Gender Prediction

In addition to user identification, gender prediction based on eye tracking data was investigated. A method for gender prediction utilizing eye tracking data of prepubescent children (185 participants) aged 9–10 [Zaidawi et al., 2020] was proposed. Despite earlier studies demonstrating that gender variations in eye movements are only found in adults [Miyahira et al., 2000b], gender prediction accuracy up to 64% was achieved in this thesis using the dyslexia dataset. The dyslexia eye tracking dataset was employed, which includes both non-dyslexic and dyslexic young children. This information provided an extra opportunity to assess the influence of

dyslexia on gender prediction. Furthermore, two datasets (VST and GOF) were used for gender prediction, to study the gender prediction in different stimuli and with more participants. The gender prediction with VST and GOF datasets are based on [Haria et al., 2022]. The contribution of using the above datasets for gender prediction are the following:

- Contrary to previous findings, it is shown that gender prediction is possible in young children, with an accuracy of up to 64% which matches the previous accuracy obtained for healthy adults in [Mercer Moss et al., 2012].

- The impact of mixing non-dyslexic with dyslexic children for gender prediction is studied. The findings show that our ML classifiers only work well for "isolated" groups of either only non-dyslexic or only dyslexic children. For mixed groups, the obtained accuracies drop dramatically. To address this new challenge, a hierarchical classifier was constructed that makes use of the high accuracy known for dyslexia prediction. The hierarchical classifier achieves accuracies that almost match those of the isolated groups.

- Several quantitative gender differences in eye movements were confirmed (e.g. female eyes move slower with longer fixations than male eyes) known for older participants [Emam and Youssef, 2012] and students in their age of puberty [Huang and Chen, 2016] are also valid for prepubescent children.

- Gender differences in eye movements of males and females were confirmed; the findings have proved that female gazers tend to gaze more towards the left eye when the stimuli are faces. Males, on the other hand, tend to look more in the right eye or the bottom of the face. This also known from the state-of-the-art [Leonards and Scott-Samuel, 2005, Sammaknejad et al., 2017, Sæther et al., 2009].

- It is confirmed that females are more explorative than males in all the three used datasets as reported also in previous work [Sargezeh et al., 2019, Li et al., 2018b].

- It is shown that, besides face images, other stimuli can also be used for gender prediction. For this, we use data of a visual search task stimulus featuring 48 participants. An accuracy of up to **69.9 %** over this dataset was achieved.

- It is shown that "distance to previous fixation" is the highest ranked ANOVA feature in the experiment of VST data. Mean distance to previous fixation for male is approximately 1.2 times more than females.

- It is shown that the state-of-the-art accuracy for gender prediction of **70.0** % can be improved to **72.5** % using our set of features and a Random Forest classifier with using GOF dataset.
- The state-of-the-art accuracy could be improved further (from 72.5% to 77.5%) by optimizing the weights of fixation and saccade classifiers. The results show that the fixation classifier has more importance for gender prediction in our used stimuli.
- Finally gender prediction is possible using different stimuli such as visual searching task, reading, and looking at face images stimuli.

1.4 Structure of the Thesis

This thesis consists of 6 chapters which are introduction, state of art, system design, user identification work, gender prediction work, and conclusion. Additionally, there is an appendix related to chapters on user identification and gender prediction. The overall outline of this thesis and some key points of each chapter is provided in Figure 1.5.

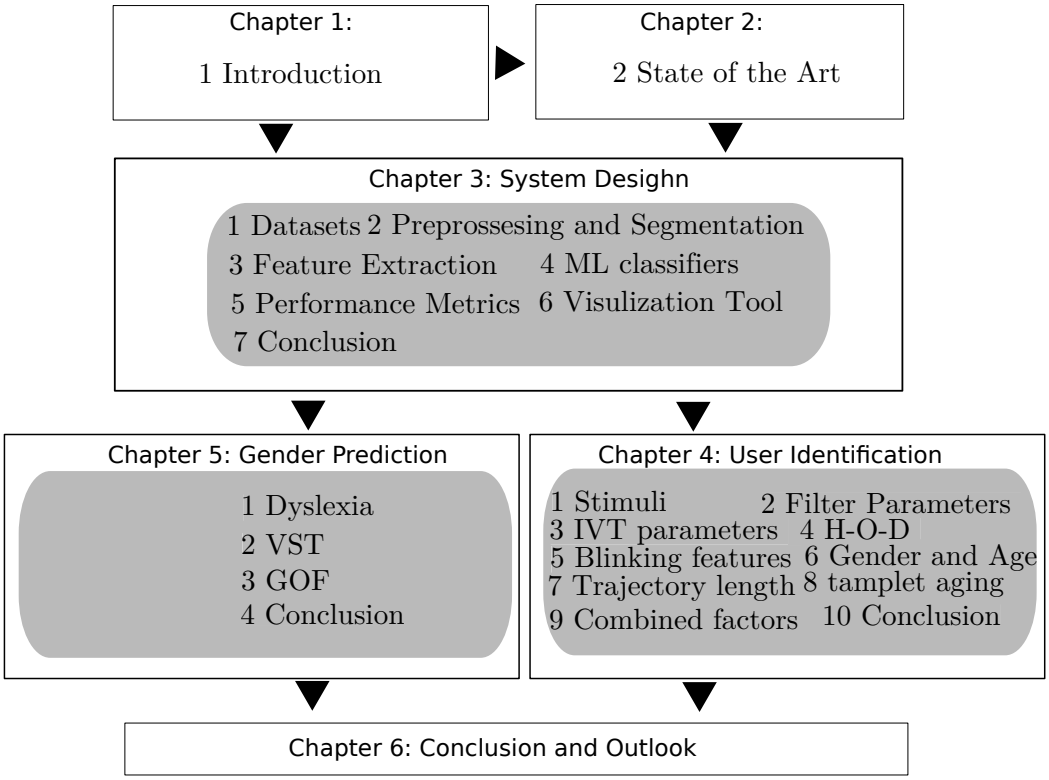


Figure 1.5: Thesis Outline

Chapter 1: Introduction serves as the entry point of this thesis and fulfills the following purposes: the motivation of the topic is shown and its relevance. It provides the main objectives of this thesis gives a brief description of the methodology developed and highlights the main contributions. Further, it describes the structure of the thesis. It shows the scientific publications that are a result of this thesis.

Chapter 2: State of the Art provides a state of the art survey on various related work in biometrics and gender prediction using eye tracking data.

Chapter 3: System Design gives an overview of the system design for both user identification and gender prediction. It explains the used datasets as well as the methods for data pre-processing and segmentation. Further, it introduces the feature extraction, Machine Learning (ML) classifiers and accuracy metrics.

Chapter 4: User Identification Experiments covers all the experiments of user identification that were conducted in this thesis. The user identification tasks were done with the Bioeye RAN, TEX, MIT, VST, and GOF datasets and the proposed approach was extensively analyzed with respect to various factors (e.g. age, gender) and extended by including higher order derivative features, blink information and tuning the important IVT parameters which brings a significant increase in the user identification accuracy.

Chapter 5: Gender Prediction Experiments covers all the experiments of gender prediction that were conducted in this study. The gender prediction task was done on the Dyslexia (this work is based on [Zaidawi et al., 2020]), VST, and GOF datasets. The gender prediction work with VST and GOF are based on [Haria et al., 2022].

Disclaimer on Content Reuse Single sentences as well as entire sections of this thesis are taken from my publications without explicit quotation. The same applies to figures and tables. However, the papers from which the content has been borrowed are clearly cited at the beginning of the chapters or sections. The text from this thesis may be reused in my upcoming publications.

1.5 Scientific Contributions

1. Schröder, C., **Al-Zaidawi, S. M. K.**, Prinzler, M. H. U., Maneth, S., and Zachmann, G. (2020). Robustness of eye movement biometrics against varying stimuli and varying trajectory length. In Bernhaupt, R., Mueller, F. F., Verweij, D., Andres, J., McGrenere, J., Cockburn, A., Avellino, I., Goguey, A., Bjøn, P.,

Zhao, S., Samson, B. P., and Kocielnik, R., editors, CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020, pages 1–7. ACM. (The first two authors equally contributed to this work)

2. **Al-Zaidawi, S. M. K.**, Prinzler, M. H. U., Schröder, C., Zachmann, G., and Maneth, S. (2020). Gender classification of prepubescent children via eye movements with reading stimuli. In Truong, K. P., Heylen, D., Czerwinski, M., Berthouze, N., Chetouani, M., and Nakano, M., editors, Companion Publication of the 2020 International Conference on Multimodal Interaction, ICMI Companion 2020, Virtual Event, The Netherlands, October, 2020, pages 1–6. ACM.

3. Prinzler, M. H. U., Schröder, C., **Al-Zaidawi, S. M. K.**, Zachmann, G., and Maneth, S. (2021). Visualizing prediction correctness of eye tracking classifiers. In Bulling, A., Huckauf, A., Gellersen, H., Weiskopf, D., Bace, M., Hirzle, T., Alt, F., Pfeiffer, T., Bednarik, R., Krejtz, K., Blascheck, T., Burch, M., Kiefer, P., Dodd, M. D., and Sharif, B., editors, 2021 Symposium on Eye Tracking Research and Applications, ETRA 2020, Virtual Event, Germany, May 25–27, 2021, Short Papers, pages 10:1–10:7. ACM.

4. **Al-Zaidawi, S. M. K.**, Prinzler, M. H. U., Lührs J., Maneth, S. An Extensive Study of User Identification via Eye Movements across Multiple Datasets (conditionally accepted in Image Communication).

5. Haria R. V. V., **Al-Zaidawi, S. M. K.**, Maneth, S. Eye movement analysis to predict gender using different sets of features. Eye Tracking Research and Application ETRA 2022 conference (submitted in December 2021) (The first two authors equally contributed to this work).

Chapter 2

State of the Art

The advent of eye tracking devices has spawned much new research. The recorded movements of the eyes can be analyzed and used for a variety of purposes and applications. For example, it has been used for disorder detection in medical research such as anxiety and depression [Armstrong and Olatunji, 2012, Alghowinem et al., 2013, Zhu et al., 2020, Shen et al., 2021], Autism Spectrum Disorders [Liu et al., 2015, Billeci et al., 2017, Eraslan et al., 2020], dyslexia [Rello and Ballesteros, 2015, Benfatto et al., 2016, Jothi Prabha and Bhargavi, 2019, Chakraborty and Sundaram, 2020, El Hmimdi et al., 2021, Raatikainen et al., 2021], and Alzheimer's Disease detection [Lagun et al., 2011, Fernández et al., 2013, Molitor et al., 2015, Nam et al., 2020, Readman et al., 2021]. Also, eye movement analysis has been used in gaming [Lin et al., 2004, Alkan and Cagiltay, 2007, Lankes and Stoeckl, 2020, Isokoski et al., 2009, Majaranta and Bulling, 2014], automotive research to study the visual attention of drivers [Singh et al., , Crundall and Underwood, 2011, Sun et al., 2016, Zandi et al., 2019], user identification [Kasprowski and Ober, 2004, George and Routray, 2016, Rigas and Komogortsev, 2017, Schröder et al., 2020, Seha et al., 2021], gender prediction using Machine Learning methods [Sargezeh et al., 2019, Moss et al., 2012, Büyük, 2021], study gender differences [Miyahira et al., 2000c, Huang and Chen, 2016, Sammaknejad et al., 2017], age prediction [Dalrymple et al., 2019], study age-related differences in people's visual exploratiVeness to certain visual stimuli [Kaspar and König, 2012, Horsley et al., 2013]. Recently eye tracking technology has been combined with Virtual Reality (VR) headset. This has led to the creation of many new opportunities in virtual marketing research and marketing [Hwang and Lee, 2018, Research, 2020, Pieters and Wedel, 2017]. Finally, we refer to [Kröger et al., 2019] to have an overview about what eye tracking data can reveal about the individual.

In this chapter, the state of the art of three of the above scopes are presented, i.e., biometrics, gender prediction, and dyslexia prediction as described in Section 2.1, Section 2.2, and Section 2.3 respectively

2.1 Biometrics

It has been observed that eye movements can be used as biometrics, i.e., as a way to identify a person within a larger pool of persons. The research on this topic started about 18 years ago with the seminal paper by Kasprowski and Ober [Kasprowski and Ober, 2004]. Since then, great improvements on the accuracy of eye movements biometrics have been achieved. These achievements have been driven by the continuous generation of high-quality datasets, and by the application of and experimentation with current methods from statistics and machine learning.

For better comparison of the various existing methods for eye movement biometrics, a competition series has been set up in 2012 [Kasprowski et al., 2012], where contestants use the same dataset to make predictions. To the best knowledge of the author, the best currently known method is the one from the winner of the most recent competition, BioEye 2015 [Rigas and Komogortsev, 2017], namely, the method of George and Routray [George and Routray, 2016]. In addition to the vast and fast growing literature on eye tracking biometrics, there are some surveys that provide a good overview e.g. [Esfahani, 2016, Rigas and Komogortsev, 2017, Galdi et al., 2016, Kröger et al., 2020, Katsini et al., 2020]. When comparing results, it is important to keep in mind that the fewer participants there are, the higher the prediction accuracies to be achieved; thus, it is hard to have a fair comparison of results that have different numbers of participants.

Let us start with the article [Kasprowski and Ober, 2004] which describes a novel approach for user identification based on eye movement characteristics. The technology assesses the human eye's response to visual stimulus which involves following a moving dot on a desktop screen. The eye movements of nine individuals was recorded for eight seconds in each trial (30 trials per participant). Their approach showed that it is feasible to identify individuals using the recorded eye movement by using OBER2 250 Hz eye tracking system. Their approach compiles behavioral and physiological aspects making it difficult to counterfeit, while it is simple to execute. Additionally, it may be combined with other methods dependent on cameras, such as iris or facial recognition. They used four different algorithms for users classifications: k-Nearest Neighbor, Support Vector Machines , Naïve Bayes, and Decision Tree. The best performing classifier was k-Nearest Neighbor with an

average False Acceptance Rate (FAR) of 1.48 %. False Acceptance Rate (FAR) is the percentage of identification instances in which unauthorized persons are incorrectly accepted. While FAR can not be directly converted into overall prediction accuracy, a low FAR is a good performance indicator for the classifier.

Another study by Holland and Komogortsev [Holland and Komogortsev, 2011] sought to evaluate the use of eye movements and their aggregated scan path characteristics in reading based biometric identification system. The eye movements of 32 participants (26 males and 6 femals) aged 18 – 40 were recorded using an EyeLink II eye tracker operating at a frequency of 1000 Hz. The study used both saccadic movements and fixations to identify different scanpath[1] features. The Equal Error Rate (EER) was used as a metric to identify each participant. EER is the percentage at which False Acceptance Rate (FAR) and False Rejection Rate (FRR) are equal. Similar to FAR, False Rejection Rate (FRR)is the percentage of identification instances in which authorized persons are incorrectly rejected. In general, the lower the equal error rate value, the higher the accuracy of the biometric system. Certain features were extracted from eye movement trajectories: mean fixation duration, mean saccadic amplitudes, mean saccade velocities, the velocity waveform (refer to the ratio of peak velocity to average velocity of a given saccade), scan path length (measured as the sum of absolute distances between the vectorial centroid of fixation points), scan path area, regions of interest, and fixation count. A statistical evaluation was implemented and the study identified the participants with an EER of 27 % based on the various scanpaths features. When the unique biometric features were utilized to identify unique individuals, the EER rates were between 30–49 %.

Further, a study by Rigas, et al. [Rigas et al., 2012] investigated eye movement characteristics like velocity and acceleration of saccades that occur in the eye movements. Eye movements of 37 participants were recorded using an OBER2 250 Hz device while following a jumping point that changes its position as stimulus. The first stage of their approach was pre-processing the data, then it was segmented into fixations which occurred while a participant observed the experimental stimulus. The next step was the isolation of signals related to fixations which was easy since the saccades produced easily detectable high magnitudes in eye movements' velocities. The results of this segmentation process was the representation of each input sample with a set of ten sub-signals, each one of them corresponding to a fixation that took place during an experimental session. In the sequence, the left

[1] A scanpath is the path joining the centroids of various fixations and should be differentiated from gaze trajectory.

and the right eye positions of each person was utilized in order to deduce the velocity and acceleration features for each of the sub-signals. After feature extraction, the comparison methodology relying on the multivariate Wald-Wolfowitz test (which is a non-parametric statistical test) was implemented. In order to perform the Wald-Wolfowitz test, each sub-signal was treated as a vector of the 8-d space (velocity left x, vel left y, accel left x, accel left y, vel right x, vel right y, accel right x, accel right y). After that, they compared the test sample with the training sample to evaluate the similarity of the feature distributions using a k-nearest neighbors (KNN) classifier. Finally, they achieved an identification accuracy of 91.5 % which largely exceed the reported results in previous attempts.

Then one of the next researches by Holland and Komogortsev [Holland and Komogortsev, 2012], sought to determine the effects of the type of stimuli and the specification of the eye tracking to the eye movement biometric identification system. This study segmented the eye movements by the velocity threshold algorithm (I-VT) into saccades and fixations. It also defined scan path as the spatial path formed by sequences of fixations and saccades. Several statistical features were extracted from the unique scanpath for each participant. This is a statistical analysis work whereby similarity scores are generated pairwise based on the individual features. Subsequently, infusion of the similarity scores is done based on the identified features and the combination used with biometric identification. The I-VT algorithm was utilized in association with a micro-saccade filter and a micro-fixation filter. Both groups were then merged to create a scanpath. The experiments of investigate the effect of stimulus type (four types were used, random, textual, simple cognitive task, complex cognitive task) using 22 participants (17 males and 5 females, aged 18–46) and another experiment to provide data recorded with low-cost eye-tracking equipment for cross-validation using 28 participants (18 males and 10 females, aged 18–36). The first experiment utilized a Tobii TX300 300 Hz eye tracking system and the other experiment used a modified PlayStation camera to track the movement of the eyes. Their results showed that utilizing textual and complex cognitive task stimuli were performing better than the other stimuli with the equal error rate not exceeding 31 %. Furthermore, they showed that user identification using a low temporal resolution eye tracker of 30 Hz is still achievable. The research is limited by the high equal error rates according to accepted biometric standards (EER should be less than 5%).

Another study to consider is by George and Routray [George and Routray, 2016] which achieved the best results in the biometrics competition that was conducted in 2015 (BioEye competition [Rigas and Komogortsev, 2017]). They utilized eye

movements of 153 participants (aged 18–46) who looked at two different stimuli: random moving dots (RAN) for 100 seconds and a poem (TEX) for 60 seconds which was recorded by an EyeLink-1000 Hz eye-tracker. Each stimulus has two recordings per participant, which were separated by a pause of 30 minutes. Additionally, there was an extra third recording which was recorded for 37 participants after one year gap which was used to study the effect of template aging (to check if the participants are having unique features which should not change over time). Their approach was first to pre-process the raw eye movements then segment them into fixations and saccades using the I-VT algorithm. The following step was to extract a large number of statistical features from both fixations and saccades. The feature number was minimized by using a backward feature selection algorithm to reduce the redundancy and correlation. After that these features were fed into a Radial Basis Function Network (RBFN) classifiers for both the fixation and the saccade features. It should be noted that the approach utilized by the researchers is superior to all other methods within the BioEye 2015 competition. The best achieved results were with using the RAN dataset with an accuracy of 89.54 % while using the TEX dataset produced an accuracy of 85.62 % over one run. While the accuracy reduced to 81.08 % when they test with the recording after one year gap which showed that this lower accuracy attributed to the template aging effects.

After that there were some biometrics surveys which present overviews of the biometrics works and implications. As a first example, the study by Esfahani [Esfahani, 2016] is a literature review of the previous studies where the use of eye movements for personal identification was studied. These studies can be divided into two groups: the first group utilizes eye movement bioelectrical signals for identification purposes and another one uses eye movement tracking in human identification. The article indicates few studies have evaluated the use of eye-tracking for identification. The approaches of these studies showed that there was discriminatory information in the eye movements that can make them useful as an alternative for conventional biometrics. However, they were using datasets with few participants and their achieved accuracies were rather low. In one of these studies reviewed, the researchers studied task independence (train and test with different tasks) in their approach to identify persons based on their eye movements. Their motivation was to prevent the participants to learn about the stimuli task which can change the participant's behavior and so this may increase the recognition errors. The authors utilized a reading task for training and watching a video as a stimulus for testing the eye-tracking task-independent scenario. The study focused on the fixations only, and the eye movement were described using a histogram indicating the angles moved by the eye during a specified period. The approach produced a time sequence of K

dimensional features that were subsequently modeled by Gaussian mixtures. The accuracy level realized was 30 % EER that fell short of the desirable 50 % EER, and these results were not sufficient for a realistic application.

Another literature review study by Rigas and Komogortsev [Rigas and Komogortsev, 2017] describes the different used approaches for analyzing eye movements, their features, and their results in the "BioEye 2015: Competition on Biometrics via Eye Movements." The research delved into different topics that include the previous studies on eye movement for biometric identification, the process of creating a high-quality database for eye movement biometrics that can be of utility to the scientific community. The study indicated that in creating a database, there are key considerations that must be prioritized. These include a variation of the stimuli utilized for the eye movement biometrics. Some of the suggested stimuli include text excerpts, video watching, jumping points, natural images, and face images. The study also indicated that the most utilized eye-tracking apparatus is the EyeLink 1000. Further, the number of participants in any research study for eye movement should be adequate to easily generalize the data. Further, the study indicated that raw data from eye movements should be processed. In most cases, data gets filtered through the decimation of the original frequency from 1000 Hz to 250 Hz through the anti-aliasing FIR filter. Other filters include Kalmar and the Savitzky-Golay. The data can also be classified into saccadic and fixation components using different approaches like velocity-threshold identification, Hidden Markov models, or Kalman filters. Other features of eye movements like amplitude, velocity, and acceleration profiles were extracted. The extracted characteristics can be utilized to develop the scan paths as formed by eye movements. They also mention the limitations, that in real life there are many factors that may affect the use of eye movements for identification purposes as opposed to the controlled factors in the experiments. As such, there is a need for testing the robustness of the identified algorithms under different capturing conditions. Further, there is a need to test for the specific variations as related to differences in gender and age. The study concluded that eye movements could augment other biometrics like the iris technology to support identification.

Another biometric study was conducted by [Li et al., 2018b], which presented a visual search task stimulus that designed to improve the performance of some eye movement biometric methods compared with the usual visual task. They recorded gaze trajectories of a total of 58 participants (24 males and 34 males) in experiments in which the participants had to choose the right answer in a series of questions (searching for numbers) using a Tobii TX300 eye tracking system running at 300 Hz.

Two trials with at least two weeks in between were recorded per participant. In each trial, there were 160 questions divided into four sessions separated by two minutes, each session consisting of 40 questions. Each participant has at least four minutes recording duration in each session. In their approach, they used four different eye movement feature extraction methods. They utilized a new type of feature i.e. texture features which are fixation density maps (divide the screen to a grid of pixels and calculate the fixation density in each box). These features were extracted by multi-channel Gabor wavelet transform (GWT) from eye movement trajectories. The highest accuracy in this work is 97.35 % (termed as rank-1 identification rate) using 10 runs (data was split into 70 % train and 30 % test) with their Fixation Density Map (FDM) method. The method for computing the fixation density map is specific to stimuli where the eye movements can be tracked on a specific grid of numbers. On the other hand, they achieved an accuracy of 80.05 % and an EER of 5.74 % when they use the complex eye movement pattern (the method of extracting eye movement features based on fixation and saccade for the biometrics identification). Their accuracies dropped significantly when they train with one trail and test with the other trail were found a time gap in between e.g. the above accuracy dropped from 80.05 % to 27.87 %.

Another recent interesting research is by Seha et al. [Seha et al., 2021] which was conducted for improving eye movement biometrics. The authors point out that static and dynamic features can improve eye movement by using data extracted from periocular shape patterns and eye-blinking patterns. The authors proposed a model that comprises four main steps, including acquisition and extraction of raw data, pre-processing, feature extraction, and class score fusion. The research used text and moving objects as stimuli. Two datasets were recorded for this study, first eye movements of 35 participants were recorded using the eye tracker Autocruis (AC) system at 30 Hz, and second dataset recorded eye movement of 22 participants using the Gazepoint three (GP3) eye tracker at 60 Hz. They integrated the eye movement features with the face recognition features e.g. shape of the eye, the distance between the eyes for their identification approach. The model uses GP3 and AC to analyze eye blinking using two approaches; light reflection and area within the eye. The study used the IV-T algorithm to segment into fixations and saccades. The researchers conducted the study within an authentication window of five minutes. The reported results with using a Linear Discriminant Analysis (LDA) classifier of using a different type of features were: 1) using fixation and saccade features to led an accuracy of 89.39 % over 3 runs with using 22 users (GP3 eye tracker) and 79.29 % over 3 runs with using 35 users (AC eye-tracker), 2) with using combined eye movement and periocular features, the model achieved an accuracy of 98.48 %

over 2 runs with using 22 users (GP3 eye tracker) and 94.57 % over 2 runs with using 35 users (AC eye tracker), 3) using blinking, the model achieved an accuracy of 93.94 % over 3 runs with using 22 users (GP3 eye tracker) and 74.6 % over 3 runs with using 35 users (AC eye tracker), 4) using combined eye movement and blinking features, the model achieved an accuracy of 96.67 % over 2 runs using 22 users (GP3 eye tracker) and 94.71 % over 2 runs using 35 users (AC eye-tracker). This study's main limitations were the limited number of participants and the other limitation was that the average blinking and saccade rate per minute in different databases were inconsistent. Other factors affected the rate of eye fixations, including fatigue, distraction, and sleepiness.

Other work by Makowski et al. [Makowski et al., 2021] developed a deep convolutional neural network to identify users based on their eye movements. The eye movement of 150 participants by using Eyelink Portable Duo eye-tracker at a sampling frequency of 1000 Hz. Each participant underwent four sessions to complete the experiments. They also used other available datasets with different stimuli: A random dot (RAN), a reading task (TEX), video viewing tasks (VD), a fixation task (FXS), a horizontal saccade task (HSS), and a gaze-driven game called Balura (BLG). The methods used in the experiment were subsumed into three major categories, including aggregational, statistical, and generative methods. The study also used the deep-eye-identification model to utilize the raw signal's micro-eye movements. The study aimed at assessing the presentation-attack problem to detect whether the gaze pattern was presented with an objective of interfering with the biometric system. The researcher concluded from their work that the micro scale eye movements contain many individual characteristic patterns that can be lost when only considering macro scale eye movements that take place less frequently. They suggested that even trackers with low precision and sampling rate can contain enough identifying information outside of saccade and fixation features to give their deep-eye-identification model an advantage over statistical methods that use selected features of fixations and saccades. Also, they conclude that unconstrained eye movements during reading are suitable better for biometric identification in comparison with watching jumping dots or videos, where the eye movements are more influenced by the stimulus; forced fixations lead to the least variability across persons and are least suitable for identification. Their best reported accuracy was 80 % with TEX stimuli.

Another study by Harezlak et al. [Harezlak et al., 2021] involved biometric identification methods based on eye movement signal. In this study, new features were examined based on nonlinear time series analysis (non-LTS). They included

its representation in the frequency domain and the largest Lyapunov exponent, which characterizes the dynamics of the eye movement signal seen as a nonlinear time series. Along with the previous conducted works velocity and acceleration features, these new features were determined. The eye movement data was recorded using the eye-tracker device (Ober Consulting Jazz-Novo 250 Hz), with a total of 24 participants that tracked a jumping point appearing on the screen in 29 different positions for 3 seconds in each position as a stimulus. Four different classifiers were investigated to check their performance such as the k-nearest neighbors algorithm, decision tree, naïve Bayes, and random forests. Good classification performance was obtained for decision trees and random forest. The efficiency of the last method reached 100 %. The outcomes were much worse in the second scenario when the training and testing set were based on recordings from different sessions; the random forest attained 72 % accuracy; and in the second section, the accuracy reached 79 %. The Decision Tree gave 74 % percent for correctly assigned vectors and dropped to 46 % when the extended feature vector was utilized.

A recent study by Porta et el. [Porta et al., 2021] conducted during the covid-19 outbreak focused on explaining the importance of contactless authentication methods. According to the authors, the best authentication methods are comprised of the eye or gaze features which has the advantage that they can also be used by people wearing mouth and nose masks, which would make traditional face recognition approaches difficult to apply. The researchers proposed a gaze-based experiment using soft biometrics. Their experiment was conducted on 44 participants, of whom 30 were male and 14 were female. The main stimuli in the experiment were animated texts and images. They also used four animations as visual stimuli (Random motion, Radial motion, Spiral motion, Volcano motion), with one or more moving objects (white squares) displayed on a black background. Initially different machine learning classifiers were considered, as well as different pre-processing techniques for transforming the dataset. The best results were obtained with Support Vector Machine, Random Forest, and Multi-Layer Perceptron. Pupil consideration, temporal, and spatial features were used for authentication. The extracted features were ranked by ANOVA F-value method and they considered an incremental number of features starting from 30, 40, 50, until 60. The study revealed that the random motion was the animation that achieved the highest identification result with an accuracy of 88.69 % using a Random Forest classifier.

Finally, an interesting research study by Lohr et al. [Lohr et al., 2021] presented a convolutional neural network (DeepEyedentificationLive) that can be utilized to identify users. In the study, the network gets trained using an established metric

learning loss function, a multi-similarity loss enabling the formation of a clustered-embedding space hence allowing identification of out-of-sample users. The research utilized data generated from 322 participants (151 females and 171 males) over a 37 months time period which became publicly available recently (for more details of collecting process of this data see [Griffith et al., 2020]). The dataset consists of seven diverse stimuli in two contiguous sessions during each round of recording (fixation task, horizontal saccade task, random oblique saccade task, reading task, two free viewing of cinematic video task, gaze-driven gaming task). All data was collected using an EyeLink 1000 eye tracker at 1000 Hz. The research involved an extensive analysis of the collected data of the participants' eye movements in the assigned tasks. Their study revealed that reasonable accuracy of authentication can be achieved during the low cognitive load task or at low sampling. The eye movement signals were downsampled with an anti-aliasing filter. The targeted sampling rates were 500, 250, 125, 50, and 31.25 Hz. In their approach they evaluated with using 14 participants that were presented in all the nine recorded rounds. Also, they found that eye movements are quite resilient against template aging after three years. The method used by the researchers allowed them to present end-to-end biometric authentication using eye movements. The statistical analysis conducted resulted in different equal error rate (EER) and false acceptance rate (FAR) values that did not meet the FIDO's alliance criteria [Schuckers et al., 2015] of 5 % FRR when FAR is fixed to 10^{-4}. The GazeBase dataset was useful due to its participant diversity in characteristics and the diverse tasks involved. Further, the duration of the study had been elongated to allow for a more robust comparison. Lastly, one of the limitations of the study was that the EER estimates for the study relied on only 14 participants who enrolled to the end of the recordings experiments of this study.

Overall, a total of eleven works were considered in the state of the art study on the topic of user identification based on machine learning methods. The key aspects of these works (number of participants, trajectory length, stimuli type, ML classifier, used metric, and the score) are summarized in Table 2.1.

Table 2.1: Overview of the basic information of Section 2.1 user identification work.

Original	Part. No.	Trajectory Length [s]	Stimuli type	Used features	ML classifier	The used metrics	The score
[Kasprowski and Ober, 2004]	9	8 in 30 trails	jumping dot	Fix.& Sac.	k-NN,	FAR	1.48 %
[Holland and Komogortsev, 2011]	32	60	reading text	Fix.& Sac.	statistical test	EER	27 %
[Rigas et al., 2012]	37	N.A.	jumping dot	Fix.& Sac.	KNN classifier	Acc.	91.5 %
[Holland and Komogortsev, 2012]	22&28	N.A.	4 stimuli	Fix.& Sac.	statistical test	EER	31 %
[George and Routray, 2016]	153	60	reading	Fix.& Sac.	RBFN	Acc.	85.62 %
	153	100	random dot	Fix.& Sac.	RBFN	Acc.	89.54 %
[Li et al., 2018b]	58	240	visual search	Fix.& Sac.	SVM	Acc.	97.35 %
[Seha et al., 2021]	22	300	text	Fix.& Sac.	LDA	Acc.	89.39 %
	35		&moving object	Fix.& Sac.	LDA	Acc.	79.29 %
	22	-	-	&periocular	LDA	Acc.	98.48 %
	35	-	-	&periocular	LDA	Acc.	94.57 %
	22	-	-	&blinking	LDA	Acc.	93.94 %
	35	-	-	&blinking	LDA	Acc.	74.6 %
[Makowski et al., 2021]	150	60	7 stimuli	Fix.& Sac.	CNN	Acc.	80.0 %
			the best TEX				
[Harezlak et al., 2021]	24	87	jumping point	non-LTS	RF	Acc.	100 %
[Porta et al., 2021]	44		several stimuli	Fix.& Sac.	RF	Acc.	88.69 %
[Lohr et al., 2021]	322		7 stimuli	Fix.& Sac.	deep NN	EER	5 %
	(14	for	evaluation)				

2.2 Gender Prediction

Humans of various genders have diverse cognitive processes that impact their visual perception [Lorigo et al., 2006, Pan et al., 2004, Rupp and Wallen, 2007, Sargezeh et al., 2019, Vanston and Strother, 2017]. Several studies have previously been conducted to study this link from a physiological and behavioral standpoint [Cárdenas et al., 2013, Emam and Youssef, 2012, Hall et al., 2010, Huang and Chen, 2016, Pérez-Moreno et al., 2016, Sammaknejad et al., 2017]. Only few research studies employed machine learning (ML) classifiers to predict gender using eye-tracking data. Both studies employ pictures as stimuli and discover consistent variations in the spatial and temporal properties of female and male eye movements. Their claimed accuracy rates are 64 % in [Mercer Moss et al., 2012] and 70 % in [Sargezeh et al., 2019]. Only adult individuals are considered in these arti-

cles. Miyahira et al. [Miyahira et al., 2000c] looked at three groups (prepubescent, postpubescent, and adult), and it was discovered that only the adult group had statistically significant gender differences in four aspects (mean gazing time, the total number of gaze points, mean and total scanning length). Another study for Miyahira et al. [Miyahira et al., 2000a] in which they analyzed eye movements involved 48 participants (24 male and 24 female) in order to find gender differences. Four parameters of eye movements were examined for this study, the mean fixation time, the total number of fixations points, the mean eye scanning length, and the total eye scanning length. Their results showed that the mean fixation time of females was significantly longer than that of males, and the total number of fixation points of females was significantly less than that of males. The mean eye scanning length of males and females did not differ.

Some of the researches that studied the gender differences are presented in the following. A study conducted in [Vassallo et al., 2009] analyzes eye movements of 50 participants (27 females and 23 males) recorded by Tobii 1750 binocular infrared eye tracker to find out whether there are differences in how females and males align their visual attention to salient facial features when they view static emotional facial expressions. Also, other pictures showed a sad, disgusted, happy, surprised, angry, or fearful expression as stimuli. The participants were presented these images of faces with a defined region of interest. Their statistical analysis results showed that even considering that males and females employ an akin pattern of scanning when perceiving other human faces, they still engage different eye movement patterns in fixations and focus on different features of the face being perceived. It was established that males fixate more on the central region of the human face consisting of the nose and cheeks while females focus more on the upper region of the face on the eyes and eye-brows. Also, it was observed that women were significantly faster than men in correctly identifying expressions. This observation is from the experiment where the participants were instructed to note their time taken to accurately label the emotional expression. When they had decided the emotion shown in the facial stimuli, they pressed a spacebar and read aloud the emotion from the labels presented on the monitor. Then the emotion was recorded by the examiner on a recording sheet.

Later some similar studies were conducted to identify “scanning strategies” that are different for men than for women on different regions of face images stimuli. For instance, the study of Coutrot et al. [Coutrot et al., 2016] in which a huge number of 405 participants (203 males, 202 females) were used to study gender differences while watching pictures of faces as stimuli. Another study by Sammaknejad et al. [Sammaknejad et al., 2017] used 33 participants (18 women and 15 men) to investigate the scan path differences for both males and females. The statistical

results in both studies showed that women are gazing longer in the eye region and in particular the left eye region.

The study in [Emam and Youssef, 2012] showed another behavior difference in the eye movements of males and females during a reading task. Most individuals think that slower reading improves comprehension. This study used empirical data to see whether men and females read differently. The primary hypothesis in this study was that there is no gender difference in reading speed. Ten students, five females and five males used ASL 6000 series head-mounted eye trackers to conduct this experiment. The findings of the investigation revealed that males read faster than females. Females spend more time fixating while reading than males.

The paper by Suroya and Al-Samarraie [Suroya and Al-Samarraie, 2016] conducted a qualitative assessment to investigate gender differences among seven dyslexic children (5 males and 2 females) in a visual prediction task using eye movement data. The eye tracker used in this research was Sensomotoric Instruments with the aim to study the children's visual behavior while predicting the next target in the visual task as stimuli. ANOVA was utilized to capture the gender differences using the following eye movements features related to fixations (fixation number, duration, and pupil diameter) and saccade duration. These features were chosen to recognize children's concentration during the visual prediction experiment. Their results showed that the dyslexic males gaze longer than the dyslexic females in contrary to the same behavior in healthy group.

The following presents papers that used machine learning methods to predict gender. The study of Moss et al. [Moss et al., 2012] employed eye movements of 52 participants (26 males and 26 females) aged 19–47. The gaze points of the participants were recorded using a Tobii 650 eye tracker at 50 Hz. Eighty static pictures from films were utilized as stimuli. The recording time was 400 seconds per individual. In their study, they have used fixation density maps as features to train a Naive Bayes (NB) classifier. Their achieved accuracy was 64 % over ten runs. They also conducted a statistical analysis using the t-test method to find significant differences in some features such as fixation duration, saccadic amplitude scan path. Their statistical results showed that females were fixating longer than males, while males' saccadic amplitudes were significantly smaller than those of females'. Furthermore, they found that females tend to be more exploratory by fixating more on non-face regions. These results showed that the males' eye movements were four percent shorter but four percent more frequent than females, and based on the entropy based measures they found the female fixation distribution to be more

spread out and longer. This exploratory behavior creates more distinctive female fixation maps, which can explain why females were categorized more correctly than males in their study.

Sargezeh et al. [Sargezeh et al., 2019] carried out different experiments on the eye movements of 45 participants (25 males and 20 females) aged 25–34. The data was recorded using an Eyelink 1000 eye tracker 1000 Hz. The stimuli were sixty indoor pictures containing a large number of objects. The total recording time per participant was 240 seconds. There features were extracted from fixations and saccades: fixation duration, the ratio of total fixation duration to total saccade duration (RFDSD), Spatial density, the number of saccades, saccadic amplitude, and scan path length. The best achieved accuracy was 70 % over five runs using a Support Vector Machine classifier and a train to test ratio of 80:20. In this study gender differences in eye movement pattern were also investigated statically. Their statistical analysis showed that the mentioned six features reveal significant differences between males and females. They observed that females were more explorative than males having larger saccadic amplitudes and longer scan paths.

Finally, two other recent works [Büyük, 2021] and [Mohammad, 2021] which are theses at Utrecht University are considered. Both studies used a huge number of participants consisting 1242 visitors (549 males and 614 females) with aged 10–60 who were visitors to the Nemo science museum in Amsterdam. The eye movements of the visitors were recorded by using Tobii 4C eye tracker at a sample rate of 60 Hz. The stimuli task included free viewing of one image for only ten seconds. In total, there were 600 gaze points per individual. In [Büyük, 2021], the author extracted 24 statistical features from fixations and saccades for e.g. mean and SD, the number of fixations and saccades, and the age category. The segmentation algorithm used in this work to segment the trajectories into fixations and saccades was the two-means clustering (I2MC) [Hessels et al., 2017]. This segmentation algorithm is robust against high noise and it is suitable for eye movements data recorded with remote or tower mounted eye trackers using static stimuli. The best achieved accuracy in this work was 70 % using a Support Vector Machine classifier. However, explanation of some information in this work approach is missing for e.g. how they selected the features, how they solved the over-fitting issue, and what was the train to test ratio in their experiments. Furthermore, they reported some limitations in this work due to the data recording, since the data was recorded in one of the exhibition halls of a museum where the eye tracker was replaced without supervision, which might lead to duplicate data by some visitors. In [Mohammad, 2021], the author used the same dataset to predict the gender using deep learning methods and a classical machine

learning algorithm namely random forest. The participants of age of 10, 20, and 60 were removed hence, the remaining participants were 1163 (female: 614, male: 549). First, for the classical machine learning approach, the I2MC algorithm was used to segment the gaze data into fixations and saccades, then several statistical features were extracted from them. The result of the predicted gender of the participants included a mean absolute error of 47 % and a root mean squared error of 68 %. Since the features are highly algorithm dependent in the classical machine learning methods, the author tried the unsupervised deep learning algorithms to predict gender in this work. Therefore, the important features of the eye movement were identified by the deep learning method from the raw data and there was no need for segmentation of fixations and saccades. Hence, the results of the analysis are not dependent on the algorithm used for gaze event classification. However, this also has the drawback that the psychological and physical meaning of the features learned by the algorithm network is not necessarily well understood and hereby limits the theoretical meaningfulness of the findings. The results showed that the gender prediction task did not have a proper prediction by using velocity profiles as the input to the neural networks. It had 51 % mean absolute error and 53 % root mean squared error for predicting the gender. While using the gaze events segmented by the I2MC algorithm, i.e., a number of fixations and fixation duration resulted in a mean absolute error of 12 % and a root mean squared error of 34 % for the gender prediction. Finally, the results of data analysis in this study revealed a negative correlation between age and number of fixations and a positive correlation between age and fixation duration.

Overall, a total of four works were considered in the state of the art study on the topic of gender prediction based on machine learning methods. The key aspects of these works (number of participants, trajectory length, stimuli type, ML classifier, used metric, and the score) are summarized in Table 2.2. It can be observed that few studies were found for gender prediction based on machine learning methods. Also, these studies considered one dataset with only one stimuli task which is viewing images that can question the generality of these approaches.

Table 2.2: Overview of the basic information of ML studies in Section 2.2 of gender prediction work.

Original	Part. No.	Trajectory Length [s]	Stimuli type	Used features	ML classifier	The used metrics	The score
[Moss et al., 2012]	52	400	images	Fix.& Sac.	NB	Accuracy	64 %
[Sargezeh et al., 2019]	45	240	images	6 Fix.& Sac.	SVM	Accuracy	70 %
[Büyük, 2021]	1242	10	images	24 Fix.& Sac.	SVM	Accuracy	70 %
[Mohammad, 2021]	1163	10	images	Fix.& Sac.	deep NN	mean absolute error	12 %
[Mohammad, 2021]	1163	10	images	from raw data	deep NN	mean absolute error	51 %

2.3 Dyslexia Prediction

While investigating some aspects of the users from their eye movements such as biometrics and gender, several factors can lead to abnormal behavior of the eye movements which should be taken into the consideration such as the existence of some eye vision disorders like dyslexia. Dyslexia is a disorder identified by the difficulty of learning reading, spelling, writing, pronouncing. Commonly dyslexia affects about ten percent of children in the population. A review study by G. Vanitha et al. [Vanitha and Kasthuri, 2021] listed the existing dyslexia prediction works in the literature which revealed that the existing earlier machine learning studies are rather few and used a limited set of data applied to reading concerning only fixations and saccades. In this thesis, some work was done to predict dyslexia hence here is a short overview of some of the related studies in the literature that investigated existing dyslexia in eye movements.

The study of Rello et al. [Rello and Ballesteros, 2015] proposed the first machine learning model to automatically predict readers with dyslexia. The study conducted an experimental investigation on eye movements of 97 participants (48 with dyslexia, 49 without dyslexia) recorded with the eye tracker Tobii 1750. The stimuli included reading 12 different texts with different font types. They employed a supervised learning method for data analysis and pattern recognition namely polynomial Support Vector Machine (SVM). The experiments were conducted by using different common eye movements features such as the total time of reading, the mean of the fixation, the number of the fixations, and the age of the participants. The achieved accuracy to predict dyslexia was 80.18 % averaged over ten folds with a train to test ratio of 90:10. Another experiment was conducted after removing the age from the features list and the achieved accuracy was 76.38 %. They reported in

their work that eye movement can be used in the future to detect dyslexia and the prediction accuracy can be improved by using better datasets. Finally, their results showed that age is an important factor that can affect reading performance.

Another study by Benfatto et al. [Benfatto et al., 2016] proposed an experimental work to classify dyslexia in children aged 9–10 years old using eye tracking data. Their study used the eye movement of 185 participants (97 with dyslexia, 88 without dyslexia) that was recorded by a goggle-based infrared corneal reflection system at 100 Hz. The used stimuli were reading a short passage of a text appropriate for to the age of the participants. In this approach, the authors segmented the trajectories first into fixations and saccades by using a dynamic dispersion threshold algorithm. f Additionally, saccades were divided into further movements i.e. progressive (left to right) and regressive (right to left) movements, and the fixations were specified accordingly, depending on the direction of the preceding saccade. The next step was extracting several statistical features based on the segmented eye movements for a total of 168 features. These features were reduced to 48 features by using a recursive feature elimination method [Guyon et al., 2002]. The achieved dyslexia prediction accuracy was 95.6 % averaged over ten runs with a train to test ratio of 90:10 and using the machine learning classifier SVM.

The same dataset was used in the work of Benfatto et al. [Benfatto et al., 2016], was used again in another study [Jothi Prabha and Bhargavi, 2019] where they developed a statistical model to predict dyslexia. The main steps of this approach were first to derive fixations and saccades (progressive and regressive saccades) from the eye movement data by using statistical measures. A total of 75 features were extracted from these eye movements events such as the duration, the mean of the position of the eye during the experiment, the standard deviation of the mean of the eye position, and the distance between two positions of the eye. These features were selected by using Principal Component Analysis (PCA). In this work, the authors proposed to use an optimized kernel function that can be used in SVM for the classification, which is done by the Particle Swarm Optimization algorithm (PSO). They then compare the results with SVM performance. The PSO algorithm is used for tuning the weights for the features and to generate an optimized kernel which is a combination of linear and quadratic kernels. The proposed classifier Hybrid Kernel SVM-PSO is the combination of different kernels and results in better accuracies than a single kernel after considering linear and quadratic kernel combinations. The best achieved accuracy was 95 % using SVM-PSO, while using the traditional SVM classifier gave an accuracy of 90 %. These results were averaged over ten runs with a train to test ratio of 80:20.

The recent research conducted by Appadurai and Bhargavi [Appadurai and Bhargavi, 2021], is another study that used the same dataset as in [Benfatto et al., 2016] and [Jothi Prabha and Bhargavi, 2019]. Also, they tried to improve the approaches used in these two studies by investigating different feature sets and using a Deep Convolution Neural Network (CNN) in addition to an SVM classifier. The approach that was used in this work can be summarized as follows: first, pre-processing the raw data of all the 185 participants and removing the blink information, then segmenting the eye movement trajectory into fixations and saccades by using two different algorithms dispersion based and using the IVT algorithm. The next step was extracting several features. To eliminate the features to achieve the least correlated features for the dyslexia prediction task, two feature sets were selected by implementing two methods. The first feature set consisted of a total of 52 features that were selected by implementing the Principal Component Analysis (PCA) method on the features of fixations and saccades segmented by the dispersion based algorithm. The second feature set consisted of a total of 40 features that were selected by implementing the Recursive Feature Elimination with Cross-Validation (RFE-CV) method on the features of fixations and saccades segmented by the IVT algorithm. These eye movement features were used in different machine learning algorithms e.g. SVM, Boosting Decision Tree, Random forest, and CNNs. The study indicated that the accuracy of the different features based on three major approaches of velocity threshold based, statistical, and dispersion-based approaches and the noted accuracies are 96 %, 92 %, and 94 %, respectively (see Table 2.4). It was noted that velocity based threshold features gave the best accuracy. The eye movement feature sets that gave the best accuracy include the total number of fixations, mean fixation saccade duration, the total number of saccades, the ratio of saccades and fixations, and the first fixation start time. The features gave an accuracy of 96 % for the Hybrid Kernel SVM-PSO model, while the XGBoost classifier achieved 95 % accuracy. Finally, the CNN gave an accuracy of 88 %.

A further study of detecting dyslexia via eye movements was conducted by Raatikainen et al. [Raatikainen et al., 2021]. The dataset utilized in the study was generated from the project eSeek that was conducted by the University of Jyväskylä. This project was done to study the reading skills of Finnish students with and without reading challenges. The eye movements of a total of 165 participants with an average age of 12.5 years was recorded with an EyeLink 1000 eye tracker at 1000 Hz. The stimuli were search tasks that included reading contextualized question and then choosing the answer out of four options that would help the participants to answer the question. Different sets of features were investigated in this work,

being extracted from fixations and saccades. These features include the mean of the fixation duration, saccadic amplitude within the areas of interest, and fixation count. The important features were selected by using the random forest features selection method, and the following incremental number of features were used in this work experiments (10, 20, 30, 35, and 40 features). The best accuracy was achieved by using the set of the top 30 features and the SVM classifier (89.7 % averaged over 100 runs). Furthermore, several different approaches for generating feature sets were tested such as the transition matrix average (TMA) and matrix with histograms (TMH) of the fixations. The best accuracy of 86.7 % averaged over 100 runs was achieved by using the TMH features set (in a total of 760 features) which were used to train a random forest classifier. The research concludes that a deeper analysis of the significance of the eye movement features is needed to give a more reliable and fast screening tool for dyslexia.

Finally, the study of El Hmimdi et al. [El Hmimdi et al., 2021] utilized machine learning to predict reading speed and dyslexia among 87 adolescents. The eye movement data of 46 dyslexic adolescents, consisting of 18 males and 28 females with an average age of 15.5 years (and a standard deviation (SD) of 2.45 years) and the eye movement data of 41 non-dyslexic adolescents youths comprising of 20 females and 21 males with an average age of 14.8 (and an SD of 2.4 years) were used in this study. Different stimuli were utilized in this work i.e. reading and non-reading tasks (saccade and vergence movements) by using a head-mounted video oculography eye tracker at 200 Hz. The reading stimuli were divided further into types of text, first, the researchers provided an Alouette text to the participants which is a text that has no meaning that is used usually to assess the reading capacity in dyslexia, and secondly a normal text which has meaning. The participants were asked to read aloud while their eye movements were recorded. The researchers then selected the ten least correlated features for dyslexia prediction. Different machine learning classifiers were used in this study, including multiple linear classification models such as logistic regression (logReg), linear Support Vector Machines (SVM), and Naive Bayes and a nonlinear SVM classifier. Furthermore, some tree-based classifiers were used as well such as random forest and decision trees. Their results showed that the best linear classifier was logReg (accuracy of 81.25 %) and the best non-linear classifier was SVM (accuracy of 80.0 %), both using the Alouette text as stimuli. The best achieved accuracies with the best mentioned two classifiers and the four stimuli are shown in Table 2.3. These results were achieved using cross validation over 5 runs.

Overall, a total of six works were considered in the state of the art study on the topic of dyslexia prediction via eye movements based on machine learning method. The key aspects of these works (number of participants, trajectory length, stimuli type, ML classifier, used metric, and the score) are summarized in Table 2.4.

Table 2.3: The accuracy of Logistic regression (logReg) and Support Vector Machine (SVM).

Classifier	Reading Alouette text	Reading Meaningful text	Saccade stimuli	Vergence stimuli
logReg	81.25 %	70.2 %	81.25 %	77.3 %
SVM	80.0 %	71.57 %	80.0 %	68.42 %

Table 2.4: Overview of the basic information of Section 2.3 dyslexia prediction related work.

Original	Part. No.	Trajectory Length [s]	Stimuli type	Used features	ML classifier	Accuracy
[Rello and Ballesteros, 2015]	97	N.A.	reading	Fix.& Sac.	SVM	80.18 %
[Benfatto et al., 2016]	185	20	reading	45 Fix.& Sac.	SVM	95.6 %
[Jothi Prabha and Bhargavi, 2019]	185	20	reading	75 Fix.& Sac.	SVM-PSO	95.0 %
[Raatikainen, 2019]	165	N.A.	search tasks	Fix.& Sac.&AOI	SVM	89.7 %
				TMH	RF	86.7 %
[El Hmimdi et al., 2021]	87	N.A.	reading	10 Fix.& Sac.	logReg	81.25 %
[Appadurai and Bhargavi, 2021]	185	20	reading	52 Fix.& Sac.	SVM-PSO	94.0 %
				40 Fix.& Sac.	SVM-PSO	96.0 %
				40 Fix.& Sac.	CNN	88.0 %

Chapter 3

System Design

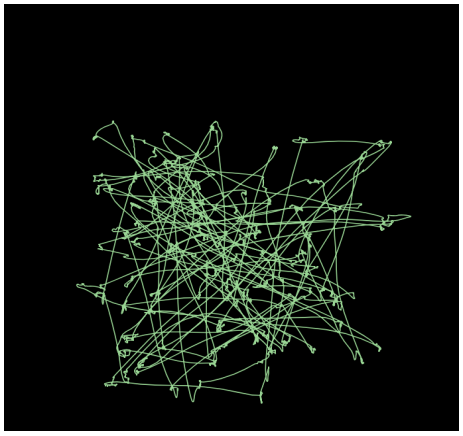
This chapter gives an overview of the system design for both user identification and gender prediction. It explains the used datasets as well as the methods for data preprocessing and segmentation. Further, we introduce the feature extraction, Machine Learning (ML) classifiers and accuracy metrics.

3.1 Datasets

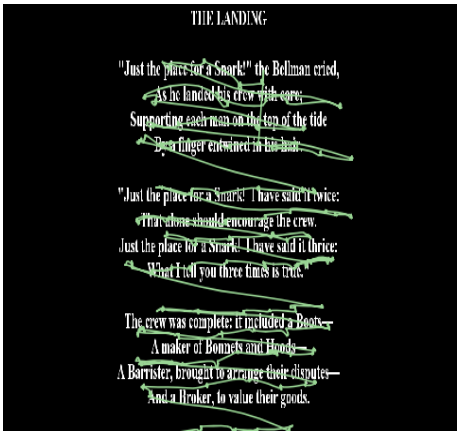
Six datasets with different stimuli were used: Bioeye TEX, Bioeye RAN, Visual Search Task (VST), Gaze on Faces (GOF), MIT, and Dyslexia. Some of these datasets were used for user identification (RAN, TEX, VST, and GOF), and some were used for gender prediction task (Dyslexia, VST, and GOF). In addition to gender a dyslexia prediction task was performed with the Dyslexia dataset. See Table 3.1 for an overview of the datsets. The following subsections describe these datasets in detail.

Table 3.1: Overview of Datasets.

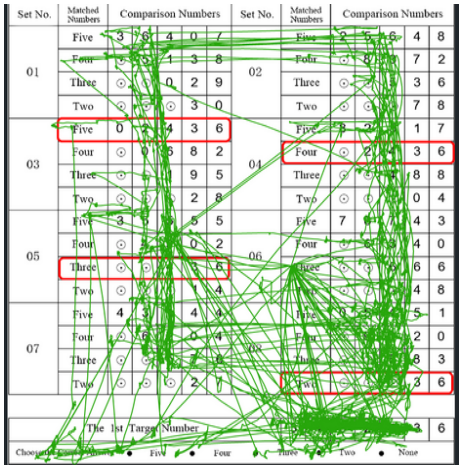
Dataset	Participants			Age	Trajectory	Blink	Prediction
	M	F	T	Range	Length [s]	Info.	Task
TEX	N.A.	N.A.	153	18–46	60	Yes	User
RAN	N.A.	N.A.	153	18–46	100	Yes	User
VST	24	34	58	21–33	240	Yes	User and Gender
GOF	193	185	378	20–72	60	No	User and Gender
MIT	N.A.	N.A.	15	18–35	3×1003 = 3009	No	User
Dyslexia	145	40	185	9–10	20	No	Gender and Dyslexia



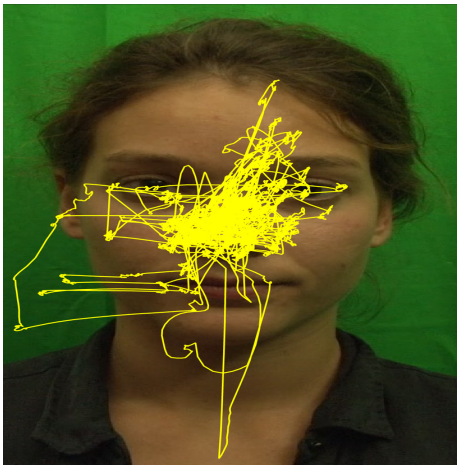
(a) RAN



(b) TEX

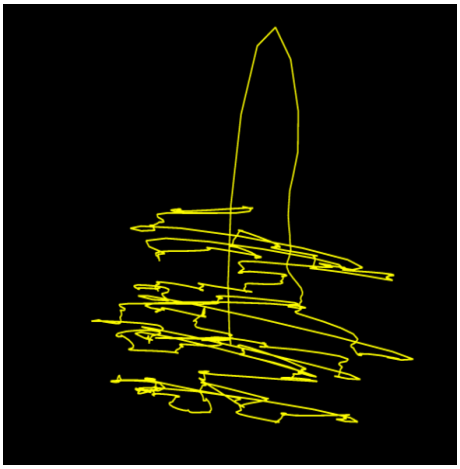


(c) VST



(d) GOF

(e) MIT



(f) Dyslexia

Figure 3.1: Various stimuli and their gaze trajectories. (a) a participant looking at random dots (RAN data), (b) a participant reading a text (TEX data), (c) a participant searching for number (VST data), (d) a participant gazing on an image of face (GOF data), (e) a participant gazing on an outdoor image (MIT data), (f) a participant reading (dyslexia data)

3.1.1 Bioeye (TEX/RAN) Data

Bioeye TEX and RAN are two datasets with different stimuli which were used in the *BioEye 2015 competition* [Rigas and Komogortsev, 2017] [1]. Each dataset has two recordings per participant, which were separated by a pause of 30 minutes in between. Both datasets consist of an eye movement data from a total of 153 participants including males and females aged 18–46. Additionally, there is one more session recorded after one year containing 37 participants.

TEX 60 second recordings of reading a poem (reading task). Figure 3.1a shows the gaze trajectory of a sample participant from this dataset along with the reading stimulus.

RAN 100 second recordings of observing a randomly moving dot (looking at a white dot moving in a dark background). Figure 3.1b visualizes the gaze trajectory of a sample participant.

Each participant was comfortably seated in front of monitor screen at a distance of $550\,\text{mm}$. The dimensions and resolution of the monitor were $474 \times 297\,\text{mm}$ and 1680×1050 pixels respectively. The participants' heads were supported with a chin rest to ensure stability during the sessions. The device used for recording was an EyeLink-1000 eye-tracker (1000 Hz) and the data was down-sampled to 250 Hz using an anti-aliasing FIR filter [Rigas and Komogortsev, 2017]. Despite this interpolation, the dataset still provides explicit information about the validity of the samples in the recording which can be attributed to device specific faults or user specific reasons (e.g. blinking, loss of attention etc.). During experiments, the device operated in a monocular mode and captured the movements from the left eye. The provided raw data is in the form of visual angles in x and y directions. These visual angles were calculated based on the user's eye angular rotation with respect to the stimuli. This dataset was used in the user identification experiments.

3.1.2 Visual Search Task (VST) Data

This dataset[2] [Li et al., 2018a] includes the recordings of gaze trajectories of a total of 58 participants (24 males, 34 females) aged 21–33. The participants performed a visual search task which consists of a series of number search questions carried out with pictures. The participants were asked to compare the target number with comparison numbers in a form to find the length of the longest matched number. Figure 3.1c shows an example of this stimulus and the gaze trajectory of one of the participants. There are four Comparison Numbers ("02436", "2436", "436", and "36")

[1] The data was provided by Oleg Komogortsev.
[2] The VST dataset was provided by Chunyong Li

which match the Target Number ("3402436"). The correct answer is "Five", which is the length of "02436". This visual search task was designed to improve the performance of some eye movement biometric methods compared with ordinary visual tasks. The recording duration was at least 4 minutes for each participant in each session. The data collection experiment was divided into two trials with at least two weeks in-between. In each trial, there were 160 questions divided into four sessions separated by two minutes, each session consisting of 40 questions. The participants took two minutes rest between these tests. Overall, $160 \times 2 = 320$ gaze trajectories for each participant were collected (18560 in total). The participants were seated at a distance of 600 mm from the monitor screen. The dimensions and resolution of the monitor were 474×297 mm and 1920×1080 pixels respectively. The device used for recording was a Tobii TX300 eye tracking system running at 300 Hz. In contrast to the Bioeye dataset, no explicit information about the validity of the gaze trajectory was provided. Nevertheless, the gaze trajectory contained some NaNs which we consider as invalid and hence a source for blink information. We interpolated across the invalid segments in order to have a connected gaze trajectory by filling the NaNs with the mean of the previous and next valid gaze points. This data has long trajectory lengths and different sessions that were recorded on the same day and others that were recorded after a two week gap. Therefore, this gives us the opportunity to study the influence trajectory length, blink (fatigue), and the gap between train and test sessions. This dataset was used in the pipeline for both user and gender prediction. The raw data is in the form of gaze points (X(t), Y(t)) in pixels relative to the screen for each timestamp "t". For the experiments, the trajectory was further divided into segments, e.g., 12 seconds, 20 seconds, 40 seconds, 1 minute, 2 minute and are used to study the effect of length of gaze trajectories on the user identification or gender prediction. These experiments are performed by taking gaze trajectory segments from beginning to end and from end to beginning.

3.1.3 Gaze on Faces (GOF) Data

This dataset[3] is provided in the study of [Coutrot et al., 2016] and involves the gaze trajectories of participants while observing faces. Overall, there were eight actors whose faces were used as stimuli consisting of an equal number of males and females. During the start and end of the video clips, the actor gazed towards the bottom of the screen for 500 ms. The data was recorded in 2016 at the Science Museum of London, UK. The dataset consists of a total number of 405 candidates from 58 countries, aged 18–72. However, participants with erratic, absent, or damaged data were eliminated

[3]GOF dataset: https://uncloud.univ-nantes.fr/index.php/s/8KW6dEdyBJqxpmo, Accessed: 2020-08-21

to avoid inaccuracies from the study. Hence, the research in this thesis was carried out using the remaining 378 participants of which 193 were males and 185 were females, aged 20–72. The participants were seated at a distance of 570 mm from an LCD monitor screen (1280×1024 pixels). The width and height of the stimuli were 429×720 pixels. Most parts of the fixed images were covered by the face measuring 280×420 pixels. The pixel size of the eye, nose, and mouth was 75×30, 80×90, 115×35 respectively.

Eye-tracking data was collected using the EyeLink 1000 kit eye-tracker at 250 Hz. Figure 3.1d shows an example of the gaze trajectory path which followed by the user with ID 167 can be seen when viewing at the face of the actress. Originally, there were 35 trials of which 3 were eliminated as they did not possess records of all the trials. Hence the used data contains a minimum of 32 trials that were adopted for this study. The participants looked at multiple images of a single actor gazing towards them for varying durations ($0.1\,s$ to $10.3\,s$) in all the 32 different trials. This dataset is used for both gender prediction and user identification tasks. The raw data is in the form of gazepoints (X(t), Y(t)) in pixels.

The selected dataset is utilized as it offers an opportunity to study a set of participants that contains a large number of male and female participants with a broad age range. In this dataset, no invalid parts in the gaze trajectories were found and hence no blink information could be deduced. The demographics of the GOF dataset

Table 3.2: Demographics of Males and Females in GOF dataset

Age Group	Males	Females	Total
20–72	193	185	378
20–30	116	132	248
31–50	59	42	101
51–72	18	11	29
20–40	151	157	308
41–72	42	28	70

is summarized in Table 3.2. For this study, 378 participants are selected when using the whole age group (20–72), 248 participants when using age group (20–30), 101 participants when using age group (31–50), 29 participants when using age group (51–72). The total number of participants comprised of an equal number of males and females. The participants were divided into three different age groups to evaluate the effect of age on gender prediction. Next, the three age groups are merged and reorganized into two age groups with broader age gaps to assess the effect of age on gender prediction and user identification.

3.1.4 MIT Data

The publicly available MIT dataset [Judd et al., 2009][4] provides three seconds long eye movement trajectories (per image) from 15 participants looking at 1003 images. The participants in this category were males and females aged 18–35. These participants freely viewed the images on a 19-inch computer screen (with resolution 1280×1024) in a dark room and a chin rest, situated at a distance of approximately 609 mm from the screen, to stabilize their head during the data recording and eye tracking. An ETL 400 ISCAN eye tracker device recorded their gaze paths (with 240Hz) on a separate computer as they viewed each image at full resolution for three seconds separated by one second of viewing a gray screen.

The original objective of the MIT dataset was the study of saliency models, i.e., models that allow predicting points of (visual) attention [Judd et al., 2009]. We used this dataset for eye movement biometrics, as a particular instance of *stimulus-independent* eye movement biometrics. The use of different images within biometrics as a way of stimulus independence was already mentioned and experimented with (using only two different images) by Darwish [Darwish, 2013]. Also, the MIT dataset is used to study the influence of the amount of train and test data on the accuracy of user prediction. In order to use the MIT dataset for our task, we split the recordings of each participant into train and test sets of certain sizes as will be explained later in Chapter 4. Furthermore, most of the studies on the application of eye movements data for biometric identification used the datasets involving reading tasks and looking at a moving object (e.g. the RIGAS dataset [Galdi et al., 2016]). Hence, we chose the MIT dataset in our study because this kind of dataset has not yet been used for eye movement biometrics. Finally, the provided raw data is in the form of visual angles in x and y directions.

3.1.5 Dyslexia Data

The dataset provided in the study of [Benfatto et al., 2016] was used. This dataset comprises children aged 9–10 years who participated in the Kronoberg reading development project, which is a longitudinal research project on reading development and reading disability on Swedish school children that ran between 1989 to 2010. The dataset includes gaze points from both eyes recorded with a goggle-based infrared corneal reflection system at 100 Hz for 185 participants consisting of 97 dyslexic (21 females (DF) and 76 males (DM)) and 88 non-dyslexic individuals (19 females (NDF) and 69 males (NDM)). The demographics of the Dyslexia dataset can be seen in Figure 3.2. The data was previously used to predict dyslexia in children. The distance

[4]MIT dataset: `http://saliency.mit.edu/datasets.html`, accessed 01.07.2018

of the screen from the seated participants was 450 mm away from an LCD monitor screen 1680×1050 pixels. The stimuli's height and width were 474×297 pixels. All participants viewed the same text, which consists of 10 sentences (8 lines) with a mean length of 4.6 words. The recording time per participant did not exceed 20 seconds. The dataset has a very unbalanced gender representation (40 females and 145 males) but to the best of our knowledge it is the only openly available eye tracking dataset for young children. Nevertheless, the dataset still provides a decent number of participants to study gender prediction in balanced groups by limiting the number of males. This way 38 non-dyslexic (19 NDF + 19 NDM) and 42 dyslexic (21 DF + 21 DM) participants were available. Figure 3.1f shows an example of this data trajectory. Lastly, the provided raw data is in the form of visual angles in x and y directions.

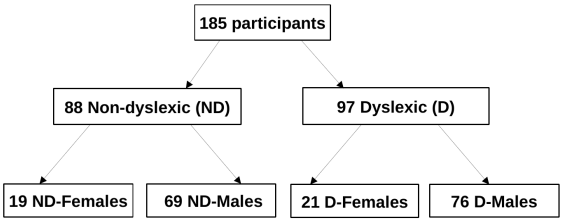


Figure 3.2: Demographics of Males and Females in Dyslexia dataset

3.2 Preprocessing and Segmentation

The raw eye tracking data usually provided either the viewing angles or pixel coordinates in x and y directions. Normally, most of the eye tracking features are computed with pixel coordinates expressed in the screen coordinate system. When the dataset provides the angular gaze points, the pre-processing stage involves converting the obtained data to screen coordinates based on head distance and geometry of the acquisition system (see Figure 3.3). While the VST and GOF datasets provide the recorded pixel coordinates directly, the RAN, TEX, MIT, and Dyslexia datasets provide viewing angles w.r.t. the x and y axes. They have been converted to screen pixel coordinates (x_{screen}, y_{screen}) as follows:

$$\begin{aligned} x_{screen} &= \left(\frac{d \cdot w_{pix}}{w}\right) \tan\theta_x + \frac{w_{pix}}{2} \\ y_{screen} &= \left(\frac{d \cdot h_{pix}}{h}\right) \tan\theta_y + \frac{h_{pix}}{2} \end{aligned} \tag{3.1}$$

The distance from the screen and viewing angles in the x and y directions are denoted by d, θ_x, and θ_y respectively. The variables w_{pix} and h_{pix} indicate the screen's

resolution and w and h the physical size (width and height). The inverse of the above equation is used to compute the viewing angle coordinates from the pixel coordinates for the RAN, TEX, MIT, and Dyslexia datasets (see Equation 3.2).

$$\begin{aligned} \theta_x &= \arctan\left(\left(\frac{x_{screen} \cdot w}{d \cdot w_{pix}}\right) - \left(\frac{w}{2 \cdot d}\right)\right) \\ \theta_y &= \arctan\left(\left(\frac{y_{screen} \cdot h}{d \cdot h_{pix}}\right) - \left(\frac{h}{2 \cdot d}\right)\right) \end{aligned} \tag{3.2}$$

This is necessary as our implementation of the IVT algorithm for segmentation in fixation and saccades works in angular coordinates.

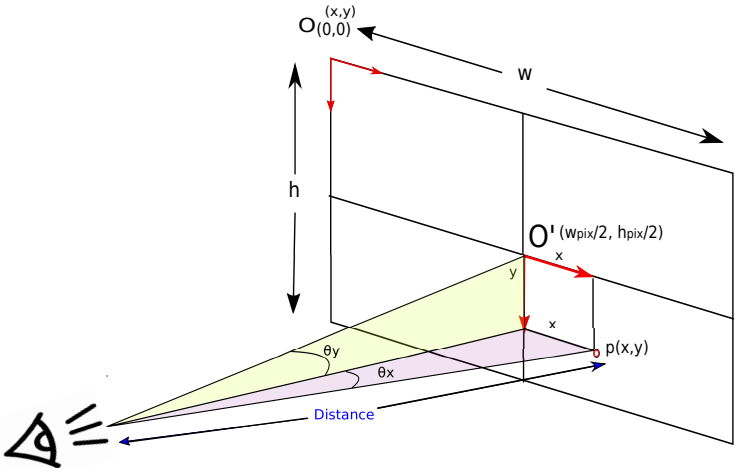


Figure 3.3: Relationship between visual angles and pixel coordinates

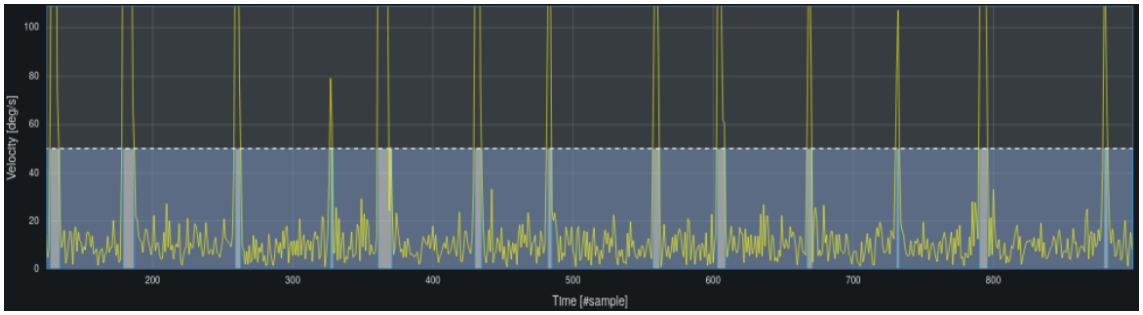
Noise may be present in the raw gaze trajectory data which amplifies further in the calculation of velocity, acceleration and other higher order time derivatives of the gaze trajectory. This noise is due to the high-frequency components in the eye movement signals, especially during saccades.

Duchowski et al. [Duchowski et al., 2016] reported in their work that the Savitzky-Golay filter is appropriate for implementing velocity-based saccade detection. It is suggesting the suitability of the Savitzky-Golay filter for the Identification Velocity Threshold algorithm (IVT) implementation of the event detection algorithm, which will be explained in the next paragraph. In this thesis, a Savitzky-Golay [Savitzky and Golay, 1964, Schafer, 2011b] filter has been implemented to reduce the influence of noise (see also [Schafer, 2011a]) and to increase the precision of the data without distortion (later in Section 4.2, it is shown that the difference in accuracy of user identification with and without filter can be as high as 4%). For every data point, it fits a symmetric polynomial through the point and a number of points in the neighborhood (frame). A polynomial order of 6 and a frame size of 15 were

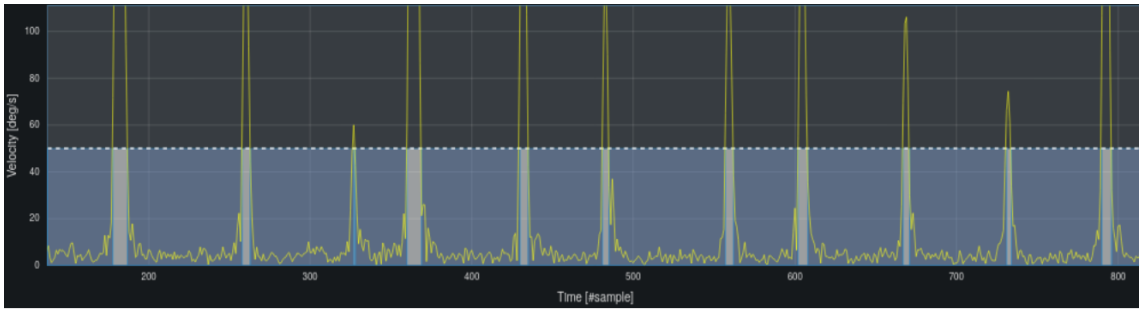
selected as used in the works of [George and Routray, 2016, Schröder et al., 2020]. These parameters were selected for all the experiments in all the datasets for both user identification and gender prediction. Figure 3.4a and Figure 3.4b show examples of velocity profiles of the participant with ID 006 from TEX dataset before and after filtering. As one can observe, the filter reduces the level of noise in the velocity data derived from the raw gaze trajectory. This leads to a better segmentation of the gaze trajectory into fixations and saccades. Another example can be seen in Figure 3.5a and Figure 3.5b for the GOF dataset from the participant with ID 191.

The next step is to segment the filtered gaze trajectories into a sequence of fixations and saccades. The term fixation is used to define the state when the eyes are focused on a specific point of interest (very slow eye movements last typically up to 600 ms). Saccades are the fast movements when the eyes move from one fixation to another (very fast movements typically last less than 100 ms). Such segmentation has shown to be beneficial for classification tasks, see e.g. [George and Routray, 2016]. Several algorithms in the literature have been used for segmentation. The following are the five most common algorithms which are used for segmentation: Identification Velocity Threshold (IVT), HMM Identification (I-HMM), Dispersion Threshold Identification (I-DT), MST Identification (I-MST), and Area-of-Interest Identification (I-AOI) see, e.g. [Salvucci and Goldberg, 2000]. The simplest identification algorithm among them to understand and implement is the IVT algorithm.

The Identification Velocity Threshold (IVT) algorithm has been utilized in numerous publications, e.g. [Holland and Komogortsev, 2011, Holland and Komogortsev, 2012, Olsen and Matos, 2012]. Also, it has been described in different ways in the literature. In [Sen and Megaw, 1984, Salvucci and Goldberg, 2000, Andersson et al., 2017], the IVT algorithm is implemented with only one parameter, i.e., the Velocity Threshold (VT). This segmentation might produce very short fixations. These very short fixations are often not meaningful because the brain requires some time to register the visual input [Olsen, 2012]. Therefore, many researchers remove the short fixations by using a second parameter named Minimum Fixation Duration (MFD). Hence, in various studies [Holland and Komogortsev, 2012, Rakoczi et al., 2013, Kasneci et al., 2021], this algorithm uses both the VT and MFD. The IVT algorithm which is described in [George and Routray, 2016] has been used in this thesis, see also Algorithm 1. In the presented implementation of the IVT algorithm, the input is visual angles (as the required thresholds are described in angular coordinates). The inverse of the mapping in Equation 3.1 is used to convert the pixel coordinates (x_{screen}, y_{screen}) into visual angles (θ_x, θ_y), see Equation 3.2. The algorithm defines as fixation all consecutive gaze points resulting in eye rotation velocities below the VT, unless the fixation would be shorter than the MFD. All other segments are identified as



(a) Data with noise



(b) Data after applying Savitzky-Golay filter

Figure 3.4: Velocity plots of a participant chosen randomly from TEX dataset with and without filtering.

saccades. Commonly used parameters in this IVT algorithm include VT = 50 °/s (for the RAN, TEX, and VST datasets) and 15 °/s (for the GOF dataset) and MFD = 100 ms [George and Routray, 2016, Schröder et al., 2020]. For the gender prediction task a velocity threshold of 50 °/s (with the Dyslexia dataset), 150 °/s (with the VST dataset), 20 °/s (with the GOF dataset) and a minimum fixation duration threshold of 100 ms were used. The parameter choices ensure that each participant has a non-zero number of fixations which is around the mean of the maximum number of fixations over all the users.

Figure 3.4 and Figure 3.5 also visualize the segmentation of fixations and saccades in the two datsets namely TEX and GOF. They show the velocities of the TEX data for the participant with ID 006 and for the GOF dataset from the participant with ID 191 distinguished by color. The horizontal bars contain information of the segmentation (saccade or fixation). The blue shading means fixation and the white means saccade.

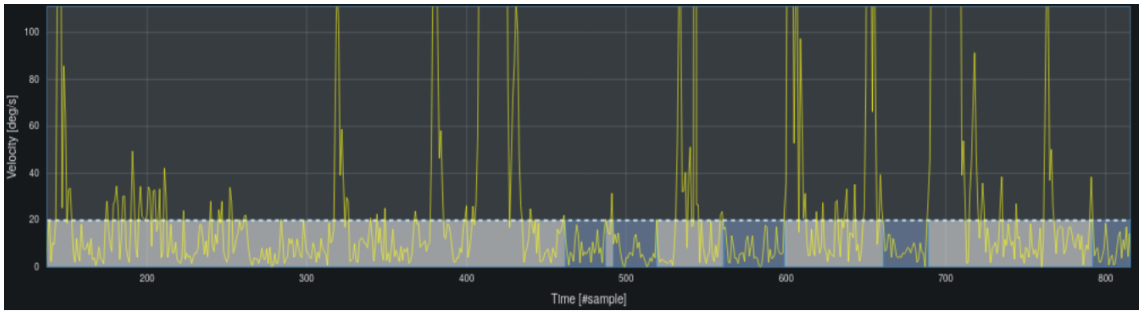
The eye tracking data can have invalid data (NaNs) or outliers (e.g invalid gaze points). This can be due to user-specific reasons such as blinking, loss of attention (micro-sleeping) or eye tracker faults (e.g. solo missed gaze points) [Rigas and Komogortsev, 2017]. However, the majority of the outliers are caused by blinks. Physiologically, the blinking behavior can encode some informa-

Algorithm 1 IVT Algorithm [George and Routray, 2016]

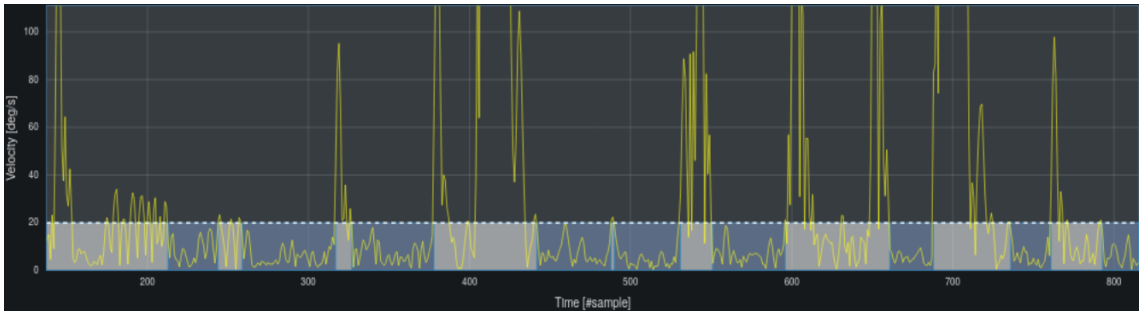
Data: [Time Gazex Gazey]
Result: Res
Constants : VT=Velocity threshold, MFD=Minimum
fixation duration
States =[FIXATION, SACCADE]
fixationStart=1
$Velocity$=smoothDiff(data) ▷ Numerical differentiation of filtered gaze coordinates
$N \leftarrow$ Number of samples of data
$lastState = NULL$
for $index \leftarrow 1$ to N **do**
  if $Velocity[index] < VT$ **then**
    currentState=FIXATION
    if $lastState \neq currentState$ **then**
      fixationStart=index
    end if
  else
    if $lastState = FIXATION$ **then**
      duration=data(index,1) - data(fixationStart,1)
      if $duration < MDF$ **then**
        for $i \leftarrow$ fixationStart to index **do**
          res[i]=SACCADE
        end for
      end if
    end if
    currentState=SACCADE
  end if
  lastState=currentState
  res[index]=currentState
end for
$Res \leftarrow$ res

tion about the participants [Kröger et al., 2020]. Actual blinking rates vary by individual averaging around 10 blinks per minute and the duration of a blink is on average between 100–400 ms according to the Harvard Database of Useful Biological Numbers[6] (see example of the blinking number and duration for all the participants of VST dataset: Figure 3.6. Blinking behavior is arguably different in men and women [Doughty, 2002] and it has been found that adults blink more often than infants [Juan, 2006]. These studies are the motivation to extract the blink information of the participants in addition to the fixations and saccades when available (which is the case of RAN, TEX and VST datasets). The blink segments are extracted from the

[6] https://bionumbers.hms.harvard.edu/ [accessed 11-August-2021]



(a) Data with noise



(b) Data after applying Savitzky-Golay filter

Figure 3.5: Velocity plot of a participant chosen randomly from GOF dataset with and without filtering

explicitly labeled invalid data (in case of RAN & TEX) and NaN segments (in case of VST) using a duration threshold between 80–500 ms, since duration of more than 500 ms are considered as micro-sleeping [Schleicher et al., 2008, Wang et al., 2011] and less than 80 ms can be device faults or other unknown reasons.

3.3 Feature Extraction

This section presents the feature extraction process for both user identification (Section 3.3.1) and gender prediction (Section 3.3.2).

3.3.1 User Identification

Feature extraction is a basic way to reduce the dimension of high-dimensional data. For each fixation, saccade, and blink (when available), various features were extracted separately. Let X and Y denote the sequences of gaze coordinates in a given fixation or saccade where $X = x_1, x_2, ..., x_N$ and $Y = y_1, y_2, ..., y_N$, and N is the number of points.

In the work presented in this thesis a number of basic features have been computed, such as duration, path length, fixation or saccade ratio (ratio of maximum

fixation or saccade angular velocity to the duration of fixation or saccade), fixation or saccade angle, amplitude, dispersion, distance with the centroid of previous fixation or saccade, angle with the centroid of previous fixation or saccade, velocity mean, and more statistical features were computed as seen in Equation 3.3 to Equation 3.14.

$$\textit{Duration} = \frac{N}{SampleRate}. \tag{3.3}$$

$$\textit{Path-Length} = \sum_{i=1}^{N-1} \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2}. \tag{3.4}$$

$$\textit{Ratio} = \frac{\max\{AngularVelocity[i] \mid 2 \leq i \leq N\}}{Duration}. \tag{3.5}$$

$$\textit{Angle} = tan^{-1}(\frac{y_N - y_1}{x_N - x_1}) \text{ between first and last points.} \tag{3.6}$$

$$\textit{Amplitude} = \sqrt{(x_N - x_1)^2 + (y_N - y_1)^2}. \tag{3.7}$$

$$\textit{Dispersion} = (\max(X) - \min(X)) + (\max(Y) - \min(Y)). \tag{3.8}$$

$$\textit{Velocity-Mean} = \frac{PathLength}{Duration}. \tag{3.9}$$

$$\textit{Mean}(\mu) = \frac{1}{N}\sum_{i=1}^{N} a_i = \frac{a_1 + a_2 + \cdots + a_N}{N}. \tag{3.10}$$

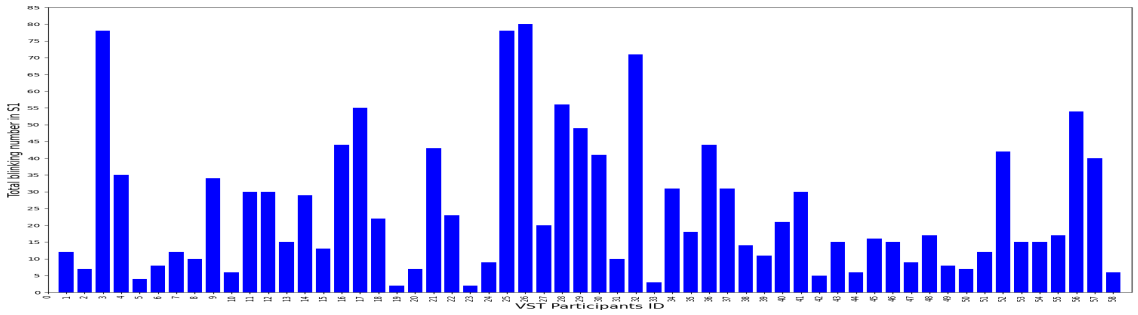
$$\textit{Median} = x_{\frac{(n+1)}{2}} \text{if } n \text{ is odd}, = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2})+1}}{2} \text{if } n \text{ is even}. \tag{3.11}$$

$$\textit{Standard-Deviation}(\sigma) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}. \tag{3.12}$$

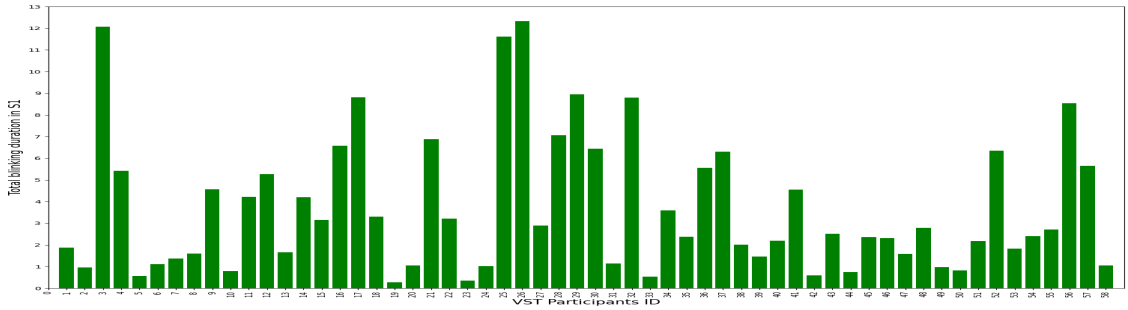
$$\textit{Kurtosis} = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i - \mu)^4}{\sigma^4}. \tag{3.13}$$

$$\textit{Skewness} = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i - \mu)^3}{\sigma^3}. \tag{3.14}$$

All these features were used in [George and Routray, 2016] and the corresponding formulae were given in that paper. The authors in [George and Routray, 2016] mainly focus on the distributional features which means that a separate value was extracted for every classified instance of fixation or saccade. Originally, they used 9 fixation features and 43 saccade features. Not all features were used in [George and Routray, 2016] (they omitted certain statistical features which are indicated with an "N" in Table 2 of that paper). The reason for this omission could be that the individual features do not add much to the classification accuracy. In our study [Schröder et al., 2020], a total of 51 features has been derived from each fixation and saccade segments including position, velocity, and acceleration features. This increased the performance by adding more features which were the combina-



(a) Blinking number



(b) Blinking duration

Figure 3.6: Blinks number and duration over the participants of VST / S1 dataset

tion of the first 51 features from Table 3.3, which shows various statistical features namely mean, median, max, standard deviation, skewness and kurtosis (in the following referred to as M3S2K features) are computed for different types of velocities and derivatives. In the context of this thesis, all the higher-order derivatives, such as velocity, acceleration, jerk, etc. (a total of 105 features, see Table 3.3) were computed using the "forward difference method" in order to improve the results further as described in Figure 3.7.

For $k \geq 1$, the k-th order derivative can only be computed for segments (fixations or saccades) of at least $(k + 1)$ points. We excluded, whenever we used such higher-order derivatives, segments that were shorter than $(k + 1)$ points (by appropriately merging the neighboring segments, in accordance to our IVT algorithm). We always removed saccades consisting of only one or two points.

To the best of my knowledge, previous approaches to eye movement biometrics include only derivatives until order two, i.e., derivatives beyond that acceleration were never used. There were works within eye tracking research that include up to fourth-order derivatives (to predict saccade movements, see [Wang et al., 2017]). In general, it is well known that humans' predictive capabilities in their perception-action loop can be captured by higher-order derivatives of the perception or action trajectories [Sargolzaei et al., 2016]. In this thesis, including higher-order derivatives beyond acceleration were demonstrated and that can indeed be beneficial for

eye movement biometrics.

Table 3.3: User identification features.

	Fix./Sac. Features		Fix./Sac. Features
1	Duration	16–21	Angular velocity*
2	Path length	22–27	Velocity X*
3	Skew X	28–33	Velocity Y*
4	Skew Y	34–39	Angular acceleration*
5	Kurt X	40–45	Acceleration X*
6	Kurt Y	46–51	Acceleration Y*
7	STD of X	52–57	Angular jerk*
8	STD of Y	58–63	Jerk X*
9	Fix/Sac ratio	64–69	Jerk Y*
10	Fix/Sac angle	70–75	Angular jounce*
11	Amplitude	76–81	Jounce X*
12	Dispersion	82–87	Jounce Y*
13	Dist. with previous Fix/Sac	88–93	Angular crackle*
14	Angle with previous Fix/Sac	94–99	Crackle X*
15	Mean velocity	100–105	Crackle Y*

*M3S2K-Statistical features:
Mean, Median, Max, Std, Skewness, Kurtosis

Lastly, seven features namely the number of blinks, duration of each blink, total of duration, mean of duration, minimum of duration, maximum of duration, and variance of duration were computed from the blink segments as shown in Table 3.4. Since a participant can have multiple blinks in their gaze trajectory, duration of each blink feature will have cardinality equal to number of blinks. On the other hand, features like mean, min, max and variance of blink duration will have cardinality one. To deal with this size mismatch, the feature vector is built with repeated entries of features with cardinality one such that its size matches the size of the duration of each blink feature.

Table 3.4: Blinks features.

	Blinks Features		Blinks Features
1	Duration	5	Minimum of the duration
2	Number of blinks	6	Maximum of the duration
3	Mean of the duration	7	Variance of the duration
4	Total of the duration		

In all cases, the features were normalized using the Z-score standardization/normalization method implemented by [Pedregosa et al., 2011] (sklearn.preprocessing.

$$
\begin{array}{lll}
\overbrace{(x_1, y_1)}^{P_1}, \overbrace{(x_2, y_2)}^{P_2}, & \ldots & , \overbrace{(x_n, y_n)}^{P_n} \quad \text{Time per sample} \\
\underbrace{d_1} \quad \underbrace{d_2} & \ldots & \underbrace{d_{n-1}} \quad \leftarrow \text{Euclidian Distance} \\
\downarrow \quad \downarrow & & \downarrow \\
v_1 \quad v_2 & \ldots & v_{n-1} \quad \leftarrow \text{Velocity} \\
\underbrace{a_1} \quad a_2 & \ldots & \underbrace{a_{n-2}} \quad \leftarrow \text{Acceleration} \\
\underbrace{J_1} \quad J_2 & \ldots & \underbrace{J_{n-3}} \quad \leftarrow \text{Jerk} \\
\underbrace{j_1} \quad j_2 & \ldots & \underbrace{j_{n-4}} \quad \leftarrow \text{Jounce} \\
\underbrace{C_1} & \ldots & \underbrace{C_{n-5}} \quad \leftarrow \text{Crackle}
\end{array}
$$

$$t = \frac{1}{samplerate}$$

$$d_i = \|P_{i+1} - P_i\|_2 \quad 1 \leq i \leq n-1$$

$$v_i = \frac{d_i}{t} \quad 1 \leq i \leq n-1$$

$$a_i = \frac{v_{i+1} - v_i}{t} \quad 1 \leq i \leq n-2$$

$$J_i = \frac{a_{i+1} - a_i}{t} \quad 1 \leq i \leq n-3$$

$$j_i = \frac{J_{i+1} - J_i}{t} \quad 1 \leq i \leq n-4$$

$$C_i = \frac{j_{i+1} - j_i}{t} \quad 1 \leq i \leq n-5$$

$$\|P_{i+1} - P_i\|_2 = \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2}$$

Figure 3.7: Computation of higher order derivative features.

StandardScaler). The method of calculation here was to determine the mean and standard deviation for each feature. Next we subtract the mean from each feature and divide the obtained value by its standard deviation. This ensures that feature distributions have mean = 0 and standard deviation = 1.

3.3.2 Gender Prediction

For the gender prediction task, the used datsets (Dyslexia, VST, GOF) were processed as explained in Section 3.2. The raw data of dayslexia was provided as angular values (VST and GOF datasets are provided as pixels values) so it was converted into pixel coordinates in order to derive a total number of 67 features (52 features

from [George and Routray, 2016] plus 3 from [Sargezeh et al., 2019] and the rest inspired by [George and Routray, 2016]). These include statistical information (e.g., mean, median, maximum, standard deviation (STD), skewness, kurtosis) from the position, velocity, and acceleration of every saccade and fixation. Also, additional features like duration, dispersion, and path length were computed (see the first 51 features in Table 3.3 as examples and see Equation 3.3, Equation 3.4, Equation 3.5, Equation 3.6, Equation 3.7, and Equation 3.8).

For the **dyslexia dataset**, all the features were calculated independently using different eye configurations including left, right, an average of both eyes features, and concatenation of the features from both eyes in one vector were studied. Averaging the features from both left and right eyes gives the best results for our experiments on this dataset. To improve accuracies (and reduce the computational load and avoid over-fitting), only a selected number of features were used, and determined by their Fisher Score [Fisher, 1936]:

$$\text{Fisher Score}(f) \quad = \quad \frac{(\text{mean}(f_{\text{class1}}) - \text{mean}(f_{\text{class2}}))^2}{\text{sd}^2(f_{\text{class1}}) + \text{sd}^2(f_{\text{class2}})}, \tag{3.15}$$

where f stands for the various values of a specific feature. For gender prediction, class1 has been taken as "male" and class2 as "female" in Equation 3.15 to calculate the Fisher Score ranking of a given set of features. Substituting males instead of class1 and females instead of class2 in Equation 3.15 for all the 185 participants is an example to calculate the top Fisher score for gender prediction. Table 3.5 shows the top five fixation and saccade features ranked by their Fisher Score (for all 185 participants). It should be noted that the prediction accuracy achieved with using the Fisher Score were similar to those obtained by ANOVA [Weerahandi, 1995, de Souza Jacomini et al., 2012]. In Table 3.6 an example of the top features by ANOVA for the Non-dyslexic participants group can be seen.

Table 3.5: Features ranked by Fisher Score (FS) for Dyslexia datset

Fixation features	Fisher Score ($\times 10^{-4}$)	Saccade features	Fisher Score ($\times 10^{-4}$)
1) Mean fixation duration	300	1) Angle to previous saccade	30
2) Ratio fixation over saccade time	200	2) Saccadic amplitude	20
3) Total fixation duration	50	3) STD of X direction	18
4) Median of acceleration X	34	4) STD of velocity Y	15
5) Mean of acceleration X	33	5) STD of acceleration Y	15 (14.89)
6) Minimum of acceleration Y	31	6) Median of velocity X	15 (14.86)

Table 3.6: Features ranked by ANOVA Score for Non-Dyslexic group (Gender prediction)

Fixation features	ANOVA Score	Saccade features	ANOVA Score
1) RFDSD	5.50	1) Skew of X direction	6.91
2) Velocity mean	5.26	2) Skew of Y direction	2.85
3) Mean of angular velocity	4.58	3) STD of Y acceleration	2.31
4) Median of ngular velocity	4.13	4) Maximum Y acceleration	2.19
5) Fixation ratio	4.06	5) STD of Y velocity	2.12
6) STD of angular velocity	3.96	6) Max of Y velocity	2.10

For the **GOF and VST datasets**, the ANOVA method is used to select the best features for the gender prediction task in each subset (e.g 12 seconds, 20 seconds, 40 seconds, 1 minute, and 2 minute from the trajectory) of the trajectory used for the experiments. (Tables 3.7 and 3.8 showed examples of the top six fixation and saccade features ranked by their ANOVA Score for VST and GOF datasets respectively). The used features were normalized using the Z-score standardization/normalization method as explained in Section 3.3.1

Table 3.7: Features ranked by ANOVA Score (AS) for the VST dataset using session 1 and 12 seconds trajectory length (start of the trajectory)

Fixation features	ANOVA Score	Saccade features	ANOVA Score
1) distance with previous Fix	5.70	1) Median of acceleration X	7.73
2) Minimum of velocity Y	5.60	2) Max of acceleration X	6.35
3) STD of velocity Y	5.55	3) Skewness of velocity X	4.81
4) STD of angular velocity	5.39	4) Skewness of Y	4.25
5) Var of angular velocity	5.09	5) Mean of acceleration X	3.85
6) Var of velocity Y	5.00	6) Kurtosis of acceleration X	3.21

In addition to the fixation and saccade features that were extracted from the full trajectory region of the stimuli (lets name them as fullR) there are different types of features that were extracted from the gaze trajectory for the **GOF dataset** by Rishabh Haria from the database group at University of Bremen. Region of Interest (ROI) features were used for gender prediction task with GOF dataset. These features were statistical features M3S2K computed in each region of interest individually. Inspired by [Coutrot et al., 2016, Sammaknejad et al., 2017, Baron-Cohen, 2002] four regions were used: left eye, right eye, nose, and mouth.

Table 3.8: Features ranked by ANOVA Score (AS) for the GOF datset using 120 seconds trajectory (age group 20–72)

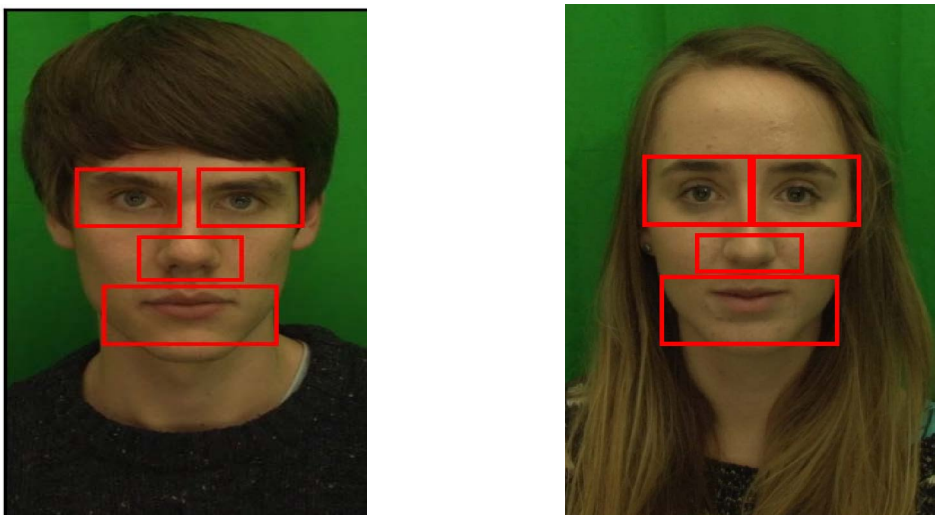
Fixation features	ANOVA Score	Saccade features	ANOVA Score
1) Maximum of angular velocity	24.40	1) Saccade ratio	39.29
2) STD of angular velocity	18.52	2) Mean of angular velocity	27.55
3) Minimum angler acceleration	18.16	3)Mean of angular acceleration	23.63
4) Mean of angular velocity	17.89	4)STD of angular velocity	20.49
5) Median of angular velocity	17.44	5) STD of angular acceleration	16.79
6) Maximum of angular acceleration	16.04	6) Maximum of angular velocity	12.51

Table 3.9: ROI features ranked by ANOVA Score (AS) for the GOF datset using 12 seconds trajectory (age group 20–72)

Fixation features	ANOVA Score	Saccade features	ANOVA Score
1) Total velocity in mouth region	7.75	1) Saccadic amplitude in nose region	14.03
2) Total distance in mouth region	7.74	2) Total velocity in nose region	9.90
3) No. of Fix in mouth region	6.98	3) Total distance in nose region	9.84
4) Total duration in moth region	5.57	4) No. of Sac in nose region	6.65
5) Mean duration in moth region	4.93	5) No. of Sac in mouth region	4.19
6) No. of Fix in nose region	3.23	6) Total duration in the right eye region	1.98

These ROIs were labeled manually as rectangles (with size of eyes 120×80, nose 120×50, and mouth 200×50 [Coutrot et al., 2016]). When computing a feature for the ROI, e.g. left eye, only the trajectories inside the left eye region were considered. For instance; Saccadic amplitude in mouth region, Total velocity in nose region. The study in [Komogortsev et al., 2010] reported that rich amount of information was present in saccades about the dynamics of the oculomotor plant. Hence, saccade amplitude and saccadic ratio were also extracted in these ROIs. The temporal properties were leveraged by using distance and angle with previous fixation and saccades as a feature. Figure 3.8 shows the labeled ROIs on two stimuli.

The ANOVA method was used again to select the best ROI features in each subset (e.g 12 seconds, 20 seconds, 40 seconds, 1 minute, 2 minute) of the trajectory used for the GOF dataset Table 3.9 shows an example of the top six ROI features ranked by their ANOVA Score for GOF datasets.



(a) Male actor (b) Female actor

Figure 3.8: Region of interest marked on the stimuli of GOF dataset

3.4 Machine Learning Classifiers

For given trajectories of eye movements of a number of participants, we seek an algorithm that, given an unseen trajectory (testing set), is able to detect which class (user id in case user identification task and male/female in gender prediction) has generated it. This type of problem is known as a *classification problem*, and the algorithm that carries out the classification is known as a *classifier*. It must be noted that the number of classes in the user identification task is equal to the number of users considered and gender prediction is a binary classification task. A maximum of three instances of machine learning classifiers were trained for the prediction task: one from fixation segments, the second from saccade segments and third from blink segments (whenever available). The final prediction probability p^i_{final} is the weighted average of the probabilities of the fixation (p^i_{fix}), saccade (p^i_{sac}) and blink (p^i_{blink}) classifiers for each class i (user ID in user identification problem or Male/Female in gender prediction):

$$p^i_{\text{final}} = p^i_{\text{fix}} w_{\text{fix}} + p^i_{\text{sac}} w_{\text{sac}} + p^i_{\text{blink}} w_{\text{blink}},$$

where w_{fix}, w_{sac}, and w_{blink} are the weights for the fixation, saccade and blink classifiers respectively. In case the blink classifier is absent, w_{fix}, w_{sac} are typically selected as 0.5 each [Schröder et al., 2020, George and Routray, 2016]. For m classes, the class having the maximum probability $p_{\text{max}} = \max\{p^i_{\text{final}} \mid i \in \{1, \ldots, m\}\}$ is the final outcome i of the classifier. Note that it never occurred that there was more than one user i having the maximum probability. The following classifiers were used in this thesis, Radial Basis Function Networks (RBFN) and Random Forests (RF) were

used in user identification task while all Radial Basis Function Networks (RBFN), Random Forests (RF), Logistic Regression (logReg), Support Vector Machine (SVM), and Naïve Bayes (NB) were used in gender prediction task. In the following a short summary of these classifiers is provided:

3.4.1 Radial Basis Function Networks (RBFN)

The machine learning classifier Radial Basis Function Networks (RBFN) [Broomhead and Lowe, 1988] has been used in [George and Routray, 2016] and also in our work in [Schröder et al., 2020]. RBFN is based on a feed-forward neural network based machine learning algorithm [Caselli et al., 2009, Derks et al., 1995]. After the input layer, the fully connected hidden layer was in the proposed network with $C \cdot K$ neurons, where C is the number of classes and K is a hyper-parameter. Then, a radial basis function activates each neuron in the hidden layer:

$$\phi(x) = e^{-\beta||x-\mu||^2} \tag{3.16}$$

Here, $||x - \mu||$ is the Euclidean distance (L_2-norm) between two points x and μ. Furthermore, β can be calculated as follows:

$$\beta = \frac{1}{2\sigma^2}, \tag{3.17}$$

where μ and σ are found for all neurons with a training procedure as follows. Their class labels partition all training data X. For each subset X_c, $c =$ class ID, we generate K seed vectors $\mu_{c,i}$, $i = 1, \ldots, K$, using K-means clustering. These cluster centers are the mean of all elements from X_c that belong to the corresponding cluster. Then, $\sigma_{c,i}$ is computed as the mean Euclidean distance of all elements in each cluster i to $\mu_{c,i}$.

We weigh the output vector from the hidden layer $\varphi(x)$ with length $K \cdot C$ in a fully connected way to produce a prediction vector y of length of C. The weights matrix W has a dimension of $K \cdot C \times C$. During training, W can be learned either by gradient descent or using the Moore-Penrose pseudoinverse to minimize the least squares error between the output vector and the one-hot encoding of the training labels.

RBFN offers properties similar to back-propagation neural networks [Lee, 1991]. They perform a weighted linear combination with a nonlinear, although confined, transformation. RBFN, on the other hand, can often be trained considerably quicker than back-propagation neural networks. An RBFN uses a linear combination of radially symmetric functions to construct an input-output (IO) mapping. As illustrated in Figure 3.9, it contains three layers with feed-forward connections between the nodes.

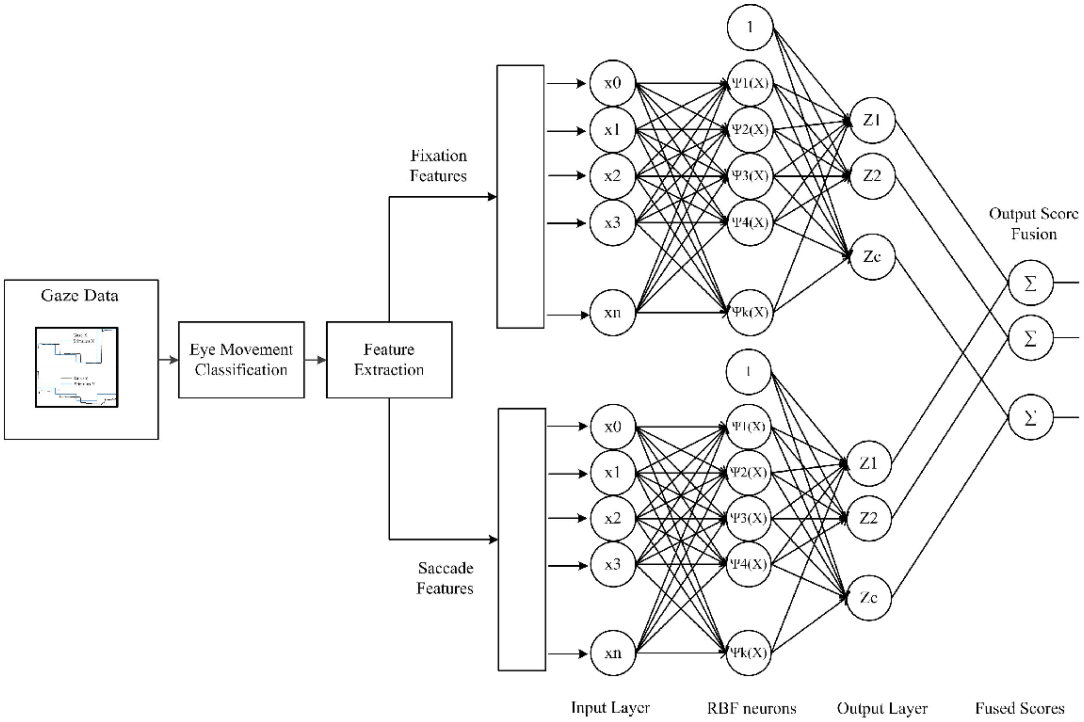


Figure 3.9: RBFN [George and Routray, 2016]

Hyper Parameters of RBFN:

- Parameter K: In RBFN, there is only one hyper-parameter which is the number of representative vectors from each class (number of clusters, K) selected by using a standard K-means algorithm. For each class, a number of cluster centers were used for fixations and saccades resulting in $K \times N$ clusters for each RBFN classifier (N is the number of classes and K is the number of the cluster centers). $K = 32$ was used as the default value, while with less than one minute trajectory length, a re-tuned value of K is implemented (e.g. in user prediction experiments in Section 4.7).

3.4.2 Random Forests (RF)

Another machine learning algorithm is the random forest (RF). Random forests are a powerful and popular classification method [Breiman, 2001, Cutler et al., 2012]. The RF classifier creates several decision trees, which are combined as weak classifiers by using the "bagging" approach [Breiman, 1996]. This approach is based on combining learning models to improve the outcomes. It uses majority voting technique to decide the final outcome from the various decision trees. As mentioned, random forests are comprised of decision trees, every tree in the forest relies on the values of a random vector selected independently and with the same distribution [Breiman, 2001].

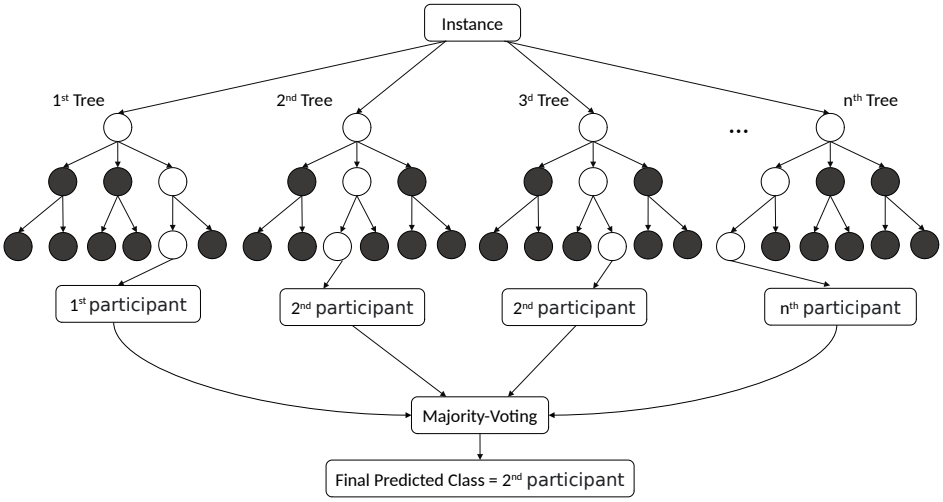


Figure 3.10: Random forest classifier

Using a random set of features to divide each node produces error rates that seem comparable to Adaboost (Adaptive Boosting) [Freund et al., 1999], but more resilient in the presence of noise [Kotsiantis et al., 2006]. Figure 3.10 represents an abstraction of random forest classifier for user prediction example. It can be noticed that the outcome of second and third decision trees in this figure is 2nd participant while the others differ from each other. Hence the majority voting decides the final predicted class as the 2nd participant.

Hyper Parameters of RF: This classifier has many parameters, e.g. maximum features, n-estimators, minimum samples split, minimum samples leaf, and maximum depth. Most of the parameters were the default values of the scikit-learn software package [Pedregosa et al., 2011] which was used in this thesis. Only two parameters were tuned by grid search method which were maximum-depth and n-estimators.

- Maximum-depth: It denotes the maximum depth of the decision tree (how many depth layers the input can be split into). Increasing the splits leads to better variation in the data.

- N-estimators: It denotes the numbers of the trees of RF algorithm. Higher number of trees are increasing the performance and making the prediction more stable however, this will increase the computation load also can over-fit the data.

3.4.3 Logistic Regression (logReg)

Logistic Regression (LogReg) [Bishop, 2006] is a machine learning algorithm used for classification and regression challenges. Based on input variables, logistic regression can estimate the possibility of a categorical output. Most logistic regression models include a binary result, including true/false and yes/no. Multinomial logistic regression can classify events with more than two discrete outputs. Logistic regression is a powerful analytical tool for classification in which we want to know whether a given data belongs in a class [Edgar et al., 2017].

This classifier employs the concepts of probability to predict the class. A *probability* is a value that measures the possibility or chance of a specific event. A probability of 0 shows that an event cannot happen, while a probability of 1 suggests that an event is very likely to happen. LogReg uses the Sigmoid function as a cost function to map predicted values into probabilities ranging from 0 to 1. The probability threshold determines the class. If the predicted value exceeds the threshold, the result is classified as class one; else, it is classified as class two. The usage of this classifier is more suitable if the dependent variables are binary, which means that they fall into one of two categories.

The Logistic Regression equation is provided below:

$$L = \log(\frac{p(X)}{1 - p(X)}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \tag{3.18}$$

where L denotes the log-odds of the event $X = 1$, $p(X)$ represents the probability of event X where X is any feature. The left-hand side shows the ratio of X being predicted correctly to X being false predicted. On the right-hand side, β_i illustrates the regression coefficients associated with the reference group β_0.

Hyper Parameters of logReg: Always the default parameters of scikit-learn package were used with LogReg in this thesis.

3.4.4 Support Vector Machine (SVM)

Support Vector Machine Networks [Cortes and Vapnik, 1995] is another machine learning algorithm used for classification purposes. This algorithm separates n-dimensional space into classes by creating a boundary line (Hyperplane), ensuring that each new data point is assigned to the correct class. The margin is the distance between the hyperplane and the nearest associated point. The optimal hyperplane has the maximum margins before hitting data points. These points are support vectors; hence the algorithm is called a Support Vector Machine. Figure 3.11 presents an example of support vectors.

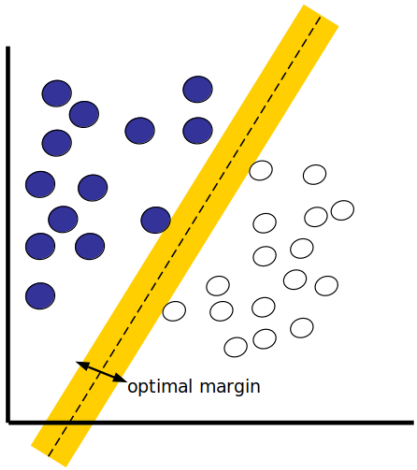


Figure 3.11: An example of Support Vector Machine Networks for two dimensional space. The support vectors, marked with yellow line, define the optimal margin of largest separation between the two classes.

This classifier employs a collection of mathematical functions known as the kernel [Noble, 2006]. The kernel's role is to collect input data (features) and convert it into the desired shape, enabling the algorithm to separate and classify the data. Each SVM algorithm uses a different kernel function. Kernels can be linear, non-linear, polynomial, radial basis function (RBF), and sigmoid functions. Kernel maps the data into feature space to make them separatable with this transformation. The feature space is a set of all probable values for a particular set of features from input data [Scholkopf et al., 1999]. Some of the kernels include linear, polynomial and radial kernels.

Hyper Parameters of SVM: In this work, RBF kernel is used for the SVM classifier which has two relevant hyperparameters:

- Regularization Parameter (C): This parameter is used to control the error. The regularization parameter determines the importance of mis-classifications. It specifies how many mis-classification examples can be avoided during SVM optimization. For high values of C, the optimizer chooses a hyperplane with a smaller margin if it correctly classifies all training points. In contrast, a relatively small value of C drives the optimizer to seek a separating hyperplane with a broader margin, even if it mis-classifies more points than other hyperplanes. See Figure 3.11 for a pictorial depiction of hyperplane and its separation margin.
- Gamma parameter: This parameter determines how far the influence of a single training sample extends; for lower gamma values, the far-away points are

considered, while for higher specified values, only points nearer to the plane are considered.

3.4.5 Naïve Bayes (NB)

Naïve Bayes (NB) [Bayes, 1968, Rish et al., 2001] is a probabilistic machine learning algorithm based on Bayes' theorem. The word Naïve means the assumption of the independence of the features from each other. This method states that a feature's values are unaffected by the attributes of other features and are not reliant on them, meaning that it presumes the independencies among the predictors (inputs for ML algorithm). While the Naïve Bayes algorithm is based on an idealistic hypothesis, it has shown to be effectively practical, frequently challenging with considerably more advanced classification approaches [Hilden, 1984, Langley et al., 1992, Friedman et al., 1997, Domingos and Pazzani, 1997]. Bayes theorem states that:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3.19}$$

where $P(A|B)$ refers to the predicted probability of event A based on event B that is already occurred, $P(B|A)$ refers to the probability of event B based on event A that is already occurred. Bayesian classifiers determine the most probable class for a given case based on its feature vector [Rish et al., 2001]. Based on the given features, Naïve Bayes predicts the probabilities for each class, and the class with the maximum probability is chosen as the predicted class. It is possible to considerably simplify the learning of such classifiers by presuming that features are independent of class, that is, $P(X|C) = \Pi_{i=1}^{n} P(X_i|C)$, where $X = (X_1, ..., X_n)$ is a feature vector, and C is a class.

Hyper Parameters of NB: In this work, no tuning of the hyperparameters for the Naive Bayes classifier provided by the scikit-learn is performed.

3.5 Performance Metrics

Metrics were applied to monitor and evaluate a machine learning model's performance during the training and testing process. They assisted to find how accurate future (unseen/out-of-sample) data will be predicted. For a single experiment, we always calculated predictions for all available user classes or gender classes. The following sections were showing the metrics used for each classification task.

3.5.1 User identification

Accuracy Prediction accuracy is a measure that quantifies a classification model's performance by dividing the number of correct predictions by the overall number of predictions. It is simple to compute and comprehend, making it the most often used statistic for assessing classifier models [Steyerberg et al., 2010]. As accuracy of user identification, the number of correct predictions were divided by the total number of predictions (equal to the number of users). As shown in Equation 3.20:

$$Accuracy = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right). \tag{3.20}$$

Where TP, TN, FP, and FN denote True Positive, True Negative, False Positive and False Negative respectively.

Standard Deviation (STD) Standard deviation is a statistical measure that estimates a dataset's distribution compared to its mean [Dudoit and Fridlyand, 2002]. By determining the variation of each data point from the mean, we determine the standard deviation as the square root of the variance [Streiner, 1996]. There is a significant deviation within the dataset if the data points are far from the mean; hence, the larger the standard deviation, the more spread out the data. The standard deviation can be calculated as follows:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \tag{3.21}$$

where x_i is the value of the i^{th} point in the dataset, $\bar{x}$ is the mean value, and n is the number of data points.

Standard Error of the Mean (SEM) The standard error of the mean is a measure of how much the sample mean (average) of the data is expected to differ from the actual population mean. Thus, the SEM is always smaller than the STD [Altman and Bland, 2005]. Together with each accuracy value we also report the standard error of the mean which is computed as

$$\sigma_\mu = \frac{\sigma}{\sqrt{k}}, \tag{3.22}$$

where σ is the standard deviation and k the number of runs.

3.5.2 Gender prediction

The accuracy, STD, and SEM that were mentioned in Section 3.5.1 are used for gender prediction problem with the following convention: TP = True Males, TN = True Females, FP = False Males and FN = False Females. In addition, the following metrics also applied for gender prediction task:

Precision (Positive Predictive Value) The precision gives an insight on how our classifier might be biased towards false positives. If the precision is low, it means the classifier has a high number of false positives. The precision is the number of true positive results divided by the number of all positive results, including those not identified correctly. It can be calculated by the following equation:

$$Precision = \frac{TP}{TP + FP} \tag{3.23}$$

where TP = True Males and FP = False Males.

Recall (Sensitivity) The recall gives an insight into how our classifier might be biased towards false negatives. If the recall is low, it means the classifier has a high number of false negatives. This is very important for e.g. if we are dealing with a cancer prediction classifier as we do not want to include any false negatives in our dataset. The recall is the number of true positive results divided by the number of all samples that should have been identified as positive. Recall can be calculated by the following equation:

$$Recall = \frac{TP}{TP + FN} \tag{3.24}$$

Here, TP = True Males and FN = False Female.

F1-Score (F-measure) The F1-score, is the harmonic mean of accuracy and recall. It is a common measure that combines precision and recall. When precision and recall are both equally important for a classification task, it is recommended to look at the F1-score. It is calculated as follows:

$$F1 - score = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall} \right) \tag{3.25}$$

Confusion Matrix The confusion matrix, a tabular representation of model predictions versus the real labels, is an important concept in classification performance. Cases in a predicted class are represented by each row of the confusion matrix, while each column defines the occurrences in an actual class [Tharwat, 2020].

		Truth: Yes	Truth: No
Prediction	Yes	TP	FN
	No	FP	TN

Figure 3.12: Sample confusion matrix

The diagonal components represent the true prediction for each class, but the off-diagonal elements represent misclassified data.

Receiver Operating Characteristic Curve (ROC Curve) A receiver operating characteristic curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier as its discrimination threshold is varied. Each confusion matrix corresponds to a ROC curve. The area under the curve (AUC) is an evaluation of a binary classifier's overall possible threshold values [Fawcett, 2006]. The AUC estimates the area under the ROC curve and is between 0 and 1 (perfect classifier). The AUC can be defined as the probability that the model rates a random positive case higher than a random negative example. The AUC can be used to identify which classifier is better. The larger the area under the ROC curve, the better is the classifier (e.g. AUC3 is the best in Figure 3.13).

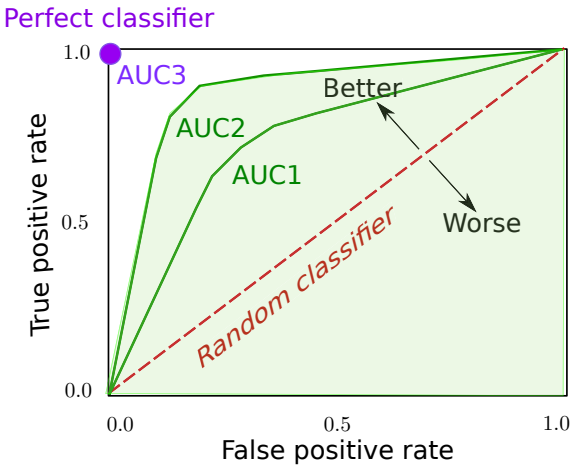


Figure 3.13: ROC curve.

3.6 Statistical analysis methods

There are multiple statistical hypothesis tests available in the literature. Each test aims to find if there is a difference in one of the statistical properties. Statistical properties include for example the standard deviation, mean or variance.

The t-test [Student, 1908, Witkovsky, 2013], ANOVA-test [Girden, 1992], Kruskal-test [Kruskal and Wallis, 1952, Daniel, 1990], and Wilcoxon signed-rank test [Woolson, 2007] are examples of these statistical methods. ANOVA and t-test are used to determine if the means of two or more groups are different.

The Wilcoxon signed-rank test and Kruskal-test are used to determine if the median of two or more groups are different and they are non-parametric methods.

The t-test and ANOVA-test can only be used with normally distributed data while the Wilcoxon signed-rank test and the Kruskal-test can be used with not normally distributed data.

In the scope of this thesis, the feature data is found to be not normally distributed (see Figure 5.6) and hence Wilcoxon signed-rank test is used as a statistical test to find if there are significant differences and is described in the following:

3.6.1 Wilcoxon-Test:

The Wilcoxon Signed Rank test is a non-parametric test (such a test also called a distribution-free test) that compares the median of two dependent samples and determines whether or not there are statistically significant differences. Wilcoxon Signed Rank test is used when the data does not meet the requirements for a parametric test (i.e. if the data are not normally distributed where mean and variance are not equal). The Wilcoxon sign test is an alternative to the dependent samples t-test that can be used to compare two groups when data do not follow the normal distribution.

The logic behind the Wilcoxon test is quite simple; first, rank the data differences to produce two rank totals, one for each condition. In case there is a systematic difference between the two conditions, then most of the high-rank scores will belong to one condition and most of the low-rank scores will belong to the other one. As an outcome, the rank totals will be different and one of the rank totals will be small. In contrast, if the two conditions are identical, then high and low ranks will be distributed fairly evenly between the two conditions and the rank totals will be fairly similar and large. The Wilcoxon test statistic W is simply the smaller of the rank totals. Taking into account how many participants we have, the SMALLER it is then less likely it is to have occurred by chance. A table of critical values of W shows us how likely it is to obtain our particular value of W purely by chance. Note that the Wilcoxon test is unusual in this respect: normally, the BIGGER the test statistic, the

less likely it is to have occurred by chance. If one has a large number of participants, W can be converted into a z-score and should be compared instead. The formula for the Wilcoxon signed-rank test is mentioned in Equation 3.26.

$$W = \sum_{i=1}^{N} Sgn(x_{2,i} - x_{1,i}) \cdot R_i, \tag{3.26}$$

where N is the sample size without the case when $x_1 = x_2$, $x_{1,i}$ and $x_{2,i}$ are the ranked pairs from the distribution, Sgn is the sign function, and R_i is the rank of i.

A p-value is a likelihood of obtaining a test statistic equal to or greater than the value specified by the Wilcoxon test, assuming that both distributions are identical. The p-value is a conditional probability, and it is based on the circumstance that the null hypothesis (H_0) is true. This point is critical since it implies that the p-value cannot determine whether H_0 is true or not [Dorey, 2010]. The null hypothesis is an assumption that claims that statistical differences between observable data or measurable events are not existent [Haldar, 2013]. It claims that the findings are coincidental and have no relevance to the validity of the hypothesis under investigation. As a result, the null hypothesis suggests that whatever we attempt to verify did not occur. A lower p-value indicates that the alternative hypothesis is more likely to be true. P-values decide whether or not the findings of the hypothesis test are statistically significant. The p-value is determined by the statistical sample distribution of the test under the null hypothesis, the sample data, and the kind of test being conducted. If a p-value less than 0.05 is achieved the null-hypothesis (no differences) is rejected which means that differences are significant between the two groups and if a p-value less than 0.01 is achieved it means there is one in a thousand chance of being wrong which is referred to as rare so the differences are highly significant [Gibbons and Pratt, 1975].

3.7 Feature selection methods

Choosing the most consistent, non-redundant, and relevant features to utilize in a model is known as feature selection. The primary purpose of feature selection is to increase a predictive model's performance while lowering its computing cost [Togaçar et al.,]. ANOVA is one of these methods which is used in this thesis and is described in the following subsection.

3.7.1 ANOVA

The ANOVA (Analysis of Variance) method aids in the selection of the important features. A feature's variance affects how much it influences the response variable. Variables with less variance are not affecting the response feature (dependent feature) [Wahba et al., 1993]. ANOVA employs the F-Distribution statistical examination to compute the probability distribution for variance analysis. It is a parametric statistical hypothesis method used to determine whether or not the means of two or more samples of data originate from the same distribution. To summarize, if the null hypothesis is rejected, it signifies that there is variance across the groups, implying that the target variable influences another variable. The first step for ANOVA feature selection is setting up a hypothesis and determining the significance level. In the null hypothesis, means are taken as equal. If there is an equal mean between groups, this feature has no impact on response and it can be removed from the considered features for the model. Next steps to perform ANOVA are as follows: Calculate the sum of squares which is the statistical technique used to determine the dispersion in data points. That is the measure of deviation and can be written as in Equation 3.27

$$\textit{Sum-sequares} = \sum_{i=0} \left(X_i - \mu\right)^2 . \tag{3.27}$$

where x_i is the value of the i^{th} point in the dataset, μ is the mean value of all points, and $x_i - \mu$ is the deviation of each item from the mean. After that implement the F-test to compare the variance within groups by total sum of squares. Then calculate the distance between each group average value g from grand means μ is $g - \mu$. Analogous to the total sum of squares as as in Equation 3.28:

$$\textit{SSB} = \sum \left(g_i - \mu\right)^2 . \tag{3.28}$$

The distance between each observed value within the group x from the group-mean g is given as x-g. Analogous to the total sum of squares (see Equation 3.29)

$$\textit{SSE} = \sum \left(X_i - g\right)^2 . \tag{3.29}$$

The total sum of squares = SSB (Between Sum of Squares) + SSE (Within Sum of Squares). Finally, the F-Value is to compare variance between the groups and variance within the groups as follows.

$$\textit{F} = \frac{SSB}{df_a} / \frac{SSE}{df_b}. \tag{3.30}$$

where df is the degrees of freedom and it is equal to number of the groups reduced by one.

3.8 Eye Movement Trajectory Visualization Tool

This section represents the summary of the visualization tool [Prinzler et al., 2021] that was used to visualize the eye movements trajectories of all the datasets used in this thesis. The development of this tool was initiated by Christoph Schröder in the Institute for Computer Graphics and Virtual Reality at University of Bremen and was further developed by Martin H.U. Prinzler in the Database Group at University of Bremen. The tool is written in Python (using the Bokeh module) that allows to interactively gauge the performance (e.g., accuracy) of different classification methods (such as RF, RBFN, SVM, etc.) on eye tracking data. The main idea is to visualize the fixation and saccade segments, and color them according to their positive/negative influence towards the prediction. Moreover, the user is able to interactively change various parameters (such as size and order of the polynomial in Savitzky-Golay filter [Schafer, 2011b] for smoothening the raw data, velocity and minimal fixation duration threshold for IVT algorithm etc.). Figures 3.14 and 3.15 show the complete view of the graphical user interface of the tool and the four main parts of this tool are described in the following:

(1) Trajectory plot: This part is used to visualize gaze point trajectories after selecting the data set and one or more users for a selected time span. Optionally, the segmentation can be shown. The fixations are shown as circles, annotated with consecutive numbers and the saccades as dashed lines. The top left of Figure 3.14 shows the trajectories of four participants from the MIT dataset with user IDs *emb, kae, po* and *tmj* with colors violet, orange, green and yellow respectively.

(2) Velocity plot: This part is shown at the bottom of Figure 3.14. It shows the velocities of the previous four users distinguished by color. For each user, to the right of the name, the horizontal bars contain information of the segmentation (saccade or fixation). A blue shading means fixation, and white means saccade. As can be seen, three users (emb, kae, and po) have their first saccade after a similar time span, while the user tmj has a later first saccade. Further, the saccade duration of the users emb, po, and tmj are similar while the user kae has a longer saccade duration. For more information on velocity visualization, please refer to Figure 3.4 and Figure 3.5.

(3) Feature table part: This part provides a direct access to all the calculated fea-

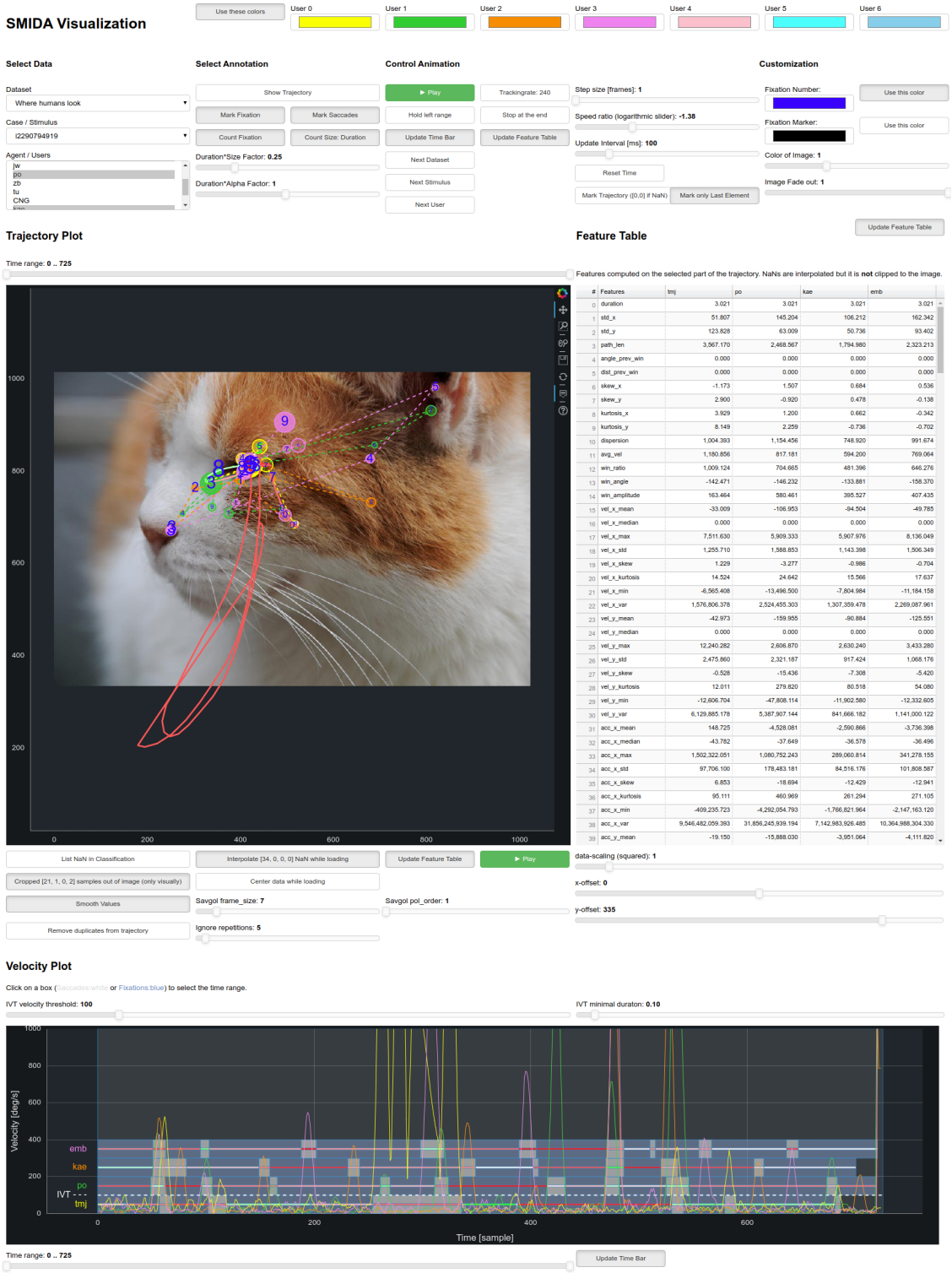


Figure 3.14: First part of our graphical user interface with *Settings* to control the behavior, *Trajectory Plot* to show stimulus and 2D tracking information, *Feature Table* to have detailed access to statistic data, and *Velocity Plot* for understanding segmentation.

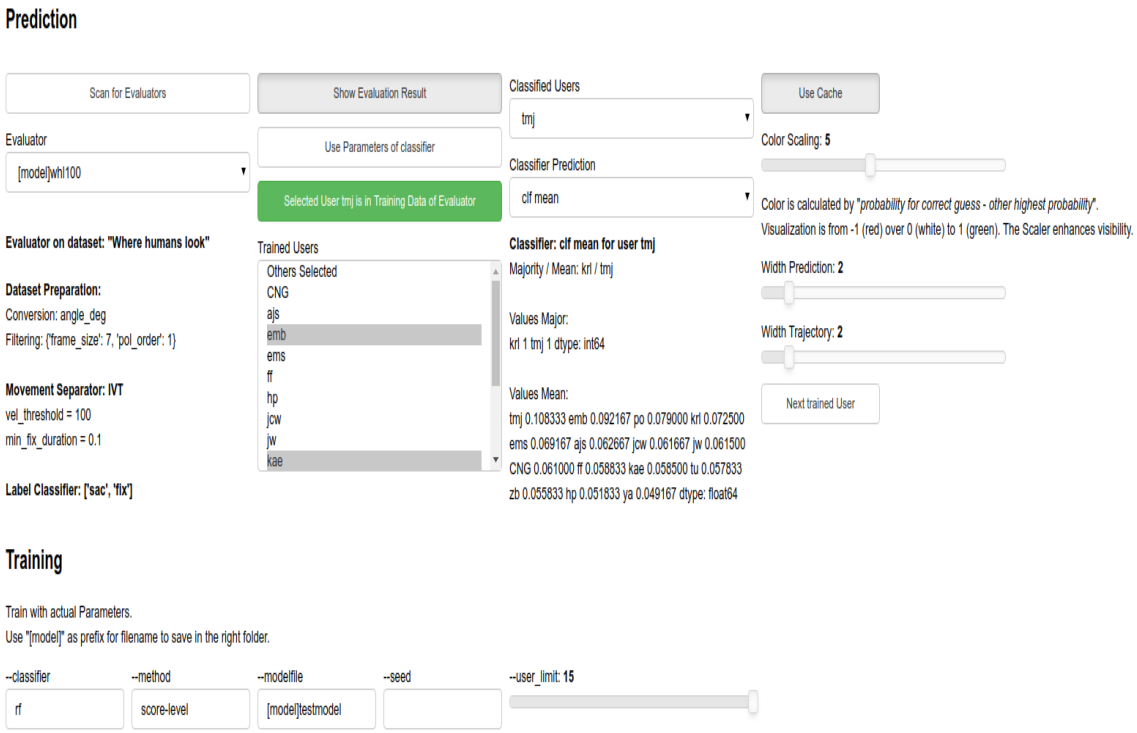


Figure 3.15: The second part of our interface with *Prediction* part to visualize the correctness of classifiers and get detailed information of results, as well as *Training* part to have direct access to our classifier training module.

tures in the selected time range in a tabular form (see right of Figure 3.14). For example, it can be seen that the second feature std_x, which means the standard deviation in horizontal direction, is much lower for the user tmj, than for the others.

(4) Visualization of Classifier Predictions: This part is used to show the prediction information (correctness of classification; by colors). Green color means a segment is correctly classified, red shows that the classifier predicted a wrong label and white depicts unclear (see Figure 3.16). To have more details of the prediction, in the part Prediction of our interface (Figure 3.15, top, third column), the absolute values of the probabilities are shown. The following table shows an excerpt:

User	*tmj*	*emb*	*po*	...
Probability	0.1083	0.0922	0.0790	...

These are the mean probabilities of a weighted combination of the classifiers for saccades and fixations. We can also select a mean over all segments regardless

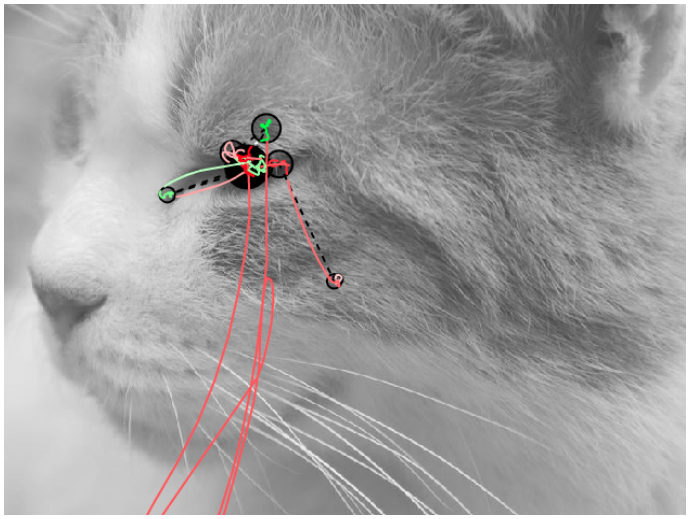


Figure 3.16: Trajectory plot with color coded correctness of classification for a participant of the MIT dataset. Color from green (beneficial) to white, to red (harmful). The segmented parts are augmented in black: saccades as dashed lines and fixations as circles.

to which group they belong, or the values of a single classifier (for a specific group, like only the fixations).

3.9 Conclusion

This chapter explains the system design (pipeline) in this study for both user and gender classification. It describes in detail the used datasets. For user prediction, all the following datasets were used: RAN, TEX, MIT, VST, and GOF. While for gender prediction, Dyslexia, VST, and GOF were used. Further, the chapter covers the pre-processing process which is the same for both tasks. After that, the segmentation algorithm was explained i.e. the IVT algorithm. Further, feature extraction was introduced. For user prediction, the used features were inspired by [George and Routray, 2016] and we add more features that improved the user identification as reported in [Schröder et al., 2020]. We then improved the user identification further by adding higher derivative features. For gender prediction, selecting the features was by using different methods of features ranking e.g. the Fisher score, and ANOVA. Finally, we mention the ML classifiers and the performance metrics used in each prediction task. For user identification, the RBFN and RF classifiers were used, while all the mentioned five classifiers were used for gender prediction.

Chapter 4

User Identification

This chapter is covering all the experiments on user identification that were conducted in this thesis. User identification was carried out with the RAN, TEX, MIT, VST, and GOF datasets. The work presented in this chapter is based on an initial collaboration with Christoph Schröder on the RAN, TEX, and MIT datasets in [Schröder et al., 2020] and later an extensive analysis of the proposed approach in [Zaidawi et al., 2022] for the above mentioned datasets. Both authors contributed equally to the work in [Schröder et al., 2020]. Later, this work was extended in [Zaidawi et al., 2022] for two more datasets namely VST and GOF and the proposed approach was extensively analyzed with respect to various factors (e.g. age, gender) and extended by including higher order derivative features, blink information and tuning the IVT parameters, each of which brings a significant increase in user identification accuracies.

The hypothesis of user identification via eye movement data is that the gaze trajectories recorded by an eye tracker encodes the user's identity. Different participants may have an inherently different eye movement behavior which can be attributed to their physiological parameters. However, the data recorded by a device may be additionally subject to the device characteristics for e.g. noise, sampling frequency etc. As explained earlier in Section 3.2 as the first step, the noise was filtered out to get filtered gaze trajectories $x(t)$. Next, the IVT algorithm was used to segment the gaze trajectory into fixations and saccades. The segments of fixations and saccades can be used to compute several features for user identification. The time differentiation of some of these features (denoted as s for saccade and as f for fixations) is physically meaningful for e.g. path length, velocities, acceleration etc. In general, one may denote n^{th} time derivative of s and f as $s^{(n)}$ and $f^{(n)}$ respectively. Additionally, blink information when available is used to extract blink features as explained in Sections 3.3.1 and 3.4. Based on these feature

sets, fixation, saccade, and blink features based classifiers are constructed using RBFN and their accuracies are combined to get the final user prediction accuracy. RBFN was used as a preferred classifier because it produced higher prediction accuracy in our previous work [Schröder et al., 2020]. For all experiments in this chap-

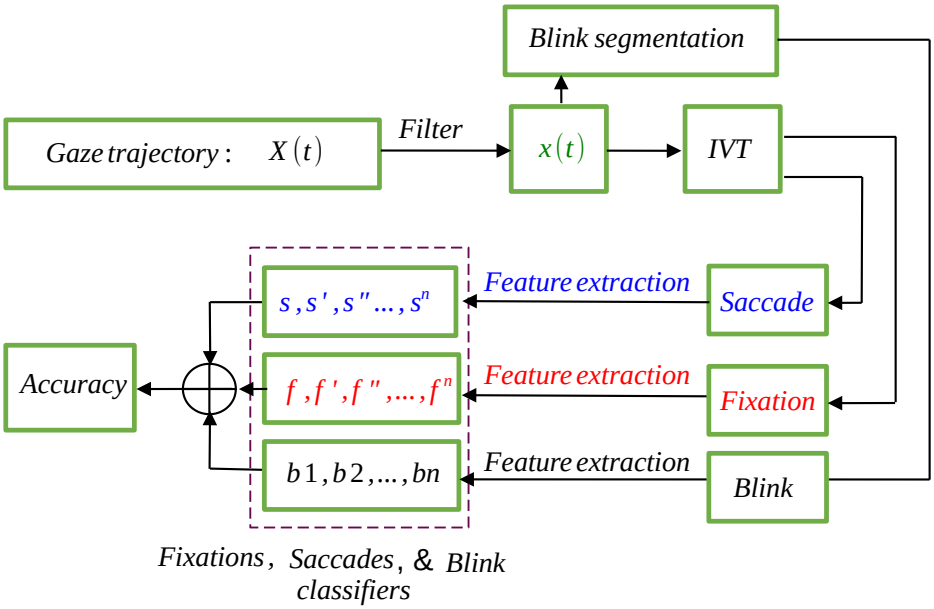


Figure 4.1: Overview of user prediction method

The structure of this chapter is as follows: Section 4.1 investigates the effect of stimuli (using the RAN, TEX, VST, and GOF datasets), Section 4.2 studies the effect of filter parameters, Section 4.3 studies the effect of IVT parameters (using the RAN, TEX, VST, and GOF datasets), Section 4.4 studies the effect of including higher-order derivatives of the gaze trajectory as features (using the RAN, TEX, VST, and GOF datasets), Section 4.5 investigates the effect of adding blink features (using the RAN, TEX, and VST datasets), Section 4.7 investigates the effect of the length of the trajectory (using GOF, VST, and MIT datasets), Section 4.6.1 investigates the effect of gender groups on user prediction (using GOF dataset), Section 4.6.2 investigates the effect of age on user prediction (using the GOF dataset), Section 4.8 investigates the effect of time gap between the train and test data (template aging using RAN, TEX, and VST datasets). Section 4.9 studies the effect of stimuli after homogenizing the datasets. Section 4.10 studies of combined factors.

4.1 Effect of Stimuli

Here, user identification was performed using four different datasets with different stimuli. The first dataset involves following a randomly moving dot (RAN), the second dataset stimulus consists of an explicit reading task (TEX), the third dataset involves an explicit visual searching task for number (VST), and the fourth dataset (GOF) involves a stimulus without an explicit task which captures the eye gaze on visual stimulus (face images). Hence, determining whether user identification is easier with one of these stimuli was the focus of this study. The default IVT parameters of VT = 50 °/s, MFD = 100 ms were used for the RAN, TEX, and VST datasets. For GOF VT = 15 °/s is applied, since the default parameters lead to zero fixations for some participants in this dataset. The first 51 features which are shown in Table 3.3 were used and an equal weighting of the saccade and fixation classifiers (RBFN) were used to compute the user identification probabilities. In the following, the user identification experiments are explained and the results are summarized in Table 4.1.

Table 4.1: Performance metrics with different datasets using 51 features over 50 runs.

Data set	Identification Accuracy	Number of participants	Trajectory Length [s]
RAN	92.62 ± 0.13 %	153	100
TEX	90.90 ± 0.10 %	153	60
VST	85.69 ± 0.16 %	58	180
GOF	77.91 ± 0.51 %	153	60

Bioeye TEX/RAN For both datasets, the ML classifier (RBFN) was trained with all the 153 participants of the second session and tested with the first session (train to test ratio 50:50 %). In our previous work [Schröder et al., 2020] an identification accuracy of 94.10 % using RAN data and 90.80 % using TEX data was achieved with **one run**. In this work, the average accuracy achieved over **50 runs** with the RAN dataset was 92.62 ± 0.13 % (maximum accuracy = 94.77 %) and with the TEX dataset was 90.90 ± 0.10 % (maximum accuracy = 92.28 %) which was a more stable prediction accuracy of our classifier compared to our previously reported results in [Schröder et al., 2020].

VST The ML classifier was trained with all the 58 participants using the training and testing sessions recorded on the same day. The accuracy achieved with this data

set was $85.69 \pm 0.16\,\%$ over 50 runs.

GOF This dataset has a large number of participants (193 males and 185 females in different age groups). The first 16 trials (first one minute of the combined trajectory of the first 16 trails) were used for training and the remaining 16 trials were used for testing. For better comparison with BioEye data, 153 participants (77 males and 76 females) are chosen randomly over 50 runs and the average accuracy was $77.91 \pm 0.51\,\%$. The prediction accuracy in GOF dataset is less than other datasets which maybe attributed to the nature of face viewing stimulus which does not trigger individual cognitive processes.

4.2 Effect of Savitzky-Golay Filter Parameters

Savitzky-Golay has been implemented to reduce the influence of noise. For every data point, it fits a symmetric polynomial through the point and a number of points in the neighborhood (frame). The filter has two parameters, polynomial order, and frame size. The frame size should be always an odd number and should be greater than the polynomial order parameter. The used parameters values in this thesis are frame size of 15 and polynomial order of 6 as mentioned previously. To study the effect of varying these parameters on the classification accuracy, three experiments were conducted (see Table 4.2): the first and second experiments were conducted by fixing one of the parameters and varying the other parameter by increasing the value with 2 steps. The results in both experiments showed that the highest accuracy is achieved with 6 for polynomial order and 15 for the frame size parameters. In the third experiment, both the parameters were varied by increments of 4 steps starting from deactivating the filter by using 0 for polynomial order and 1 for the frame size parameters. The results showed not much difference in the accuracy when the two parameters had values close to each other.

4.3 Effect of IVT Parameters

The IVT algorithm is a crucial component of the user prediction architecture. It segments the gaze trajectory into fixations and saccades (see Figure 4.1) which are later used to calculate different features in each category of the segments. As mentioned previously in Section 3.2, the IVT algorithm has two parameters namely the velocity threshold (VT) and the minimum fixation duration (MFD). In this section experiments were conducted in order to study the effect of changing the IVT parameters. A systematic parameter variation was conducted to determine

Table 4.2: Variation of polynomial order and the frame size parameters against the accuracy of user identification over 153 participants 50 runs in BioEye RAN dataset. (VT is fixed to 27 °/s)

Fixed frame size			Fixed Pol. order			Vary with 4 steps		
Pol. order	Frame size	Acc.%	Pol. order	Frame size	Acc.%	Pol. order	Frame size	Acc.%
0	1	91.56	0	15	90.04	**0**	**1**	**91.56**
6	7	91.54	2	15	93.35	4	5	91.63
6	9	93.97	4	15	94.26	8	9	91.57
6	11	94.97	**6**	**15**	**95.96**	12	13	91.54
6	13	94.72	8	15	94.72	16	17	91.56
6	**15**	**95.96**	10	15	93.93			
6	17	95.20	12	15	92.63			
6	19	95.05	14	15	91.58			
6	21	94.31	16	17	91.56			

which IVT parameters lead to the highest accuracy using the RBFN classifier (this classifier selected because it gives the higher accuracies).

An initial set of experiment was carried out with the RAN and TEX datasets. The experiments were conducted by varying the two IVT parameters, velocity threshold (VT) and minimum fixation duration (MFD), separately. In the first stage, varying the VT with a fixed MFD of 100 ms was performed and in the second stage fixing the VT (at the value with the highest accuracy from the previous stage) and variation of the MFD was performed. The parameter range under consideration was 10–100 °/s for VT and 50–150 ms for MFD. In each stage, first a broad variation is done in steps of 10, then a fine variation was executed in steps of 1. For each setting, a cross validation was performed with random 80 % subsets of the users (i.e. 122 participants) for a total of 50 runs. Figures 4.2a and 4.2b show the accuracy of user identification along with the number of fixations as the velocity threshold is varied. It can be noted that the highest accuracy was always obtained around the highest number of fixations. The best accuracy was achieved with a VT and MFD respectively of 26 °/s and 98 ms for TEX, 27 °/s and 96 ms for RAN. With MFD = 100 ms the obtained accuracies are almost identical (the difference was in the order of 0.05 % see Table 4.3).

The above approach requires a cross validation in the initial stages to compute the best parameters which is a time consuming process. Therefore, in an alternative approach, the VT parameter was tuned to achieve the highest number of fixations. This was done as follows. First prepare a plot such as Figure 4.3 and the zoomed plot Figure 4.4 (which is for the RAN dataset). The x-coordinate represents the number of fixations over all the participants in the data set and the y-coordinate shows

Table 4.3: Variation of MFD against the accuracy of user identification over 122 participants 50 runs in BioEye dataset. (VT is fixed to 27 °/s)

BioEye RAN			BioEye TEX		
Vel. threshold	MFD	Acc.%	Vel. threshold	MFD	Acc.%
27	0.05	94.62 ± 01.36 %	26	0.05	93.13 ± 1.37 %
27	0.06	94.63 ± 1.34 %	26	0.06	93.18 ± 1.35 %
27	0.07	95.03 ± 1.30 %	26	0.07	93.26 ± 1.37 %
27	0.08	95.73 ± 1.14 %	26	0.08	93.31 ± 1.12 %
27	0.09	95.57 ± 1.15 %	26	0.09	93.44 ± 1.56 %
27	**0.095**	**96.14 ± 1.07 %**	26	0.095	93.73 ± 1.38 %
27	**0.096**	**96.14 ± 1.07 %**	26	0.096	93.73 ± 1.38 %
27	0.097	96.09 ± 0.96 %	26	**0.097**	**94.18 ± 1.30 %**
27	0.098	96.09 ± 0.96 %	26	**0.098**	**94.18 ± 1.30 %**
27	0.099	96.09 ± 0.96 %	26	**0.099**	**94.18 ± 1.30 %**
27	0.1	96.09 ± 1.07 %	26	**0.1**	**94.18 ± 1.30 %**
27	0.101	96.04 ± 1.07 %	26	0.101	93.73 ± 1.45 %
27	0.102	96.04 ± 1.07 %	26	0.102	93.73 ± 1.45 %
27	0.103	96.04 ± 1.07 %	26	0.103	93.73 ± 1.45 %
27	0.104	96.04 ± 1.07 %	26	0.104	93.73 ± 1.45 %
27	0.105	95.98 ± 1.11 %	26	0.105	93.42 ± 1.51 %
27	0.11	95.09 ± 1.11 %	26	0.11	93.34 ± 1.73 %
27	0.12	94.70 ± 1.27 %	26	0.12	93.19 ± 1.43 %
27	0.13	93.83 ± 1.34 %	26	0.13	92.44 ± 1.38 %
27	0.14	94.13 ± 1.07 %	26	0.14	91.68 ± 1.49 %
27	0.15	93.72 ± 1.26 %	26	0.15	91.01 ± 1.49 %

the velocity thresholds. The colored stripes are a representation of the number of fixations of all the participants in the data set for each velocity threshold. The beginning of the line indicates the minimum number of fixations, while the end of the line indicates the maximum number of fixations on each specific velocity threshold. The blue dots on the colored stripes mark the mean fixation number. From the plot, the VT value where the number of fixations peaks is determined. Around this VT value, a local search is performed and the VT value which gives the highest accuracy is selected.

The above described procedure was performed and the following results for the four data sets using 51 features in fixation and saccade classifiers (see Table 4.4) were obtained:

RAN The highest number of fixations in RAN data set occurred with a velocity threshold of 24 °/s with accuracy of 94.89 %, while the best accuracy of 95.96 % was achieved with a VT of 27 °/s and MFD of 100 ms over 50 runs.

TEX The best accuracy of 93.23 % was achieved with a VT of 26 °/s and MFD of 100 ms, while the highest number of fixations occurred with a VT of 30 °/s with

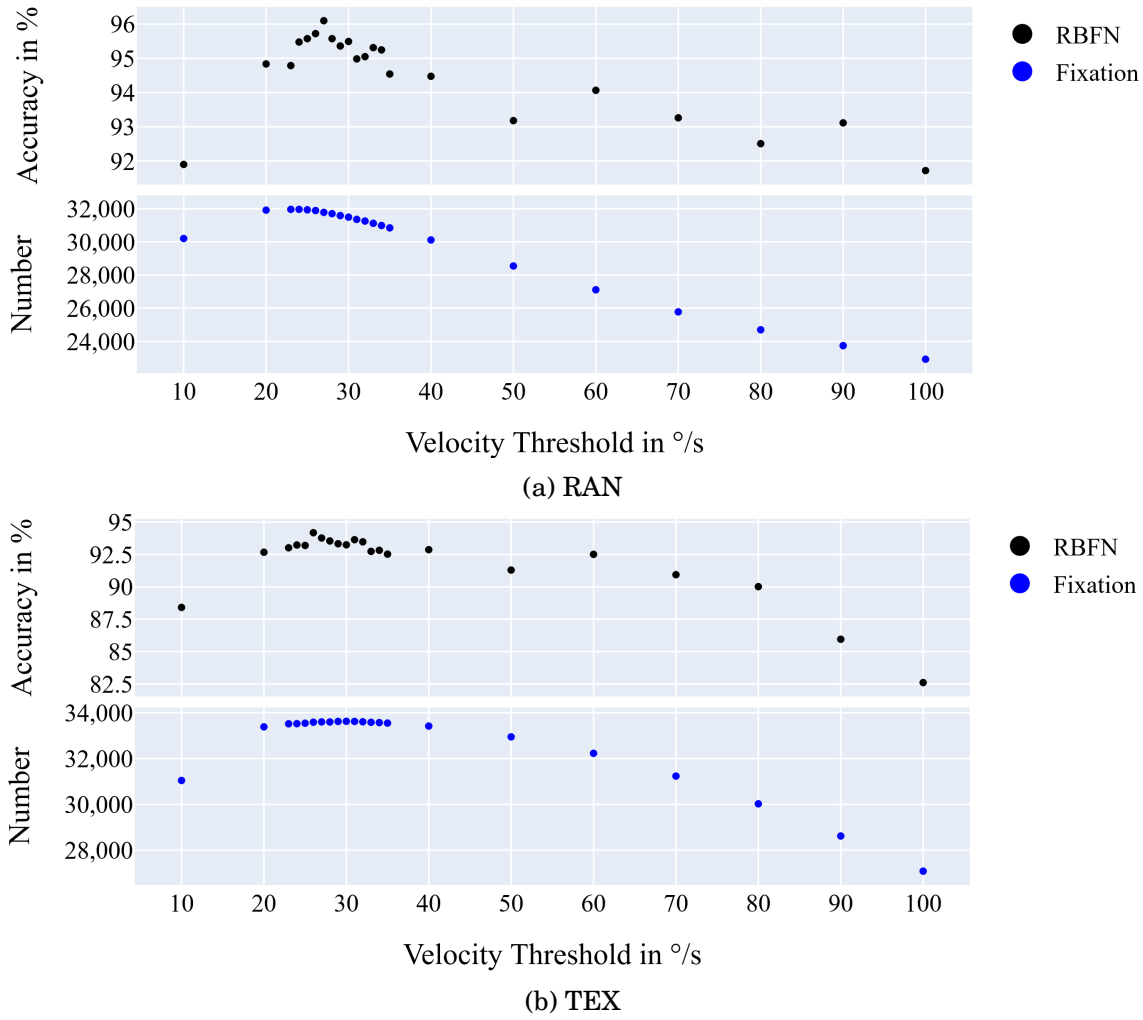


Figure 4.2: Variation of Velocity thresholds of BioEye Data with fixing MFD at 100 ms.

accuracy of 92.24 % using TEX data set.

VST The best accuracy of 94.82 % was attained with a VT of 100 °/s, MFD of 100 ms and the highest number of fixation occurred at a VT of 120 °/s with accuracy of 94.31 %.

GOF In this data set, the VT of 21 °/s produced the highest number of fixations with accuracy of 82.35 % and the best accuracy of 83.45 % was achieved with the VT of 22 °/s and MFD of 100 ms over 152 participants (76 males and 76 females) which were selected randomly over 50 runs from all the data users.

In all the above cases, the accuracies of the user identification were significantly increased (by 3.34 % for RAN, 2.33 % for TEX, 9.03 % for VST and 4.76 % for GOF) by selecting the optimal IVT parameters (compare with Table 4.1).

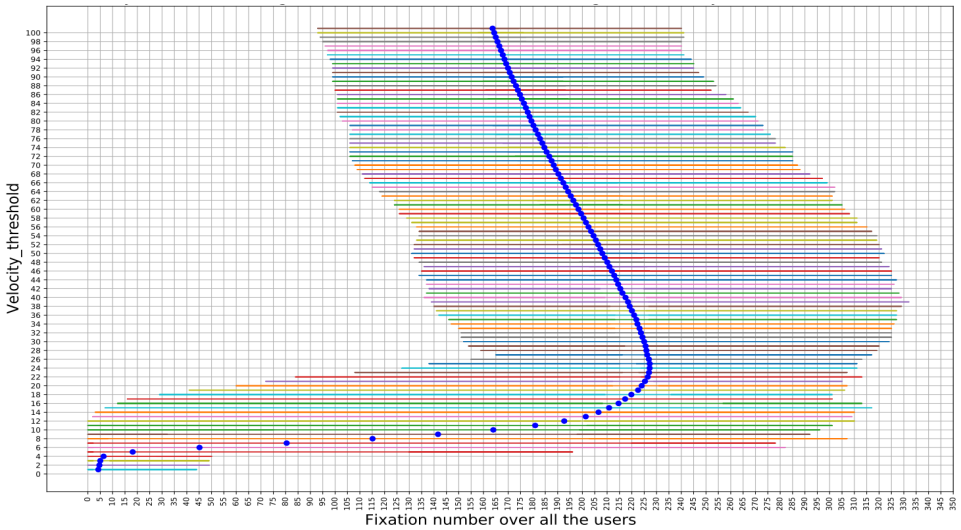


Figure 4.3: Velocity threshold of IVT against number of fixation for training session (RAN data set / 153 users)

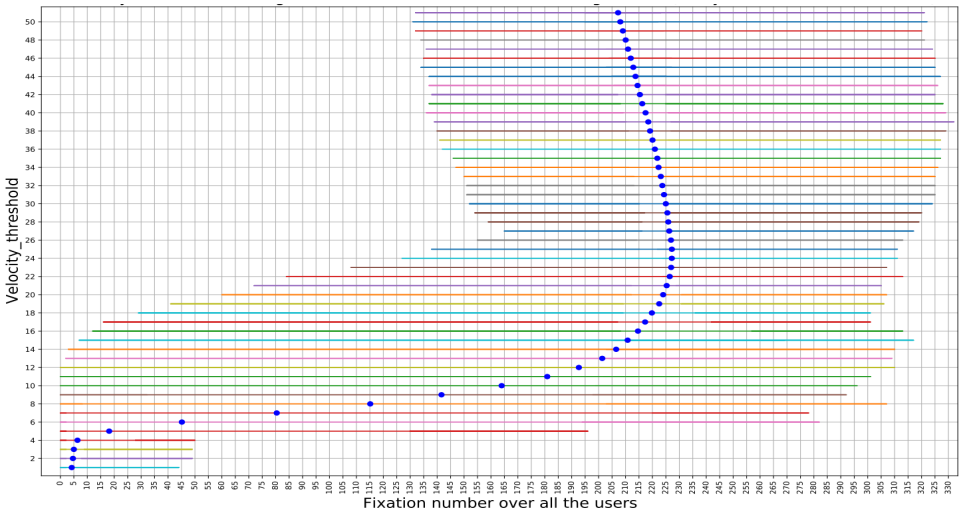


Figure 4.4: Velocity threshold of IVT against number of fixation for training session (RAN data set / 153 participants).

As one can see from Table 4.4, all the accuracies around the highest number of fixations were close to each other. Therefore, choosing any VT in this range will not significantly affect the final accuracy. From Table 4.4 an observation made was that in both approaches, the highest accuracy was around the highest number of fixations. Furthermore, these accuracies around the highest number of fixations were adjacent to each other. Therefore, the selection of any VT within the highest range would not alter the final accuracy to a great extent.

Table 4.4: The velocity threshold against the highest number of fixation over all the participants in session one, prediction accuracy of using 51 features and over 50 runs in all the data sets (except for VST, 10 runs were used in this experiment).

RAN			TEX		
Vel. threshold	Fix. No.	Acc. %	Vel. threshold	Fix. No.	Acc. %
23	31959	94.27	23	33515	92.25
24	**31961**	94.89	24	33520	92.48
25	31933	94.93	25	33540	92.65
26	31889	94.86	**26**	33584	**93.23**
27	31779	**95.96**	27	33597	92.65
28	31700	95.03	28	33594	92.56
29	31582	94.91	29	33617	92.31
30	31490	95.20	**30**	**33623**	92.24
31	31355	94.56	31	33617	92.86
32	31255	94.71	32	33602	92.84
VST data			GOF data		
Vel. threshold	Fix. No.	Acc. %	Vel. threshold	Fix. No.	Acc. %
80	24987	93.10	16	52514	80.51
90	27883	93.27	17	54289	80.83
100	29818	**94.82**	18	55399	81.56
110	30715	94.80	19	56065	82.19
120	**31021**	94.31	20	56408	82.51
130	30684	92.84	**21**	**56490**	82.35
140	30077	94.48	**22**	56391	**83.45**
150	29180	94.30	23	56079	83.25
160	28307	94.27	24	55701	83.40
170	27419	91.72	25	55268	83.23

4.4 Effect of Higher-Order Derivatives

The features that were used in the user identification experiments were inspired by the work of George and Routray [George and Routray, 2016]. They used 9 fixation features and 43 saccade features. In our earlier work [Schröder et al., 2020] as well as the next work [Zaidawi et al., 2022], 51 features were used in fixation as well as the saccade classifier. These include statistical information (e.g., mean, median, maximum, standard deviation (STD), skewness, kurtosis - see [George and Routray, 2016]) from the position, velocity, and acceleration of every saccade and fixation, as described in Section 3.3.1 and as seen in Table 3.3. However, the commonly adopted choice of including features until the second time derivative of the gaze trajectory seems to be arbitrary. Higher order derivatives of gaze trajectories may still encode useful information that can improve the user prediction accuracy.

Humans have prediction capabilities in their perception-action loop. This predictive behavior is usually captured in higher order derivatives of the perception or action trajectories [Sargolzaei et al., 2016]. In eye tracking research, fourth order derivatives were exploited to predict saccade movements [Wang et al., 2017]. Hence, the effects of including higher order derivatives of the gaze trajectory as features e.g. jerk, jounce, and crackle on the user identification accuracy were studied in this thesis. The hypothesis was that for a given dataset, inclusion of higher-order derivatives until a certain order should increase the accuracy after which it may become stagnant or start to decrease due to dominating noise characteristics [Gibaldi and Sabatini, 2021] which commonly occurs while computing the higher order derivatives of the gaze trajectory signal. In this way, certain important features that might be hidden in higher-order derivatives can be reliably captured.

In order to study the effect of higher order derivative features on the accuracy of user identification, the start was with using a minimal set of first 13 position based features and duration as a general feature. See the first 14 features in Table 3.3. Next was adding 19 velocity based features and then, step by step, including 18 statistical features based on each of the other higher order derivatives (acceleration, jerk, jounce, crackle). For this study, the default parameters for the Savitzky Golay filter (window length = 15, polynomial order = 6) were used for all the considered datasets. For RAN, TEX and VST, the IVT VT parameter was chosen as 50 °/s and for the GOF dataset, it is 15 °/s since the default parameters (VT = 50 °/s and MFD = 100 ms) lead to zero fixations for some participants in this dataset. The results reported in Table 4.5 show the accuracy with an increasing number of higher order derivative features for the four different datasets. The accuracy of user identification increases until the inclusion of jounce for the RAN, TEX and GOF datasets. For the crackle based features, the accuracy decreases again. The accuracies of user identification (with our default 51 features) were increased for the RAN dataset from $92.62\,\%$ to $94.58\,\%$, for the TEX dataset from $90.90\,\%$ to $91.96\,\%$, and for the GOF dataset from $77.91\,\%$ to $81.02\,\%$ by including until jounce level features. For the VST dataset, the accuracy rises only until the jerk level features and decreases already for jounce (from $85.69\,\%$ to $86.52\,\%$).

Figures 4.5a and 4.5b show the higher order derivatives of the gaze trajectories of the RAN and VST datasets. From the plots of the RAN dataset, it can be observed that X and Y trajectory components look qualitatively different until the jounce level (see the nature and occurrence of peaks in X and Y components for example) while jounce and crackle look very similar due to amplified noise content. Similar plots were obtained for GOF and TEX datasets (see Figure 4.5c and Figure 4.5d). However, for the VST dataset, the intense noise dominates much earlier i.e. trajectories after jerk look qualitatively similar due to high noise content and hence may not bear any

Table 4.5: Performance metrics over 50 runs with varying number of features of higher order derivatives of the gaze trajectory of all the datasets.

Deriv. order	Feat. No.	RAN Acc.% (153 participants, VT=50 °/s, MFD= 100 ms)	TEX Acc.% (153 participants, VT=50 °/s, MFD= 100 ms	VST Acc.% (58 participants, VT=50 °/s, MFD= 100 ms	GOF Acc.% (153 participants, VT=15 °/s, MFD= 100 ms
0. Position	14	83.02 ± 0.22	84.30 ± 0.17	84.79 ± 0.22	46.82 ± 0.44
1. Velocity	33	89.67 ± 0.15	89.81 ± 0.11	85.03 ± 0.15	69.46 ± 0.54
2. Acceleration	51	92.62 ± 0.13	90.90 ± 0.10	85.69 ± 0.16	77.91 ± 0.51
3. Jerk	69	93.54 ± 0.10	91.73 ± 0.12	$\mathbf{86.52 \pm 0.23}$	80.08 ± 0.50
4. Jounce	87	$\mathbf{94.58 \pm 0.10}$	$\mathbf{91.96 \pm 0.08}$	85.13 ± 0.17	$\mathbf{81.02 \pm 0.53}$
5. Crackle	105	94.15 ± 0.09	90.86 ± 0.14	83.82 ± 0.14	80.54 ± 0.54

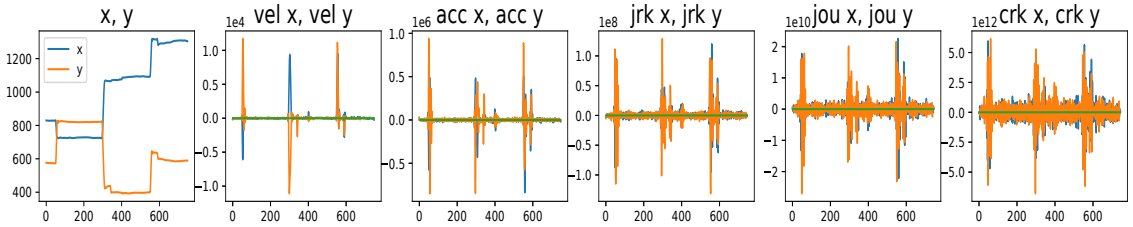
meaningful features. It was to be noted that the magnitude of X and Y trajectory components in these plots was not relevant as we normalize the features as explained in Section 3.3.

The shown observations attest our hypothesis that including higher order derivatives (up to a certain level) is indeed a useful way to capture meaningful information that contribute to an increase in the accuracy of user identification.

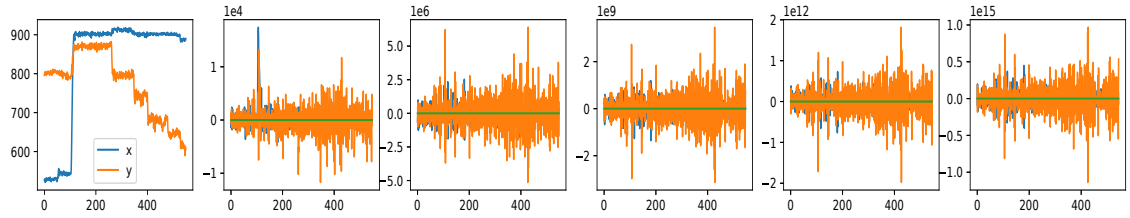
4.5 Effect of Blinking Features

As explained in Sections 3.2 and 5.2.2, the eye movement data may have invalid data (NaNs) or outliers which were considered as a source for blink information. This can be due to user-specific reasons such as blinking, loss of attention (micro-sleeping), or eye tracker faults (e.g. solo missed gaze points) [Rigas and Komogortsev, 2017]. The mean of gaze points across the invalid segments (either explicitly labeled in the dataset or segments containing NaNs) was used for interpolation in order to have a connected gaze trajectory. However, the blinks were the majority of the outliers. A loss of attention (micro-sleeping) is a temporary episode of sleep or drowsiness which may last for a fraction of a second or up to 30 seconds where an individual fails to respond to some arbitrary sensory input. Physiologically, the blinking behavior can encode some information about the participants [Kröger et al., 2020]. Actual blinking rates vary by individual averaging around 10 blinks per minute and the duration of a blink is on average between 100–400 ms according to the Harvard Database of Useful Biological Numbers[1] [Ramot, 2001, Taschenbuch Verlag Schiffman, 2001]. Blinking behavior is arguably different in men and women [Doughty, 2002] and it was found

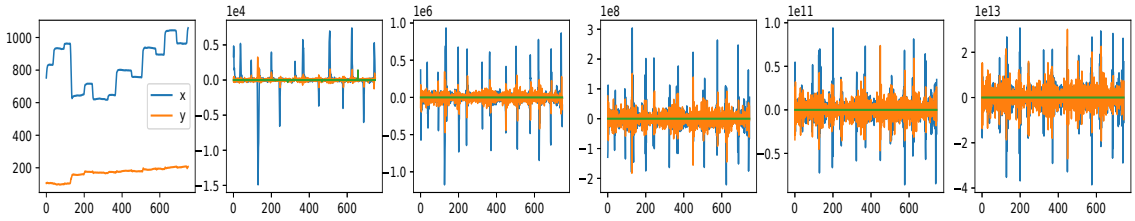
[1] `https://bionumbers.hms.harvard.edu/` [accessed 11-August-2021]



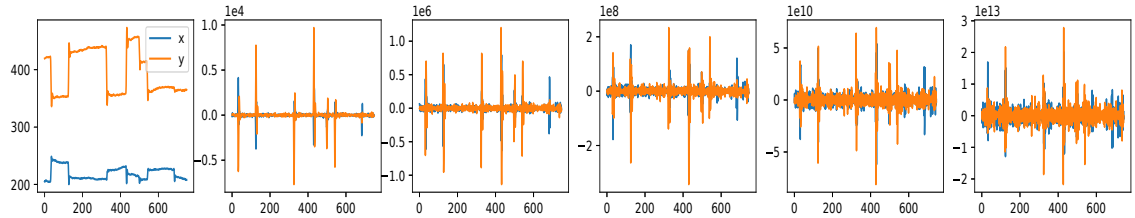
(a) RAN dataset



(b) VST dataset



(c) TEX dataset



(d) GOF dataset

Figure 4.5: Higher order derivatives of filtered XY gaze trajectory in different datasets

that adults blink more often than infants [Juan, 2006].

The effect of including a blink classifier was studied in this section, as introduced in Section 3.4 in addition to fixation and saccade classifiers. The blink segments were extracted from the explicitly labeled invalid data in the case of RAN and TEX and NaN segments in case of VST (see Section 3.2 for more details). The features that were explained in Section 3.3.1 and listed in 3.4 were used for the blink classifier (Figures4.7 and Figure 4.8 are examples of the number of blink over all the users in train and test sessions of the VST dataset). Since the identification accuracy of the blink classifier was lower than the fixation and saccade classifiers, weighing all of the three classifiers equally to compute the final accuracy was not giving good results.

In order to find the optimal weights for the three classifiers, first, the traditional way of performing hyper-parameter optimization was conducted i.e. performing a grid search. A grid search algorithm guided by the performance metric (find the maximum accuracy was used in this study), which is measured by cross-validation on the training dataset was carried out. An extensive search for the best weights was performed through 200 steps between 0 and 1. The sum of the three weights should be equal to one. The blink weight ($w3$) is equal to $1 - w1 - w2$, where $w1$ is saccade weight and $w2$ is fixation weight. Figure 4.6 shows an example of using global search with the VST dataset. As can be seen from the figure the highest accuracies occur between the weights range of 0.6 and 0.3 for the fixation and saccade classifiers. This way of searching is computationally expensive, so to maximize the objective function (i.e. prediction accuracy) and reduce the calculation time, the Nelder-Mead Method [Nelder and Mead, 1965] was used which is a popular direct search method (based on function comparison [Gao and Han, 2012]) and is suited for optimization problems for which derivatives may not be known and gradient based optimization algorithms are not a viable alternative.

As shown in Table 4.6, the accuracy increased by $1.35\,\%$, $1.25\,\%$, and $0.5\,\%$ through the use of the blink classifier in VST, RAN, and TEX respectively. It is obvious that the trajectory length has a high impact on the blink number and the accuracy of the blink classifier. The blink classifier accuracy was $3.92\,\%$ for TEX (trajectory length = 60 s), $9.15\,\%$ for RAN (trajectory length = 100 s), and $13.79\,\%$ for VST (trajectory length = 180 s) datasets.

Finally, other experiments were conducted with optimizing the weights of only the fixation and saccade classifiers to study the importance of these two classifiers before including the blink classifier. This is to verify whether fixation and saccade classifiers indeed contribute equally to the final classification accuracy. As can be seen from the results in Table 4.6 optimizing the weights did not bring advantage for the user identification accuracy. The accuracy increased only by $0.01\,\%$ for RAN, $0.08\,\%$ for TEX, and $0.20\,\%$ for VST. As observed from these results that both of the fixation and saccade classifiers are equally important for the user identification accuracy.

4.6 Effect of Gender and Age Groups on User Identification using the GOF data

The GOF dataset has a large number of participants with gender and age information. This demographic (see Table 4.7) provided the motivation to study the effect of gender and age on the accuracy of user identification.

Table 4.6: Performance metrics over 50 runs using blink classifier in **RAN, TEX, and VST datasets (VT = 50 °/s and MFD = 100 ms).**

Dataset	Sac/Fix/blink features	CLF weights Sac/Fix/Blink	Identification Accuracy
RAN	51/51/0	0.5/0.5/0.0	92.62 ± 0.13 %
RAN	51/51/0	0.525/0.475/0.0	92.63 ± 0.13 %
RAN	51/51/7	0.408/0.578/0.015	93.87 ± 0.12 %
TEX	51/51/0	0.5/0.5/0.0	90.90 ± 0.10 %
TEX	51/51/0	0.506/0.494/0.0	90.98 ± 0.11 %
TEX	51/51/7	0.4453/0.5453/0.0094	91.40 ± 0.10 %
VST	51/51/0	0.5/0.5/0.0	85.69 ± 0.16 %
VST	51/51/0	0.525/0.475/0.0	85.89 ± 0.20 %
VST	51/51/7	0.568/0.3938/0.0381	87.04 ± 0.24 %

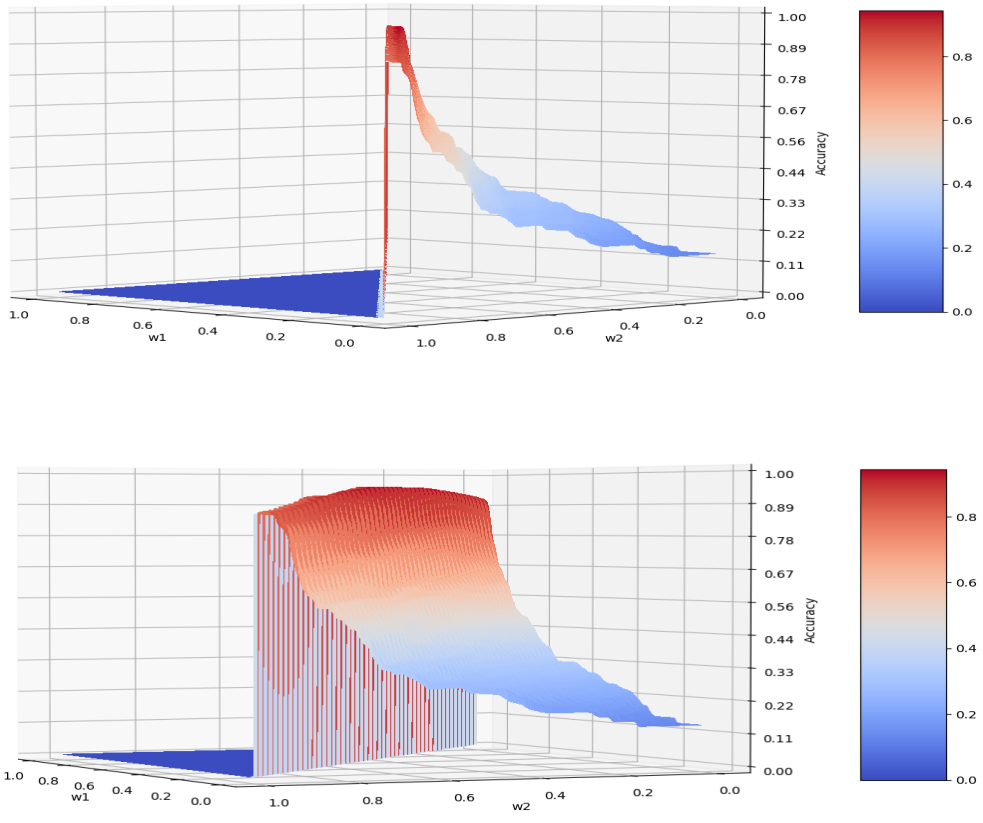


Figure 4.6: Weights optimization VST dataset (top) and (lower) showing the accuracy plot from different angles

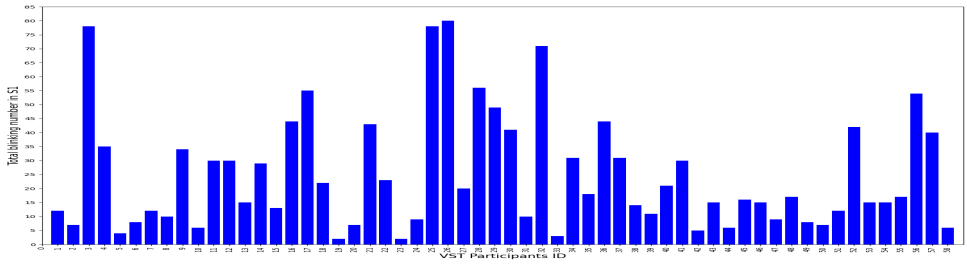


Figure 4.7: Number of blinks over each participant in S1 of VST dataset

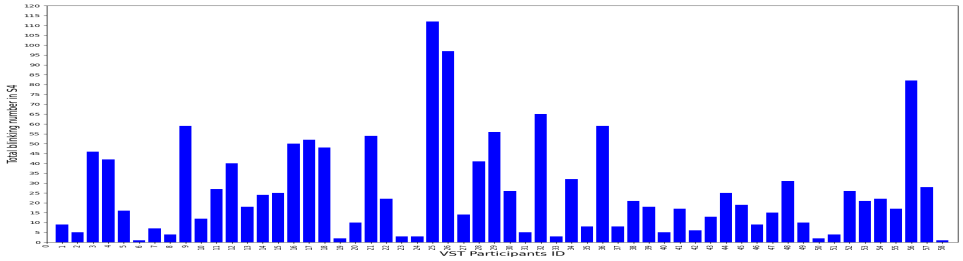


Figure 4.8: Number of blinks over each participant in S4 of VST dataset

Table 4.7: Demographics of Males and Females in GOF dataset.

Age Group	Males	Females	Total
20 – 72	193	185	378
20 – 40	151	157	308
41 – 72	42	28	70

4.6.1 Gender

To study the effect of gender, user identification experiments were performed in three different groups with each 150 participants: a balanced group with 75 Males and 75 Females, a group exclusively with male participants and a group exclusively with female participants. The participants of the groups were randomly chosen from the complete dataset 50 times each. The average of the user identification accuracies reported in Table 4.8. For the mixed group, the accuracy achieved was $83.25 \pm 0.48\,\%$. In isolated groups of males and females, the accuracy was found to be higher in the female group ($88.85 \pm 0.32\,\%$) when compared to the male one ($77.37 \pm 0.50\,\%$). These results show that user identification is biased towards gender and was found to be more accurate in females than males. We speculate that this is due to the fact that females have more exploratory behavior than males which provides richer information to predict their identity.

4.6.2 Age

To study the effect of age, the participants were split into two age groups 20–40 years and 41–72 years. Both age groups contain 56 participants with an equal number of males and females. The user identification can be performed in the older age group ($91.43 \pm 0.47\,\%$) with a five percentage points higher accuracy in comparison to the younger age group ($85.96 \pm 0.79\,\%$) see Table 4.8. We speculate that this effect might be due to decrease in neuroplasticity with increasing age which might make the eye movement behavior in older age group more indicative of their individuality than younger age group.

Table 4.8: Effect of gender and age groups on average user identification accuracy over 50 runs with VT = 22 using 51 features.

Age group	Number of participants	M	F	Identification Accuracy
20 – 72	150	75	75	$83.25 \pm 0.48\,\%$
20 – 72	150	150	0	$77.37 \pm 0.50\,\%$
20 – 72	150	0	150	$88.85 \pm 0.32\,\%$
20 – 72	56	28	28	$87.36 \pm 0.65\,\%$
20 – 40	56	28	28	$85.96 \pm 0.79\,\%$
41 – 72	56	28	28	$91.43 \pm 0.47\,\%$

4.7 Effect of Length of Gaze Trajectory and Fatigue

For real-world applications of gaze biometrics, it is relevant to estimate the amount of training data that is needed per user for reliable user identification. As explained later in the gender prediction experiments in Chapter 5, it is important to investigate the effect of trajectory length and how that will influence the prediction accuracy (in this chapter for user identification task) as it is important to know the minimum trajectory length that is sufficient for user identification with high accuracy. The MIT, VST and RAN datasets provide longer gaze trajectories of more than 50 minutes, 4 minutes and 1 min 40 seconds respectively in comparison to 1-minute trajectory length in both GOF and TEX datasets. Hence, the MIT, VST and RAN datasets were used here to study the effect of length of gaze trajectories. Furthermore, when shorter trajectory lengths were considered, we select them from both the start and the end of the full trajectory to additionally study the effect of possible fatigue.

4.7.1 MIT dataset

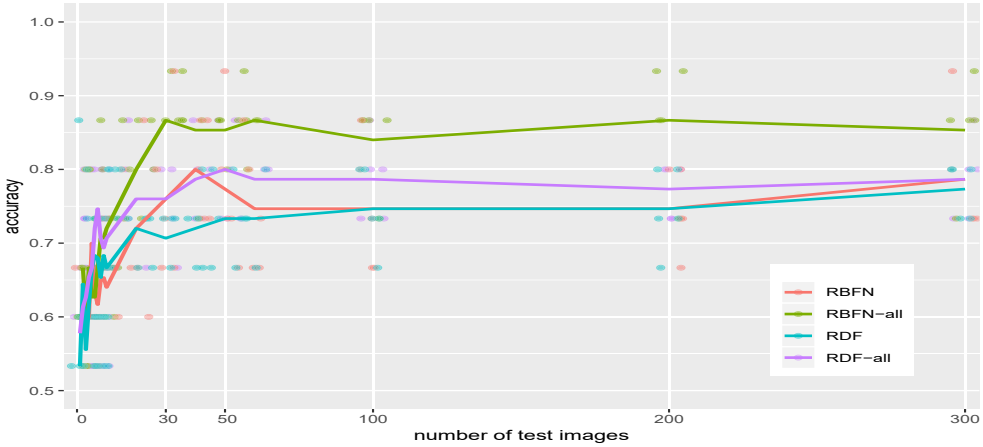
In this section, experiments were conducted to analyze how the user identification accuracy depends on the amount of training data as well as the amount of testing data using MIT dataset [Schröder et al., 2020]. This data provides three seconds trajectory length per image ("let each image stimuli be denoted as a sample"), each participant viewed 1003 images. To study the effect of trajectory length on the accuracy, a different amount of data were used for both training and testing. The final accuracy was computed from averaging the prediction accuracy over 5 runs. The following is how these experiments were conducted:

First, a fixed number of testing samples per participant of 30 (90 seconds of continuous trajectory) was used. The number of training samples was varied up to 700 images. Second, the opposite was implemented, by fixing the number of training samples per participant to 30 samples. Then number of test samples was varied up to 300 images. The Radial Base Function Networks (RBFN) and Random Decision Forest (RDF) classifiers were used with two sets of features: the features which were used in [George and Routray, 2016] (RBFN and RDF: 43 for fixation and 9 for saccade) and our set of features (RBFN-all and RDF-all: 51 for fixation and 51 for saccade). In these experiments, the hyper-parameters were the K for the RBFN classifier i.e. $K = n \quad \forall \quad n < 20$ samples (n being the number of samples) and $K = 32$ for 20 samples and more. RDF depends on several parameters, such as the number of trees, the number of features considered at each split, the maximum tree depth, and others. Preliminary experiments found that most of the default parameters from the *scikit-learn* software package [Pedregosa et al., 2011] work very well. No limit to the maximum tree depth, at least two samples per split, consider $\sqrt{F}$ features at each split, where F is the length of the feature vector. The only parameter chosen was the number of trees, which was set to 400 for all our experiments.

As shown in Figure 4.9, when the number of training samples was varied with fixing the number of samples for testing to 30, the accuracies were increases in general (see Figure 4.9a). RBFN with all features performed best using 30 training images with average accuracy and standard derivation of $86.67\%(4.71\%)$ over the 5 runs that randomly selected of sample subsets. While with 300 training samples, the RDF, both with the full feature (RDF-all) set as well as with the smaller subset (RDF), start to outperform both RBFN methods.

Similarly to the previous experiment, when increasing numbers of samples were used for testing and the number of training samples were fixed to 30, the performance of all the classifiers with the two sets of features increased (see Figure 4.9b). For all classifiers (RBFN-all, RDF, RDF-all) except RBFN with the limited features (RBFN), the performance does not increase with more than 40 test samples.

(a) Fixed testing samples



(b) Fixed training samples

Figure 4.9: Accuracy of different methods, by number of training images (a) and by the number of testing images (b). The number of testing or training images is fixed to 30, respectively. The lines mark the method's mean values.

Therefore, it can be noticed that while the RBFN methods (RBFN and RBFB-all) work well with fewer training samples, their performance was worse than the RDF methods (RDF and RDF-all) with more than 200 training samples. The results of these experiments suggested that 120 seconds (40 samples) of trajectory data is sufficient to identity participants with the maximum possible performance of these methods.

These results represent a representative subset of our experiments. While we only show plots for 30 training and testing samples, also other combinations were analyzed. The general form of the curves is the same – only the accuracy increase with more training samples.

Finally, in contrast to [George and Routray, 2016], the results showed that the performance of both RBFN and RDF with the complete features set (RBFN-all and RDF-all in Figure 4.9) were actually better than with the restricted feature set (RBFN and RDF) used in [George and Routray, 2016].

4.7.2 VST dataset

VST data provided longer gaze trajectories in comparison with RAN, TEX, and GOF datasets (explained also in Section 5.2.2). The experiments of user prediction using this data were performed with different trajectory lengths for both train and test sessions (i.e. 12, 20, 40, 60, 120, 180, and 240 seconds). Also, these sub-trajectories have been taken from the start and the end of the recording. The parameters for IVT algorithm used for this dataset were VT = 150 and MFD = 100 ms (using all the 58 users over 300 runs). RBFN classifier was used with two sets of features: the features set used in [George and Routray, 2016] (RBFN: 43 for fixation and 9 for saccade) and our set of features (RBFN-all: 51 for fixation and 51 for saccade). Session 1 was chosen for training and session 4 was used for testing. Both of train and test sessions were recorded on the same day. All the results of the experiments were reported in Table 4.9 and Figure 4.10 plots the accuracies of user identification with RBFN and RBFN-all classifiers when different trajectory lengths were considered from the beginning and end of the full trajectories.

The accuracies were about 50 % until using 40 seconds trajectory length and started to increase as the trajectory length was increased. The best accuracy achieved here was 94.6 % with 180 seconds from the end of the trajectory length and with longer trajectory (i.e. 240 seconds) the accuracy does not improve any further. The best accuracies have always been achieved when the 51 feature set (RBFN-all) was used. them

In order to study the effect of fatigue or losing attention for the user prediction task as implemented in Section 5.2.2, all the above experiments were performed with taking all the different trajectory lengths (i.e. 12, 20, 40, 60, 120, and 180 seconds) from the end of the recording. It is important to note that better accuracies were achieved when the sub-trajectories were taken from the end of the recording in most cases (see Table 4.9). This hints that predicting the user identity can be more accurate when the user is losing the attention or start to have fatigue especially when engaged in cognitively demanding task like visual searching of numbers task. This increase in accuracy can be due to the increase in blink at the end of the trajectory which is a sign of fatigue [Benedetto et al., 2011].

All the experiments in this section were repeated with removing the invalid data (NaNs) from the data to study the importance of this data. The results (see Ta-

ble 4.10) are showing a decrease in the accuracies which confirms that these parts of the data (blinking) can encode some information about the participants as reported in the study of Jacob Leon [Kröger et al., 2020]. The best accuracy was decreased by $5,1\%$, from $94.6\,\%$ in Table 4.9 to $89.5\,\%$ in Table 4.10.

Finally, as mentioned in Section 3.4, RBFN K (hyper parameter of RBFN) value of 32 was used for all the user identification experiments with trajectory length of one minute or more. With less than one minute trajectory length, re-tuning of the K parameter was performed using grid search method on the train data. Figure 4.11 shows an example of the RBFN K parameter tuning in user identification task with 40 seconds of the trajectory length in VST data set. The best value for the RBFN K parameter in Figure 4.11 was 10, as can be seen this value gave the highest test accuracy where the smallest difference between the test accuracy and it's train accuracy was achieved. Therefore, the classifiers hyper-parameters values that were used in each trajectory segment: RBFN "K" value of 1 with 12 sec, RBFN "K" value of 2 with 20 sec, and RBFN "K" value of 10 with 40 sec.

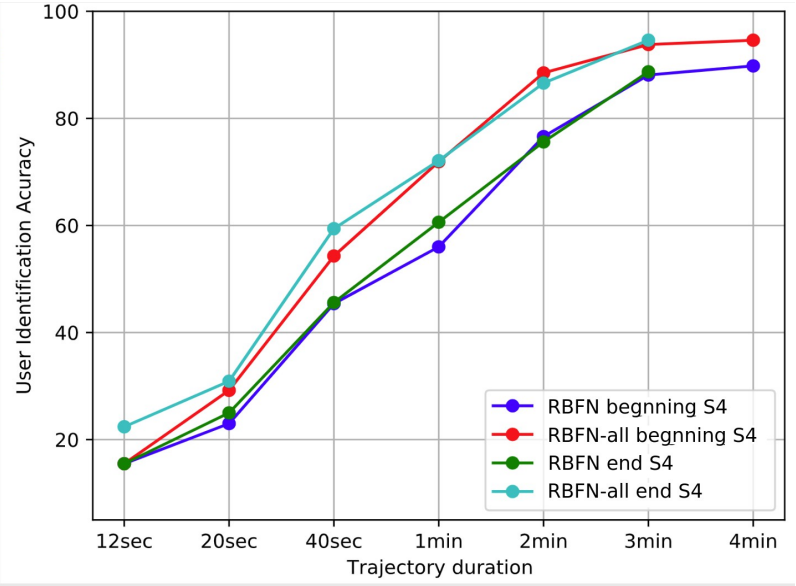


Figure 4.10: Train with S1 and Test with S4 after 2 hours gap

4.7.3 RAN dataset

For this dataset, the considered trajectory lengths were 60 seconds (from the start and the end) and 100 seconds respectively. The achieved accuracies for 60 seconds of trajectory length from the start and the end are $84.43\,\%$ and $82.37\,\%$ respectively. While for 100 seconds of trajectory length, the accuracy was $92.62\,\%$ as reported pre-

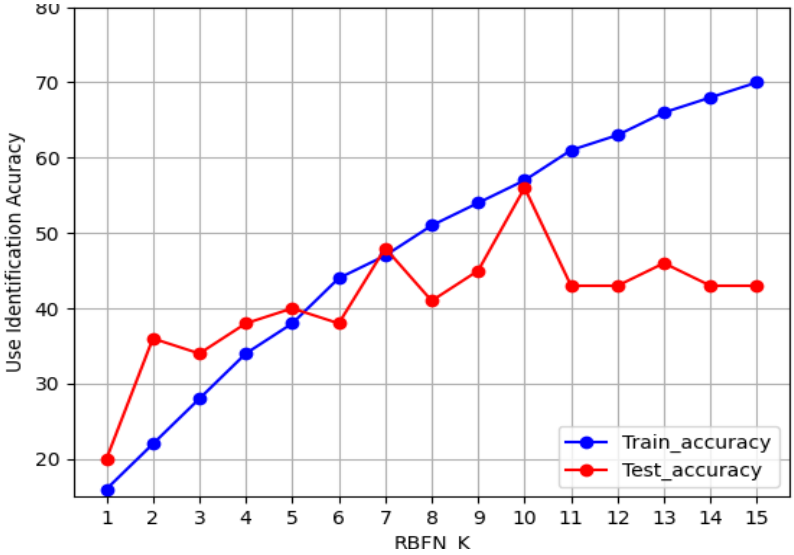


Figure 4.11: RBFN K parameter against the accuracy as examples of tune K parameter in VST dataset (40 sec trajectory length)

viously in Table 4.1. For these experiments, the same features and IVT parameters reported in Section 4.1 were used. Similar to the VST dataset, the accuracy increases with increase in trajectory length, however, the accuracy does not increase when shorter trajectories were selected from the end. This was the case because of the nature of non cognitively demanding stimulus of randomly moving dot in the RAN dataset. With the shorter trajectory length of 60 s, the accuracy dropped from 92.62 % to 84.43 %. For more details see Section 4.9 and Table 4.13. Lastly, RBFN "K" value of 32 was used in this section work.

4.8 Effect of the Time Gap Between Train and Test Data (Template Aging)

In this section a study of the effect of the time gap between train and test data on the user identification accuracy using the RAN, TEX, and VST data sets is conducted. The effect was also referred to as template aging [George and Routray, 2016]. The following Sections explain this study in each used dataset.

4.8.1 BioEye datasets

Out of 153 participants, the RAN and TEX datasets have 37 participants for which recordings were also available after one year. This can be used to study the template aging effect in eye movement biometrics. The first experiment in this section was conducted with these 37 participants in both train and test sessions where the

Table 4.9: Accuracy and standard derivation over 50 seeds of different trajectory lengths using VST data (with keeping peaks data)

CLF	Train/Test Session	Trajectory length (sec.)	Accuracy Start %	Accuracy End %
RBFN	S1/S4	12	15.5 ± 0.0	15.5 ± 0.0
RBFN-all	S1/S4	12	15.5 ± 0.0	22.4 ± 0.0
RBFN	S1/S5	12	12.1 ± 0.0	15.5 ± 0
RBFN-all	S1/S5	12	17.2 ± 0.0	15.5 ± 0.0
RBFN	S1/S4	20	23.0 ± 0.2	25.0 ± 0.3
RBFN-all	S1/S4	20	29.2 ± 0.3	30.9 ± 0.3
RBFN	S1/S5	20	15.2 ± 0.3	12.5 ± 0.3
RBFN-all	S1/S5	20	21.0 ± 0.2	25.6 ± 0.3
RBFN	S1/S4	40	45.4 ± 0.5	45.6 ± 0.5
RBFN-all	S1/S4	40	54.3 ± 0.5	59.4 ± 0.5
RBFN	S1/S5	40	25.3 ± 0.5	24.3 ± 0.3
RBFN-all	S1/S5	40	39.0 ± 0.4	43.3 ± 0.3
RBFN	S1/S4	60	56.0 ± 0.5	60 ± 0.4
RBFN-all	S1/S4	60	71.9 ± 0.4	72.1 ± 0.3
RBFN	S1/S5	60	36.6 ± 0.3	38.7 ± 0.4
RBFN-all	S1/S5	60	53.8 ± 0.4	55.0 ± 0.4
RBFN	S1/S4	120	76.6 ± 0.3	75.6 ± 0.3
RBFN-all	S1/S4	120	88.5 ± 0.3	86.6 ± 0.2
RBFN	S1/S5	120	59.7 ± 0.3	53.8 ± 0.5
RBFN-all	S1/S5	120	70.0 ± 0.3	66.3 ± 0.3
RBFN	S1/S4	180	88.1 ± 0.2	88.7 ± 0.3
RBFN-all	S1/S4	180	93.8 ± 0.2	$\mathbf{94.6} \pm 0.1$
RBFN	S1/S5	180	70.8 ± 0.3	67.5 ± 0.3
RBFN-all	S1/S5	180	74.7 ± 0.3	76.0 ± 0.2
RBFN	S1/S4	240	89.8 ± 0.1	-
RBFN-all	S1/S4	240	$\mathbf{94.6} \pm 0.1$	-
RBFN	S1/S5	240	72.9 ± 0.3	-
RBFN-all	S1/S5	240	79.8 ± 0.3	-

gap between the sessions is one year. The default IVT parameters were used in this study i.e. VT = 50 and MFD = 100 ms. For the RAN dataset, the average accuracy over **50 runs** was $83.51 \pm 0.18\,\%$ (maximum accuracy is 86.48 %) and with TEX the accuracy was $75.24 \pm 0.32\,\%$ (maximum accuracy is 81.08 %). In comparison with the work of George and Routray [George and Routray, 2016] running the same experiment yielded $81.08\,\%$ and $78.38\,\%$ with **one run** for RAN and TEX respectively. The accuracy has improved by $5.40\,\%$ with using RAN dataset and $2.70\,\%$ with using TEX dataset with **one run** and after using our 51 feature set.

The second experiment with this dataset was by training with all 153 users in session 2 and testing with a session recorded after one year gap which has only 37

Table 4.10: Accuracy and standard derivation over 50 seeds of different trajectory lengths using VST data (without keeping peaks data)

CLF	Train/Test Session	Trajectory length (sec.)	Accuracy Start %	Accuracy End %
RBFN	S1/S4	12	12.1 ± 0.0	20.7 ± 0.0
RBFN-all	S1/S4	12	17.2 ± 0.0	20.7 ± 0.0
RBFN	S1/S5	12	13.8 ± 0.0	13.8 ± 0.0
RBFN-all	S1/S5	12	13.8 ± 0.0	20.7 ± 0.0
RBFN	S1/S4	20	21.2 ± 0.2	26.0 ± 0.2
RBFN-all	S1/S4	20	29.6 ± 0.3	34.7 ± 0.3
RBFN	S1/S5	20	16.2 ± 0.2	18.8 ± 0.2
RBFN-all	S1/S5	20	21.6 ± 0.3	26.1 ± 0.3
RBFN	S1/S4	40	42.4 ± 0.4	41.9 ± 0.4
RBFN-all	S1/S4	40	50.9 ± 0.3	54.3 ± 0.5
RBFN	S1/S5	40	24.1 ± 0.4	27.6 ± 0.4
RBFN-all	S1/S5	40	35.8 ± 0.4	43.8 ± 0.4
RBFN	S1/S4	60	56.0 ± 0.3	57.9 ± 0.5
RBFN-all	S1/S4	60	67.6 ± 0.3	69.2 ± 0.3
RBFN	S1/S5	60	35.1 ± 0.5	38.9 ± 0.4
RBFN-all	S1/S5	60	50.6 ± 0.4	54.8 ± 0.4
RBFN	S1/S4	120	72.1 ± 0.3	68.3 ± 0.4
RBFN-all	S1/S4	120	83.2 ± 0.3	80 ± 0.3
RBFN	S1/S5	120	53.7 ± 0.4	50.8 ± 0.4
RBFN-all	S1/S5	120	64.5 ± 0.3	65.0 ± 0.3
RBFN	S1/S4	180	80.7 ± 0.3	81.3 ± 0.3
RBFN-all	S1/S4	180	90.2 ± 0.2	89.5 ± 0.3
RBFN	S1/S5	180	62.3 ± 0.2	66.8 ± 0.3
RBFN-all	S1/S5	180	73.7 ± 0.2	77.1 ± 0.3
RBFN	S1/S4	240	85.2 ± 0.2	-
RBFN-all	S1/S4	240	92.6 ± 0.2	-
RBFN	S1/S5	240	67.0 ± 0.3	-
RBFN-all	S1/S5	240	77.9 ± 0.2	-

users. The accuracy of $76.22 \pm 0.37\,\%$ over 50 runs with RAN using dataset. When we use the earlier optimal VT of 27, the accuracy did not improve (71.03 $\pm$0.34 % over 50 runs). Hence, this suggests that retuning of the VT for this experiment is required which may improve the accuracy. While with TEX dataset the accuracy of $63.84 \pm 0.31\,\%$ was achieved over 50 runs. The last accuracy was improved to $74.22 \pm 0.23\,\%$ when the velocity threshold of IVT was changed to the earlier optimal value of 26. The main results of the mentioned experiments in this section are shown in Table 4.11.

Table 4.11: Effect of the time gap between train and test data on average user identification accuracy with VT = 50, MFD = 100 ms with using 51 features.

Dataset	Time gap	No. runs	Identification Accuracy
RAN	30 minutes	50	$94.92 \pm 0.12\,\%$
RAN	1 year	50	$83.51 \pm 0.18\,\%$
TEX	30 minutes	50	$93.41 \pm 0.19\,\%$
TEX	1 year	50	$75.24 \pm 0.32\,\%$
RAN	30 minutes	1	$97.30\,\%$
RAN	1 year	1	$86.48\,\%$
TEX	30 minutes	1	$94.59\,\%$
TEX	1 year	1	$81.08\,\%$
RAN *	1 year	1	$81.08\,\%$
TEX *	1 year	1	$78.38\,\%$

* The results are from [George and Routray, 2016]

4.8.2 VST dataset

This dataset has a time gap of more than two weeks between the two trials and two hours between four sessions in each trial. Therefore, this data is also used to study the template aging on the user identification accuracy. As can be seen from Tables 4.9, and 4.10, Figures 4.10, and 4.12, the same experiments as explained in Section 4.7.2, were repeated. The accuracies were higher when the test data was a session (session 4) that was recorded on the same day as the train session (session 1). While the accuracies decreased in the case of testing with a session (session 5) that was recorded after a gap of two weeks from recording the first session (train session). For example the best accuracies of user identification were noted as $94.6 \pm 0.1\,\%$ with two hours time gap between train (session 1) and test (session 4) sessions, this accuracy was decreased to $(76.0 \pm 0.2)\,\%$ when two weeks time gap between train (session 1) and test (session 5) sessions were used. The last accuracy was improved to $85.0 \pm 0.19\,\%$ when the velocity threshold of the IVT was changed to optimal value of 100.

As expected, the accuracy of user identification decreases for all RAN, TEX, and VST datasets when there was a significant time gap between train and test sessions. This may be attributed to changing physiological parameters of the participants, device characteristics, and some other inexplicable effects.

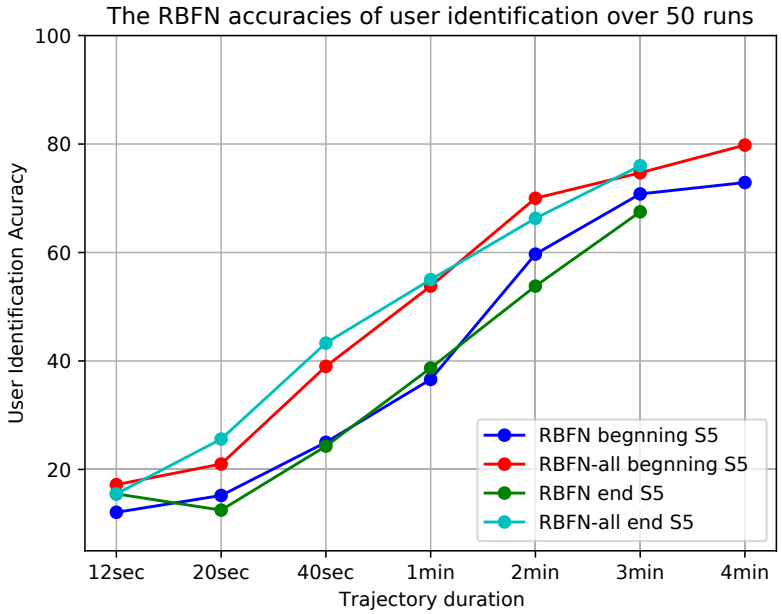


Figure 4.12: Train with S1 and Test with S5 after 2 weeks gap

4.9 Effect of Stimuli after Homogenizing the Different Datasets

In this section the question which stimuli is the best for user prediction is revisited. This was previously studied in Section 4.1 where a naive investigation indicated that RAN dataset provides the best user identification accuracy. For a fair comparison, the different aspects of these datasets, i.e., number of participants, trajectory length, and age group were homogenized. As mentioned previously, each dataset has a different number of participants (RAN: 153, TEX: 153, VST: 58 and GOF: 378). The minimum number of participants among all the datasets was selected (i.e. 58 participants). The participants were drawn at random when the datasets have a higher number of available participants. A fixed trajectory length (first minute) was selected to test the effect of the stimuli on the user identification accuracy. Since, RAN, TEX and VST datasets have participants only in the age group 20–40, hence, only participants of the corresponding age group 1 of the GOF dataset were selected. The experiments were carried out with both default and optimal IVT parameters of the respective datasets.

The results of these experiments are reported in Table 4.12. As observed from the results, the TEX data accuracy was the highest: 93.23 % with the optimal IVT parameter VT = 26 °/s. This was according to the state of art e.g. [Friedman et al., 2017, Lohr et al., 2020] since the eye movements during reading have been used to achieve some of the best biometric performances. However, it was important to consider the

number of fixations (if recorded with the same eye tracker) before concluding, which stimuli was the best, as the number of fixations played a vital role in attaining a higher accuracy. The higher the number of fixations, the better the accuracy achieved. Table 4.13 noted the number of fixations for the RAN datasets for a trajectory length of 1 minute. It was much lower than in the TEX dataset, which might influence the accuracy. The RAN stimuli in the Bioeye dataset were displayed every one second resulting in 100 big saccades and some smaller ones in between. Therefore, the total mean number of fixations (output of IVT algorithm) for the chosen number of users, during the first 60 seconds was 18831 (producing an accuracy of 92.82 %) when using the best VT i.e. 27, while for the whole 100 seconds the number of fixations was 31779 (producing an accuracy of 95.96 %) , which was much closer to the number of fixations of the TEX dataset i.e. 33584 (producing an accuracy of 93.23 %) . The RAN dataset could achieve higher accuracy than TEX if the fixation number can be increased for the first one minute of its trajectory length. If the randomly moving dot would have been shown for 100 times in a time span of 60 seconds there would have been more saccades and fixations possibly leading to greater accuracy.

Table 4.12: A comparison of performance metrics over 50 runs using 51 features with 58 participants of all the datasets with default and optimal IVT parameters.

Datasets	VT, (MFD = 100 ms)	Identification Accuracy
RAN	$50\,°/s$	$90.37 \pm 0.48\,\%$
TEX	$50\,°/s$	$90.90 \pm 0.10\,\%$
VST	$50\,°/s$	$71.62 \pm 0.35\,\%$
GOF	$15\,°/s$	$84.48 \pm 0.79\,\%$
RAN	$27\,°/s$	$92.82 \pm 0.38\,\%$
TEX	$26\,°/s$	$93.23 \pm 0.13\,\%$
VST	$100\,°/s$	$73.59 \pm 0.39\,\%$
GOF	$22\,°/s$	$85.21 \pm 0.64\,\%$

4.10 Study of Combined Factors

In the previous sections of this chapter, various user identification experiments were performed and the effect of various factors (IVT parameters, higher order derivative features etc.) affecting the accuracy of the user identification were studied. In this section, the effects of the above factors were combined which led to further improvement in user identification accuracy.

Table 4.13: Identification accuracy, number of fixations in BioEye data sets with two different VT and different trajectory length of RAN data.

RAN (MFD = 100 ms)			
Vel. threshold	Fix. No.	Identification Accuracy	Trajectory length
50	16891	84.43 ± 0.23 %	60 sec (start)
27	18831	89.75 ± 0.16 %	60 sec (start)
50	17451	82.37 ± 0.20 %	60 sec (end)
TEX (MFD = 100 ms)			
50	32944	90.90 ± 0.10 %	60 sec
26	33584	93.23 ± 0.13 %	60 sec

4.10.1 Combination of IVT Parameter Tuning and Higher-Order Derivatives Features

In Section 4.3 it was shown how tuning the IVT parameters can lead to a significant increase in the accuracy of user identification. Similarly, it was observed in Section 4.4 that including higher order derivatives gave better accuracies for user identification.

Here, a study of the combined link between these two factors affecting the accuracy of user identification. The best IVT parameters with the approach described in Section 4.3 were taken for the four datasets (RAN, TEX, VST, and GOF) and the user identification experiments with increasing number of higher order derivative features are repeated. Table 4.14 shows the average of accuracies over 50 runs of an increasing number of higher derivative features of fixation and saccade using the best IVT parameters for the mentioned four different datasets. Similarly, Figure 4.13a, b, c, and d compare the accuracies between the default and optimal IVT parameters for RAN, TEX, VST and GOF datasets respectively for including an increasing number of higher order derivative features. It can be noticed that for the RAN and VST datasets while using the optimal IVT parameters, the best accuracy were observed with fewer features which was until acceleration level ($95.96 \pm 0.09\,\%$ for RAN and $94.72 \pm 0.08\,\%$ for VST) in comparison to the case when the default IVT parameters were used, the best accuracies were: $94.58 \pm 0.10\,\%$ for RAN until jounce level and $86.52 \pm 0.23\,\%$ for VST until jerk level (see Table 4.5). For TEX and GOF datasets, the best accuracy were achieved with including jounce ($93.39 \pm 0.09\,\%$ for Tex) and jerk ($84.48 \pm 0.44\,\%$ for GOF) level features while when the default IVT parameters were used, the best accuracies were until jounce level: $91.96 \pm 0.08\,\%$ for TEX and $81.02 \pm 0.53\,\%$ for GOF (see Table 4.5).

Overall, one can conclude that finding the best IVT parameters and including higher order derivative features were somewhat complementary: When the default threshold is used, including higher order derivatives always increases the best accuracy for all the four datasets. However, finding the best IVT threshold may yield similar best case accuracies with lower derivative features as this was the case for RAN and VST datasets. Nevertheless, for TEX and GOF datasets, the accuracy still increases when jounce and jerk level features in TEX and GOF datasets respectively were included even after using the optimal IVT parameters. This recommends that the default velocity threshold can satisfy increasing the accuracy when using the higher derivative features. Also, we can see from the Table 4.5 that using only the higher derivative features without tuning the IVT parameters can cover the accuracy lost by using the wrong IVT velocity threshold.

Table 4.14: Performance metrics over 50 runs with varying number of features of higher order derivatives of the gaze trajectory of all the datasets with their best Velocity Threshold (VT).

Deriv. order	Feat. No.	RAN Acc.% (153 participants, VT=27 °/s, MFD= 96 ms)	TEX Acc.% (153 participants, VT=26 °/s, MFD= 98 ms	VST Acc.% (58 participants, VT=100 °/s, MFD= 100 ms	GOF Acc.% (153 participants, VT=22 °/s, MFD= 100 ms
0. Position	14	88.08 ± 0.19	85.77 ± 0.13	91.86 ± 0.22	57.53 ± 0.53
1. Velocity	33	93.71 ± 0.15	91.09 ± 0.16	94.41 ± 0.13	76.55 ± 0.55
2. Acceleration	51	**95.96± 0.09**	93.23 ± 0.13	**94.72 ± 0.08**	82.67 ± 0.52
3. Jerk	69	95.16± 0.12	93.12 ± 0.12	93.59 ± 0.12	**84.48 ± 0.44**
4. Jounce	87	95.37± 0.08	**93.39 ± 0.09**	91.59 ± 0.20	84.08 ± 0.46
5. Crackle	105	94.89± 0.10	93.04 ± 0.11	90.34 ± 0.19	83.20 ± 0.48

Finally, other experiments were done to study the effect of removing only the angular features i.e. the six M3S2K statistical of the angular feature with higher-order derivative level (the M3S2K features from 16 to 21 of the angular velocity features and the features from 16 to 21 of the angular acceleration features, see Table 3.3). The aim of these experiments was to investigate if we still can keep the high accuracies achieved from combining the effect of IVT parameter tuning and higher-order derivatives features as reported in Table 4.14 with a fewer number of features to reduce the computational load. As can be seen in Table 4.15, RAN, TEX, and VST datasets were used using velocity and acceleration level. After removing the angular features, 27 features were used instead of 33 in velocity level and 39 features instead of 51 in acceleration level. The results show the following pattern: with the RAN dataset, the accuracy dropped by 4.3 % while it decreased by only 1.1 % and 0.7 % for

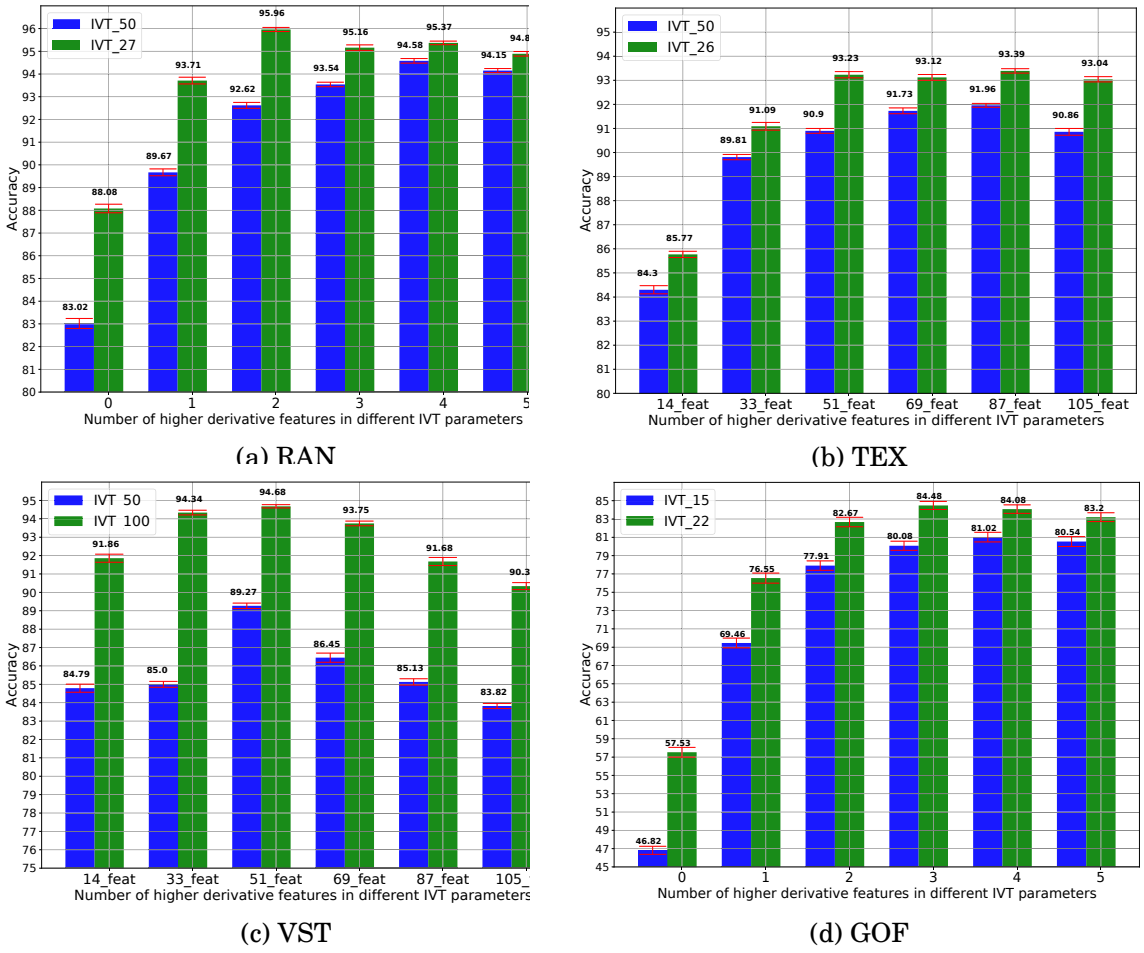


Figure 4.13: Performance metrics over 50 runs with different number features of higher order derivatives of the gaze trajectory for different datasets.

TEX and SVT respectively with the velocity level. While with acceleration level, with the RAN dataset, the accuracy decreased by 3.2 % while it decreased by only 1.1 % and 0.6 % for TEX and SVT respectively. That can be explained as including the angular features are more important while using random searching stimuli than with using reading stimuli.

4.10.2 Combining IVT Tuning and Higher-Order Derivatives with Blink Classifier

To investigate if we can further increase the accuracies achieved in Section 4.10.1, a study for the case of including the blink classifier with the optimal IVT threshold and set of higher order derivatives features was performed. As mentioned in Section 4.5, the Nelder-Mead optimization method was employed to find the optimal weights for combining the fixation, saccade and blink classifiers. The achieved accuracy for the

Table 4.15: Performance metrics over 50 runs with varying number of features of higher order derivatives of the gaze trajectory of different datasets with their best Velocity Threshold (VT) after removing the angular features.

Deriv. order	Feat. No.	RAN Acc.% (153 participants, VT=27 °/s, MFD= 96 ms)	TEX Acc.% (153 participants, VT=26 °/s, MFD= 98 ms	VST Acc.% (58 participants, VT=100 °/s, MFD= 100 ms
1. Velocity	27	89.45 ± 1.09	90.05 ± 0.98	93.72 ± 1.13
2. Acceleration	39	92.77 ± 0.88	92.14 ± 0.83	94.13 ± 1.03

three datasets were reported in Table 4.16. The accuracy improved by $0.70\,\%$ for the RAN dataset yielding the final accuracy of $96.64 \pm 0.07\,\%$, by $0.15\,\%$ for the TEX dataset leading to $93.53 \pm 0.11\,\%$, and by $0.07\,\%$ for the VST dataset leading to $94.79 \pm 0.03\,\%$.

Table 4.16: Performance metrics over 50 runs using blink classifier in RAN data (VT = 27 °/s and MFD = 96 ms), TEX data (VT = 26 °/s and MFD = 98 ms), and VST data (VT = 100 °/s and MFD = 100 ms).

Dataset	Fix/Sac/blink features	CLF weights Sac/Fix/Blink	Identification Accuracy
RAN	51/51/7	0.5/0.5/0.0	$95.96 \pm 0.09\,\%$
	51/51/7	0.543/0.447/0.010	$\mathbf{96.64 \pm 0.07\,\%}$
TEX	87/87/0	0.5/0.5/0.0	$93.39 \pm 0.09\,\%$
	87/87/7	0.529/0.466/0.005	$\mathbf{93.53 \pm 0.11\,\%}$
VST	51/51/0	0.5/0.5/0.0	$94.72 \pm 0.08\,\%$
	51/51/7	0.513/0.477/0.010	$\mathbf{94.79 \pm 0.03\,\%}$

4.11 Conclusion

This chapter covered user identification experiments using different datasets (RAN, TEX, MIT, VST, and GOF datasets). The proposed approach performed consistently and robustly well with different datasets of varying stimuli. This approach improved the current state-of-the-art performance in gaze biometrics. The best achieved accuracies were $96.64\,\%$ in RAN, $93.53\,\%$ in TEX, $94.72\,\%$ in VST, and $84.48\,\%$ in GOF datasets. An extensive study has been carried out to investigate different factors (for example, the IVT parameters, higher-order derivative features, blinks, gender and age, template aging, stimuli, etc.) that can affect user identification performance.

The effect of stimuli was investigated by implementing user identification exper-

iments using different stimuli. The results using the default IVT parameters and all trajectory lengths that were available for each dataset showed that RAN was giving the highest accuracies $92.62\,\%$ with RAN, $90.90\,\%$ with TEX, $85.69\,\%$ with VST, and $77.91\,\%$ with GOF and with using the optimal IVT parameters the accuracies are $95.96\,\%$ with RAN, $93.23\,\%$ with TEX, $94.72\,\%$ with VST, and $82.67\,\%$ with GOF. While after homogenizing the datasets by taking the same number of users, default IVT threshold and trajectory length in all the four stimuli the results showed that both RAN and TEX were giving the best accuracies in comparison with VST and GOF stimuli ($90.37\,\%$ with RAN, $90.90\,\%$ with TEX, $71.62\,\%$ with VST, and $84.84\,\%$ with GOF) and with using the optimal IVT parameters the accuracies are $92.82\,\%$ with RAN, $93.23\,\%$ with TEX, $73.59\,\%$ with VST, and $85.21\,\%$ with GOF..

Another factor was investigated to improve the accuracy i.e. the effect of IVT parameters. The IVT algorithm has two parameters, velocity threshold (VT) and the minimum fixation duration (MFD). A systematic parameter tuning was conducted to determine which IVT parameter leads to the highest accuracy. The results showed that tuning the VT parameter and fixing the MFD to 100 ms was giving the highest impact on the user identification accuracy. The accuracy was increased by 3 % for RAN, by 2 % for TEX, by 9 % for VST, and by 5 % for GOF datasets (95.96 % with RAN, 93.23 % with TEX, 94.72 % with VST, and 82.67 % with GOF).

The effect of higher-order derivatives was studied and the results showed that the accuracies were increased by 2 % for RAN, by 1 % for TEX and VST, and by 3 % for the GOF dataset after adding higher derivative features.

Additionally, the effect of blinking features was investigated and the outcome showed that adding a blinking classifier can increase the accuracy by 1 % for RAN and VST, and by 0.5 % for TEX.

Furthermore, combining the three factors which had the greatest impact on the accuracy improvement (IVT parameters, higher-order derivative features, and blink classifier) was implemented and that increased the accuracy by 4 % for RAN, by 3 % for TEX, and by 9 % for VST datasets.

Other findings in this chapter suggested that the user identification works better in the solo female group (accuracy of 88.85 %) than in the solo male group (accuracy of 77.37 %) with the GOF dataset and hence the user identification was biased towards gender. Similarly, it worked better in the older age group (accuracy of 91.43 % with age group from 41 to 72) than in a younger age group (accuracy of 85.96 % with age group from 20 to 40) of participants.

We studied the impact of the trajectory length on user identification. This was done with three datasets (MIT, VST, and RAN). The results of the MIT dataset suggested that only 120 seconds (40 samples) of trajectory data was necessary to achieve a good accuracy. Same pattern is observed with the VST and RAN datasets.

In addition, the accuracy of user identification decreases for all RAN, TEX, and VST datasets when there was a significant time gap between train and test sessions. This may be attributed to changes in the physiological parameters of the participants. Therefore, more work is needed on improving the accuracy of user identification when there is a time gap between train and test sessions. This is crucial for the usability of eye movement biometrics for real-world applications.

Finally, the effects of the above experiments were combined which lead to further improvement in user identification accuracy. For example combining IVT tuning and higher order derivative with blink classifier yielded the final accuracy of $96.64 \pm 0.07\,\%$ for the RAN dataset (see Table 4.16), $93.53 \pm 0.11\,\%$ for the TEX dataset, and $94.79 \pm 0.03\,\%$ for the VST dataset.

Chapter 5

Gender Prediction

This chapter covers all the experiments of the gender prediction that were conducted in this study. The gender prediction task was carried out on the datasets that provide gender information i.e. Dyslexia (this work is based on [Zaidawi et al., 2020]), VST, and GOF. The gender prediction work with VST and GOF is based on [Haria et al., 2022]. Rishabh Haria from the Database Group at the University of Bremen has contributed to use of the new type of ROI features with the GOF dataset.

5.1 Dyslexia Dataset

In this section we discuss gender and dyslexia prediction based on eye movements, using the Dyslexia dataset (see Section 3.1.5). Previous studies have suggested that eye tracking data can be used to predict gender, however, there is a limited number of such studies available and the reported accuracies were rather low (we are only aware of two studies that explicitly mentioned accuracies: 64 % in [Moss et al., 2012] and 70 % in [Sargezeh et al., 2019]). Moreover, it was believed that the accuracies were lower for young persons than older ones [Miyahira et al., 2000b]. The motivation to use dyslexia dataset was that it involves young children aged 9–10, so it can be used to either confirm or disprove the hypothesis of the previous study. For the Dyslexia dataset, different groups of participants are considered i.e. the dyslexic participants group (DG), the non-dyslexic participants group (NDG), and the mixed participants group (MG). Five machine learning classifiers namely, SVM, RBFN, NB, logReg, and RF were used to predict the gender of a given set of participants. In some experiments 38 participants had been selected in both cases in order to have a fair comparison between the results of the two isolated groups (dyslexic and non-dyslexic). In order to deal with the unbalanced nature of the dataset, we employed a down-sampling method where we kept the size of rare class (females) and randomly

selected an equal number of participants from the abundant class (males) in each iteration. In contrast to using a fixed subset, random sampling gives a more accurate estimate of the error and prevents a selection of best-performing samples. For every experiment, the used participants were randomly chosen from all available participants for the specific categories. In all gender prediction experiments, the gender ratio is always balanced and the ratio of non-dyslexic to dyslexic candidates is always the same in the male and the female group. To minimize the influence of each individual candidate in a set, a cross validation with 1000 runs was performed. For every run, the number of desired candidates was randomly chosen from all available participants. The selected candidates were randomly split into a training-set ($80\,\%$) and a test-set ($20\,\%$). Two instances of each classifier were trained; one to predict the gender of fixation segments and the other to predict the gender of saccadic segments. The final prediction was the average of the probabilities of all segments of a given participant. The *gender prediction accuracy* was calculated as the maximum percentage of correct classifications achieved by any of the mentioned five classifiers. As explained in Section 3.3.2, the top features that were used in these experiments were ranked by their Fisher Scores. Tables 5.1, 5.2, and 3.5, are showing the top selected features in NDG, DG, and MG groups respectively. A maximum of twelve top features was considered since the accuracy started to drop as we increased the number of features. The experiment was performed using incremental numbers of features ranging from 2 to 12.

Table 5.1: Features ranked by Fisher Score (FS) for Dyslexia datset of Non-dyslexic group (Gender prediction)

Fixation features	Fisher Score ($\times 10^{-4}$)	Saccade features	Fisher Score ($\times 10^{-4}$)
1) Mean of velocity	1886	1) Skewness of X	2365
2) Mean of angular velocity	1472	2) Skewness of Y	1074
3) Fixation amplitude	1469	3) STD of acceleration Y	1035
4) Dispersion	1442	4) Maximum of acceleration Y	970
5) Median of angular velocity	1441	5) STD of velocity Y	922
6) STD of Y	1282	6) Variance of velocity Y	902

5.1.1 Prediction in Isolated Groups

The hypothesis in this study was that the gender prediction features for non-dyslexic participants are different from those of dyslexic participants. Hence, gender predic-

Table 5.2: Features ranked by Fisher Score (FS) for Dyslexia datset of Dyslexic group (Gender prediction)

Fixation features	Fisher Score ($\times 10^{-4}$)	Saccade features	Fisher Score ($\times 10^{-4}$)
1) Fixation ratio	2664	1) Saccade number	2499
2) STD of angular velocity	2163	2) Median of angular velocity	1151
3) Variance of angular velocity	2148	3) Total saccade duration	1019
4) Maximum angular velocity	2020	4) Maximum of velocity X	916
5) Mean of the velocity	1850	5) Maximum of acceleration X	789
6) Mean of angular velocity	1673	6) STD of acceleration X	572

tion experiments were performed first in isolated groups i.e. groups of only non-dyslexic participants or groups of only dyslexic participants, referred to as non-dyslexic and dyslexic groups hereafter. As mentioned in Section 5.1, the top Fisher Score features for both groups were calculated separately using Equation 3.15. The obtained orders of features are different, but similar to Table 3.5 as can be seen in Tables 5.1, and 5.2. Multiple experiments were run with increasing numbers of top features (to reduce the search space, always equal numbers of fixation and saccadic features were used). A plot of the achieved accuracies with different classifiers against the number of top features is shown in Figure 5.1. Table 5.3 shows the gender prediction accuracies for different classifiers with different amounts of features for the non-dyslexic group. A similar experiment was performed for the dyslexic group as shown in Table 5.4. It can be observed that the logReg classifier peaks with *four* features (two "fixation features" and two "saccadic features") while the other classifiers peak with different numbers of features in both dyslexic and non-dyslexic groups.

For the saccadic features, the difference between non-dyslexic and dyslexic groups can be explained as follows: skews in X and Y directions are important to distinguish differences in reading speed; dyslexic participants do not read the text, i.e., their eyes do *not* follow the lines of text, that is why other features were more important there.

By using groups of 38 non-dyslexic and dyslexic participants and their respective top four features, the gender prediction accuracies (with logReg) were:

- **63.8** $\pm$ **0.5** % for groups of non-dyslexic participants and
- **60.7** $\pm$ **0.6** % for groups of dyslexic participants.

The accuracy of 63.8 % was surprising, as a previous

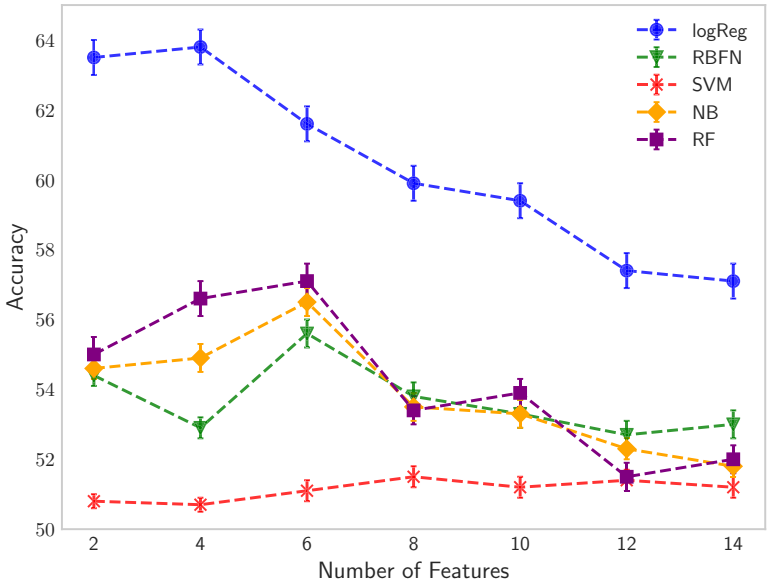


Figure 5.1: Prediction accuracies for all classifiers against numbers of features with 38 non-dyslexic participants.

study [Miyahira et al., 2000b] suggested that the gender differences in eye movement data were observed only in the adult group of participants and not statistically significant in pre-puberty children. This may suggest that the gender prediction would be hardly possible via eye-tracking in young children. This study experiments suggested that gender differences in eye movement are significant even in pre-puberty children, making the gender prediction possible using eye tracking data. We are not aware of any studies which deal with gender prediction in dyslexic participants using eye tracking data. It should also be noted that dyslexic participants have a condition that makes it difficult to perceive letters and their order [Stein, 2014]. In this case, the dyslexic participants may actually perceive a text similar to an image. Hence, for non-dyslexic participants it is a reading task, while for dyslexic it is more like seeing an image. Nevertheless, gender prediction in dyslexic participants is possible but with an accuracy lower than that of non-dyslexic participants.

Another experiment was conducted with all the available 21 dyslexic females and 21 dyslexic males randomly selected over each run. The results showed that the best accuracy was $\mathbf{61.5} \pm \mathbf{0.5}\,\%$ using again only four features (2 fixation and 2 saccade features) and with the logReg classifier. In comparison with using 38 dyslexic (19 males and 19 females) participants, the accuracy improved by 1 % (from $60,7\,\%$ to $61.5\,\%$).

In order to check the robustness of our classifiers, we additionally computed met-

Table 5.3: Gender prediction accuracies with standard error of the mean for different classifiers with different amounts of features in non-dyslexic group of 19 males and 19 females.

#Features (fix + sac)	logReg %	RBFN %	NB %	SVM %	RF %
2 (1+1)	63.5 ± 0.5	54.4 ± 0.3	54.6 ± 0.3	50.8 ± 0.2	55.0 ± 0.5
4 (2+2)	$\mathbf{63.8 \pm 0.5}$	52.9 ± 0.3	54.9 ± 0.4	50.7 ± 0.2	56.6 ± 0.5
6 (3+3)	61.6 ± 0.5	55.6 ± 0.4	56.5 ± 0.4	51.1 ± 0.3	57.1 ± 0.5
8 (4+4)	59.9 ± 05	53.8 ± 0.4	53.5 ± 0.4	51.5 ± 0.3	53.4 ± 0.4
10 (5+5)	59.4 ± 0.5	53.3 ± 0.4	53.3 ± 0.4	51.2 ± 0.3	53.9 ± 0.4
12 (6+6)	57.4 ± 0.5	52.7 ± 0.4	52.3 ± 0.3	51.4 ± 0.3	51.5 ± 0.4

Table 5.4: Gender prediction accuracies with standard error of the mean for different classifiers with different amounts of features in dyslexic group of 19 males and 19 females.

#Features (fix + sac)	logReg %	RBFN %	NB %	SVM %	RF %
2 (1+1)	60.5 ± 0.6	52.3 ± 0.5	58.7 ± 0.5	58.7 ± 0.5	51.7 ± 0.5
4 (2+2)	$\mathbf{60.7 \pm 0.6}$	54.6 ± 0.5	58.3 ± 0.5	58.6 ± 0.5	52.0 ± 0.6
6 (3+3)	57.0 ± 0.6	52.9 ± 0.5	59.2 ± 0.5	58.5 ± 0.5	51.6 ± 0.6
8 (4+4)	56.9 ± 0.6	54.2 ± 0.5	58.1 ± 0.5	58.4 ± 0.5	51.2 ± 0.6
10 (5+5)	57.0 ± 0.6	56.9 ± 0.5	58.3 ± 0.5	58.4 ± 0.5	51.8 ± 0.6
12 (6+6)	56.8 ± 0.6	57.9 ± 0.5	58.2 ± 0.5	58.7 ± 0.5	51.9 ± 0.6

rics such as recall, precision and F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUROC) [Sokolova et al., 2006, Baratloo et al., 2015]. These metrics for our best performing classifier i.e. LogReg in the non-dyslexic group are: Recall $61.7 \pm 0.6\,\%$, Precision $65.7 \pm 0.6\,\%$, F1-Score $62.4 \pm 0.5\,\%$, AUROC 0.67 ± 0.006 and in dyslexic group were: Recall $59.3 \pm 0.7\,\%$, Precision $61.3 \pm 0.6\,\%$, F1-Score $59.6 \pm 0.6\,\%$, AUROC 0.62 ± 0.01. All these values were in a similar range and hence the classifiers were not biased towards predicting false positives or false negatives. An overview of these metrics for different classifiers in non-dyslexic group and in dyslexic group are shown in Tables 5.5 and 5.6 respectively.

Figure 5.2 shows an example of confusion matrix using one of the runs of the best accuracy with the logReg classifier and 38 participants in the non-dyslexic group. The figure shows the correct and wrong predictions in both test and training sets. It can be seen that the training set consists of 30 participants (15 males and 15 females) while the test set consists of 8 participants (4 males and 4 females) for the test set only two females were predicted wrong, so the accuracy of this run was $75\,\%$. On the other hand for the train set, as can be seen in Figure 5.2, 12 participants (4

males and 8 females) were predicted wrongly, Furthermore, Figure 5.3 shows ROC examples (using two runs) of the best accuracy with logReg classifier and 38 dyslexic groups. As explained in Section 3.5.2, the AUROC estimates the area under the ROC curve and it is between 0 and 1. The AUCROC can be used to identify which classifier is better as the larger the area under the ROC curve, the better the classifier. In Figure 5.3a the ROC curve was better since the model rated a random positive case higher than a random negative and the AUROC was larger than the once in Figure 5.3b. In Figure 5.3b the ROC is close to the diagonal, so the classifier behaves randomly since the values are close to 0.5.

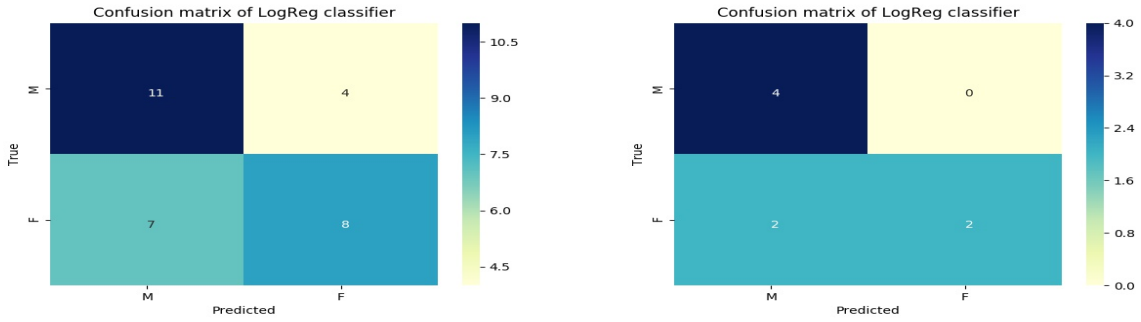
Table 5.5: Gender prediction metrics in non-dyslexic group of 19 Males and 19 femails.

Metrics	logReg %	RBFN %	NB %	SVM %	RF %
Recall	61.7 ± 0.6	18.8 ± 0.9	28.7 ± 1.0	21.4 ± 1.1	42 ± 0.9
Precision	65.7 ± 0.6	24.9 ± 1.1	41.7 ± 1.3	18.3 ± 0.9	55.8 ± 1.0
F1-Score	62.4 ± 0.5	19.0 ± 0.8	30.5 ± 0.9	17.5 ± 0.9	45.2 ± 0.8
AUROC	0.67 ± 0.006	0.64 ± 0.006	0.61 ± 0.006	0.52 ± 0.007	0.60 ± 0.006

Table 5.6: Gender prediction metrics in dyslexic group of 19 Males and 19 femails.

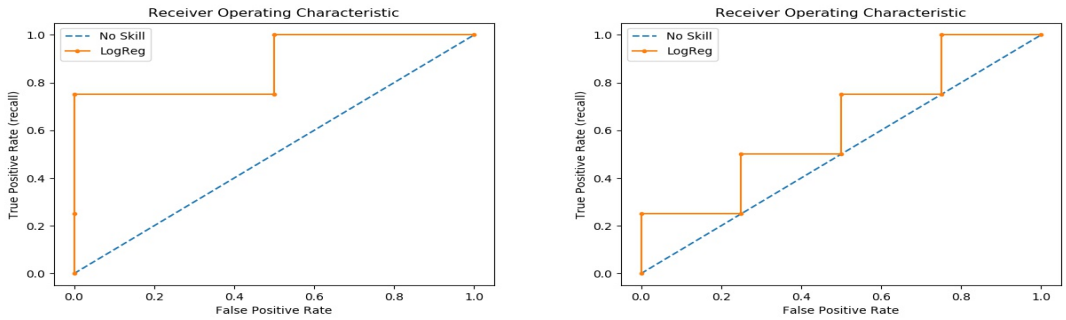
Metrics	logReg %	RBFN %	NB %	SVM %	RF %
Recall	59.3 ± 0.7	44.1 ± 0.8	47.2 ± 0.7	50.0 ± 0.8	52.3 ± 0.8
Precision	61.3 ± 0.6	55.0 ± 0.9	61.9 ± 0.7	58.5 ± 0.8	52.2 ± 0.7
F1-Score	59.6 ± 0.6	46.4 ± 0.7	52.1 ± 0.6	52.5 ± 0.7	50.6 ± 0.6
AUROC	0.62 ± 0.01	0.59 ± 0.01	0.56 ± 0.01	0.48 ± 0.005	0.56 ± 0.01

Feature Ranking As mentioned in Section 3.3.2 in addition to the Fisher Score, other methods were used to rank the features for dyslexia data i.e. ANOVA, Chi, and RFECV. Hence, gender prediction experiments (with non-dyslexic group) were conducted with increasing numbers of top features for each method to compare the results with those results while using the top features of the Fisher Score method. As seen in Table 5.7, ranking the features with this data using the Fisher score was better than the results of RFECV and similar to those obtained by ANOVA and Chi. Therefore, the Fisher score was used for all the experiments here.



(a) Train set (b) Test set

Figure 5.2: Confusion matrix using the first run of the best accuracy with logReg classifier and 38 non-dyslexic group, using the four top features



(a) Run number 1 (b) Run number 500

Figure 5.3: ROC examples (from run number 1 and 500) of the best accuracy with logReg classifier and 38 dyslexic group, using the four top features

Table 5.7: Gender prediction accuracy with standard error of the mean for different feature ranking methods with their top four features of 19 females and 19 males with only non-dyslexic participants.

#Features (fix + sac)	Fisher score %	ANOVA %	Chi %	RFECV	Classifier
4 (2+2)	**63.8 ± 0.5**	63.7 ± 0.5	63.7 ± 0.5	54.6 ± 0.5	logReg

Hyper-Parameters Finally, the classifier's hyper-parameters values used in the isolated groups experiments are: $K = 2$ for RBFN, "gamma" value of 0.1 and $C = 0.1$ for SVM, and "Max-depth" value of 10 and "N-estimators" value of 200 for RF. These are found using a grid search on the hyper-parameters.

5.1.2 Prediction in Mixed Groups

The experiments for gender prediction in mixed groups of non-dyslexic and dyslexic participants are described in this section. Similar to the previous section, the top four Fisher Score features from Table 3.5 were used which include the top two fixation and saccadic features. The gender ratio was always kept balanced in each group. The results are shown in Table 5.8. As can be seen, accuracies dropped when dyslexic and non-dyslexic children were mixed. Especially the logReg classifier (which achieved the best accuracies before) dropped to an accuracy of below 50 % for groups containing 8 dyslexic and 28 non-dyslexic children. It can be observed that in the border cases of isolated non-dyslexic or dyslexic groups, the gender prediction accuracies were higher and dropped down more and more with mixed groups. This showed that gender prediction in mixed groups was hardly possible, using these five standard classifiers in contrast to isolated non-dyslexic or dyslexic groups of participants. The drop of accuracy was hypothesized in mixed groups to be the result of "contradictory" features for non-dyslexic and dyslexic participants (see Section 5.1.5). It also explains why the accuracies for isolated non-dyslexic and dyslexic groups trained with the mixed group top features were lower than the ones reported in Section 5.1.1. To better demonstrate the nature of contradictory features, the top non-dyslexic features were used in the isolated dyslexic group of 38 participants and the top dyslexic features were used in non-dyslexic isolated group of participants. The gender prediction accuracy drops to **48.5** $\pm$ **0.5** % in non-dyslexic participants (from **63.8** $\pm$ **0.5** %) and to **55.9** $\pm$ **0.6** % in dyslexic participants (from **60.7** $\pm$ **0.6** %). These experiments show that the top features were inherently different for the two isolated groups and hence, a hierarchical classifier may be used to improve the gender prediction accuracy in mixed groups.

A fair comparison can be made if we use the same number of participants, i.e., 38 in the border cases (i.e., non-dyslexic or dyslexic): the accuracies (with logReg) then are: **61.4** $\pm$ **0.5** % for the non-dyslexic group (see Table 5.9), and **59.7** $\pm$ **0.6** % for the dyslexic group (see Table 5.10) for more information of all the classifiers results with different number of features. In non-dyslexic group: Recall 75.0 ± 0.7 %, Precision 58.6 ± 0.4 %, F1-Score 65.4 ± 0.5%, AUC 0.616 ± 0.006. In dyslexic group: Recall 51.6 ± 0.6 %, Precision 62.5 ± 0.7 %, F1-Score 55.6 ± 0.6 %, AUC 0.596 ± 0.007. The accuracies in the non-dyslexic or dyslexic group trained with their specific features are higher than the reported accuracy in mixed group by **2.4** % in the non-dyslexic group and **1** % in the dyslexic group.

Moreover, the accuracy for larger mixed group of 78 participant (19 non-dyslexic males, 19 non-dyslexic females, 19 dyslexic males, and dyslexic females) is achieving the best case accuracy of **56.4** $\pm$ **0.3** % as can be seen in Table 5.8. This motivates

us to build a hierarchical classifier for gender prediction in mixed groups which is presented in the next subsection.

Table 5.8: Accuracies with standard error of the mean for different classifiers to predict gender in mixed groups with varying proportions of non-dyslexic and dyslexic children.

Dyslexic / Non-Dyslexic	logReg %	RBFN %	NB %	SVM %	RF %
0 / 36	60.8 ± 0.5	54.2 ± 0.5	59.2 ± 0.5	59.8 ± 0.5	52.0 ± 0.6
8 / 28	47.9 ± 0.5	54.9 ± 0.5	51.6 ± 0.5	54.4 ± 0.5	49.4 ± 0.5
18 / 18	50.1 ± 0.5	53.9 ± 0.5	54.6 ± 0.5	52.5 ± 0.5	50.8 ± 0.5
28 / 8	54.1 ± 0.5	53.6 ± 0.5	55.9 ± 0.5	53.1 ± 0.5	51.9 ± 0.6
36 / 0	59.6 ± 0.6	53.4 ± 0.5	56.4 ± 0.5	56.0 ± 0.6	52.4 ± 0.5
38 / 38	56.4 ± 0.3	52.4 ± 0.3	53.7 ± 0.3	54.2 ± 0.3	49.8 ± 0.4

Table 5.9: Gender prediction accuracy with standard error of the mean for different classifiers with different amount of features of 19 females and 19 males with only non-dyslexic participants.

#Features (fix + sac)	logReg %	RBFN %	NB %	SVM %	RF %
2 (1+1)	53.6 ± 0.6	50.3 ± 0.5	49.6 ± 0.5	49.1 ± 0.5	$\pm 0.$
4 (2+2)	$\mathbf{61.4 \pm 0.5}$	54.9 ± 0.5	59.7 ± 0.5	59.4 ± 0.5	$\pm 0.$
6 (3+3)	59.5 ± 0.5	52.0 ± 0.5	59.0 ± 0.5	57.8 ± 0.5	$\pm 0.$
8 (4+4)	59.5 ± 05	53.4 ± 0.5	58.8 ± 0.5	58.1 ± 0.5	$\pm 0.$
10 (5+5)	59.4 ± 0.5	55.8 ± 0.5	57.8 ± 0.5	58.3 ± 0.5	$\pm 0.$
12 (6+6)	59.3 ± 0.5	55.9 ± 0.5	57.5 ± 0.5	58.0 ± 0.5	$\pm 0.$

Table 5.10: Gender prediction accuracy with standard error of the mean for different classifiers with different amount of features of 19 females and 19 males with only dyslexic participants.

#Features (fix + sac)	logReg %	RBFN %	NB %	SVM %	RF %
2 (1+1)	53.5 ± 0.6	50.0 ± 0.5	55.1 ± 0.5	50.6 ± 0.5	51.4 ± 0.6
4 (2+2)	$\mathbf{59.7 \pm 0.6}$	51.9 ± 0.6	56.5 ± 0.5	55.4 ± 0.6	52.3 ± 0.5
6 (3+3)	57.3 ± 0.6	48.8 ± 0.5	56.5 ± 0.5	54.9 ± 0.5	51.9 ± 0.5
8 (4+4)	57.2 ± 0.6	49.8 ± 0.5	56.4 ± 0.5	55.3 ± 0.6	51.7 ± 0.5
10 (5+5)	57.1 ± 0.6	50.1 ± 0.5	56.0 ± 0.5	55.0 ± 0.6	51.4 ± 0.5
12 (6+6)	57.0 ± 0.6	51.6 ± 0.5	55.9 ± 0.5	55.1 ± 0.5	51.2 ± 0.5

Hyper-Parameters Finally, the same hyper-parameters values in Section 5.1.1 were used in the mixed group of participants.

5.1.3 Hierarchical Classifier

This section introduces a hierarchical classifier that is able to predict gender in mixed groups of non-dyslexic and dyslexic children, with accuracies that are considerably higher than those of Section 5.1.2. Hierarchical classifiers are studied in several surveys [Aly, 2005, Freitas and Carvalho, 2007, Lorena et al., 2008]; further, see also [Schwenker, 2000, Cesa-Bianchi et al., 2006, Hao et al., 2007, Wei et al., 2017, Yao et al., 2018]. In [Kinsman et al., 2010], a hierarchical classifier used to solve the classification and fixation labeling problem of a large number of images of items.

The idea of the hierarchical classifier of this study is to first predict dyslexia by developing a dyslexia prediction classifier for a mixed group of participants which can be used to predict whether a participant is non-dyslexic or dyslexic (as shown in [Benfatto et al., 2016] this can be done with high accuracies above 90 %). Our dyslexia prediction classifier achieved an accuracy of 92.8 % (see Table 5.12). The next step is to use the according classifier (trained on non-dyslexic or on dyslexic children) which works best for the individual task (see Figure 5.5). The targeted gender prediction classifier for that participant using the approach is described in Section 5.1.1. The hypothesis is that such a hierarchical classifier should increase the gender prediction accuracy in a mixed group of participants. The curve of gender prediction accuracies with different number of mixed group of participants can be seen with blue color in Figure 5.4. The following are the details of the three individual classifiers, CLF1, CLF2, and CLF3 of the hierarchical classifier.

5.1.3.1 Dyslexia prediction classifier (CLF1)

Dyslexia prediction is the first classifier in our hierarchical classifier. The following are the details of how this classifier was built. First, the Fisher Score was computed again by substituting “non-dyslexic” for class1 and “dyslexic” for class2 in Equation 3.15 as following:

$$\text{Fisher Score}(f) \quad = \quad \frac{\left(\text{mean}(f_{\text{non-dyslexic}}) - \text{mean}(f_{\text{dyslexic}})\right)^2}{\text{std}^2(f_{\text{non-dyslexic}}) + \text{std}^2(f_{\text{dyslexic}})}, \tag{5.1}$$

where f stands for the various values of a specific feature. The top features with their Fisher Score for dyslexia prediction task can be seen in Table 5.11.

Then, an increased number of top features (for 176 participants: 138 males and 38 female, both with equal numbers of 88 non-dyslexic and 88 dyslexic children) was selected and the dyslexia prediction accuracy was computed. The top two features

Table 5.11: Features ranked by Fisher Score for dyslexia prediction

Fixation features	Fisher Score	Saccade features	fisher Score
1) Total duration of Fix	4.30	1) Mean duration od Sacc	1.25
2) Distance with previous Fix	1.40	2) Maximum of angular acceleration	1.14
3) Fixation ratio over saccade time	0.58	3) Dispersion	1.13
4) Fixation ratio	0.49	4) total of angular velocity	1.13
5) Mean duration of Fix	0.48	5) Maximum of angular velocity	1.11
6) Mean of angular acceleration	0.42	6) STD of X direction	1.03

ranked by their Fisher Score in mixed participants were giving the best accuracies, see Table 5.12. The Naïve Bayes (NB) classifier with only two features was found to give the best accuracy of **92.8** $\pm$ **0.2**% (While with using all the available 185 participants, the best accuracy was 92.3 $\pm$ 0.2 %, for both NB and SVM classifiers and with using only the top two features classifiers). Hence, we decided to use the top two features and NB classifier for CLF1 of the hierarchical classifier and balanced number of 88 dyslexic and 88 non-dyslexic participants that were selected randomly over 1000 runs.

Table 5.12: Dyslexia prediction accuracy with standard error of the mean for different classifiers with the top two Fisher Score features (train to test ratio 80:20).

# Participants No.	logReg %	RBFN %	NB %	SVM %	RF %
185 (97 D, 88 ND)	92.2 ± 0.2	91.4 ± 0.2	92.3 ± 0.2	92.3 ± 0.2	87.1 ± 0.2
176 (88 D, 88 ND)	92.3 ± 0.2	91.8 ± 0.2	$\mathbf{92.8 \pm 0.2}$	92.3 ± 0.2	87.3 ± 0.2

Note that the dyslexia prediction accuracy in [Benfatto et al., 2016] was 95.6 % $\pm$ 4.5 %, slightly higher than ours, but the numbers are not directly comparable, as they used a 90:10 ratio of training to testing data of all the 185 participants. Their accuracy for SVM classifiers is based on recursive feature elimination SVM-RFE and selects 48 features from the original feature set of 168 features over 100 runs. Also, they used a different segmentation algorithm which is the dynamic dispersion threshold for fixation and saccade segmentation [Benfatto et al., 2016]. In contrast, we used 185 participants with a 90:10 training to test ratio with only two features over 1000 runs. The IVT algorithm was used for fixation and saccade segmentation. Different methods for features selection were implemented to select the top features i.e. Fisher

Score, ANOVA, and Chi as seen in Table 5.11, Table 5.14 and **??** respectively. The best dyslexia prediction accuracy achieved in this study was **93.8** $\pm$ **0.2** % by using only two features that were selected by Chi method and logReg classifier. When using the top two features from Fisher Score and ANOVA methods, the best accuracy was 92.9 % $\pm$ 0.3 % with logReg classifier in both cases as can be seen in Table 5.13.

Table 5.13: Dyslexia prediction accuracy with standard error of the mean for different classifiers with different methods of feature ranking and their top two features of the total number of 185 participants, 97 dyslexic and 88 non-dyslexic (train to test ratio 90:10).

#Features (selection)	logReg %	RBFN %	NB %	SVM %	RF %
Fisher	**92.9** $\pm$ **0.3**	91.6 ± 0.3	92.7 ± 0.3	92.6 ± 0.3	87.8 ± 0.3
ANOVA	**92.9** $\pm$ **0.3**	89.6 ± 0.3	92.9 ± 0.3	92.6 ± 0.3	87.8 ± 0.3
Chi	**93.8** $\pm$ **0.2**	85.2 ± 0.4	93.5 ± 0.2	93.8 ± 0.2	89.8 ± 0.3

Finally, the same classifiers hyper-parameters values as in Section 5.1.1 were used for dyslexia prediction except for RBFN K value was 32.

Table 5.14: Features ranked by ANOVA Score for Dyslexia prediction

Fixation features	ANOVA Score	Saccade features	ANOVA Score
1) Total duration of Fix	401.9	1) Mean duration of Sacc	118.9
2) Number of Fix	295.4	2) Total angular velocity	107.7
3) Distance with previous Fix	131.3	3) Dispersion	106,2
4) Total acceleration	84.1	4) Maximum of angular acceleration	106.1
5) Fixation ratio	44.6	5) Maximum of angular velocity	103.6
6) Mean duration of Fix	44.0	6) STD of X direction	95.7

5.1.3.2 Gender Prediction in Isolated Groups (CLF2 and CLF3)

For gender prediction in dyslexic (Table 5.4) and non-dyslexic groups (Table 5.3) respectively, we pick the best performing ML classifier which was logReg with its corresponding top four features.

From these three classifiers, a voting and weighting based hierarchical classifier for gender prediction was built. The idea behind a voting based classifier is to take binary decisions using the three classifiers whether a participant is non-dyslexic or dyslexic (outcome of CLF1) and consequently non-dyslexic male or healthy female (outcome of CLF2) or dyslexic male or dyslexic female (outcome of CLF3). The non-

dyslexic and dyslexic males were then grouped into a common male category and non-dyslexic and dyslexic females were grouped into a common female category respectively. These predictions were then compared with true labels in order to compute the accuracy of this classifier as explained before. Figure 5.5 shows a flowchart describing the idea of the voting based hierarchical classifier derived using CLF1, CLF2, and CLF3.

A weighted hierarchical model was also implemented. It is observed that it brought no benefit compared to the voting based model (see Table 5.15). In the weighting based classifier, the idea was to give both entities (non-dyslexic and dyslexic probabilities in CLF1) the opportunity of collaborating on the decision making for the gender prediction in CLF2 and CLF3. This was done by ranking the relative importance of each decision through the use of weights (probabilities in CLF1) based on the individual performance of each entity. The probabilities for a participants to be male (P_M) and female (P_F) is given by:

$$P_M = P_{ND} * P_{NDM} + P_D * P_{DM} \tag{5.2}$$

$$P_F = P_{ND} * P_{NDF} + P_D * P_{DF} \tag{5.3}$$

where P_{ND} and P_D denote the probabilities for a participant to be non-dyslexic and dyslexic respectively (outcome of CLF1), P_{NDM} and P_{NDF} are the probabilities for being non-dyslexic male and non-dyslexic female (outcome of CLF2), P_{DM} and P_{DF} are the probabilities for being dyslexic male and dyslexic female (outcome of CLF3). If the resulting probability of being male is greater than probability of being female i.e. $P_M > P_F$, then predicted label is assigned as male and vice-versa.

The results from the voting based and weighting based hierarchical classifiers are reported in Table 5.15. In general, both voting and weighting based hierarchical classifiers perform better than the non-hierarchical classifier of Table 5.8. It can be observed that the voting based classifier performs better than the weighting based classifier for this hierarchical classification problem. Moreover, the accuracy peaks for the largest mixed group achieving an accuracy of $\mathbf{63.0 \pm 0.3}$ % which almost reaches our top accuracy for isolated groups of Section 5.1.1 (i.e. $\mathbf{63.8 \pm 0.5}$ % in non-dyslexic group and $\mathbf{60.7 \pm 0.6}$ % in dyslexic group).

Figure 5.4 compares the accuracy of gender prediction in a mixed group of participants with varying proportions of healthy and dyslexic participants between hierarchical (voting based) and non–hierarchical classifiers based on the results reported in Table 5.15 and Table 5.8. Left extreme of the graph shows non-dyslexic only group and right extreme of the graph shows dyslexic only group and in the middle we can see equal number of healthy and dyslexic participants. One can observe that the accuracy of the non-hierarchical classifier experiences a larger dip as compared to

Table 5.15: Accuracies with standard error of the mean for different classifiers to predict gender in mixed groups with varying proportions of non-dyslexic and dyslexic in hierarchical classifier.

Dyslexic (DM, DF) / Non-Dyslexic (HM, HF)	Voting accuracy %	Weighted accuracy %
88 (69, 19) / 88 (69, 19)	63.1 ± 0.3	60.0 ± 0.3
38 (19, 19) / 38 (19, 19)	62.1 ± 0.4	59.5 ± 0.4
36 (18, 18) / 36 (18, 18)	62.1 ± 0.4	59.4 ± 0.4
8 (4, 4) / 28 (14, 14)	58.7 ± 0.5	57.4 ± 0.5
18 (9, 9) / 18 (9, 9)	57.9 ± 0.6	56.7 ± 0.6
28 (14, 14)/ 8 (4, 4)	59.6 ± 0.6	59.0 ± 0.6

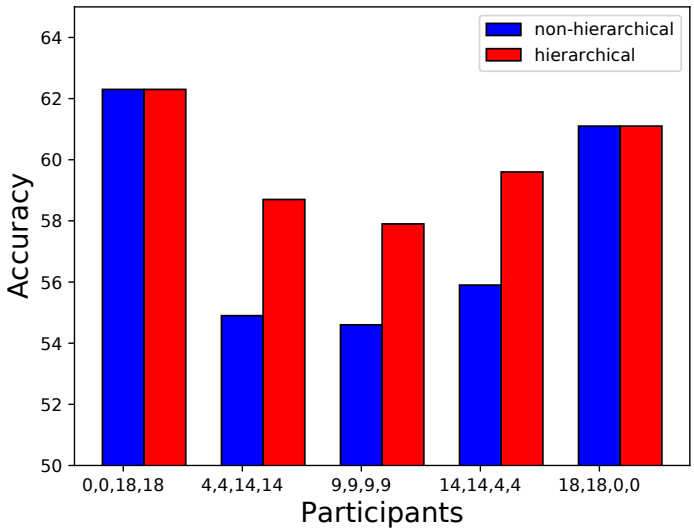


Figure 5.4: Comparison of gender prediction accuracy between hierarchical (voting based) and non–hierarchical classifiers with varying proportions of healthy and dyslexic participants in mixed groups.

the hierarchical classifier where the curve is more flat. This confirms our hypothesis that a hierarchical classifier performs better than the non-hierarchical classifier in a mixed group of participants for the gender prediction task.

5.1.4 Comparison with related works

Gender prediction, from eye movement data in adult, non-dyslexic participants has been reported in [Moss et al., 2012] and [Sargezeh et al., 2019]. The achieved accuracy in this study in the case of non-dyslexic participants was the same as in [Moss et al., 2012] over 10 runs and 6 % lower than [Sargezeh et al., 2019] which

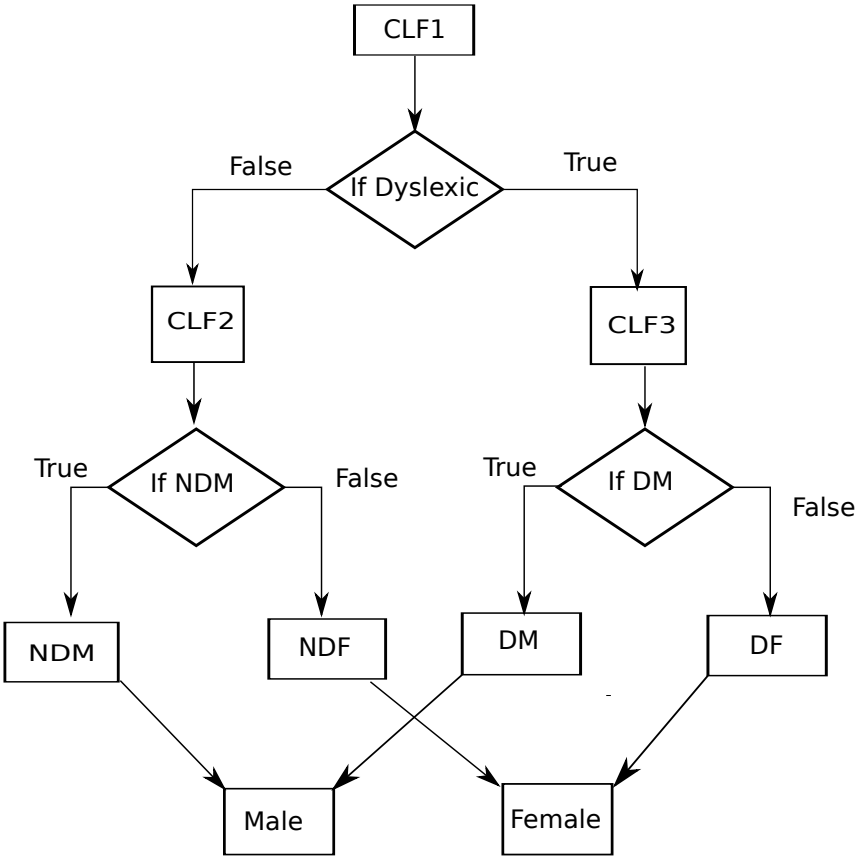


Figure 5.5: Voting based hierarchical classifier flowchart

uses 5 runs. It should be mentioned that these accuracies are not directly comparable because of the difference in available data, stimuli, and age of the participants. In [Moss et al., 2012] and [Sargezeh et al., 2019] the total recording time per participant is 400 s (52 non-dyslexic adults) and 240 s (45 non-dyslexic adults) respectively. Furthermore, while the works [Moss et al., 2012] and [Sargezeh et al., 2019] used images as stimuli, we use a reading stimulus. We are not aware of any previous gender prediction studies that use reading stimulus apart from [Emam and Youssef, 2012] which suggest that males read faster than females and [Obaidellah and Haek, 2018] which suggest that males can analyze algorithmic problems faster than females. Also, in contrast to [Miyahira et al., 2000b], our study demonstrates that gender difference in eye movement is already prominent in young children aged 9–10 in a mixed group of healthy and dyslexic participants.

For further improving dyslexia prediction accuracy, ten important features were selected including four fixation features namely minimum vertical acceleration (along y axis), variance in vertical velocity (along y axis), mean fixation duration, total fixation duration and six saccade features namely, path length (the length of path

Table 5.16: Dyslexia prediction accuracy with different classifiers in different eye configurations and train to test ratio 90:10 to predict presence of dyslexia

Eye Configrations	RBFN	RF	SVM	NB
Left eye data	94.53%	93.38%	94.29%	92.73%
Average eyes data	**95.71 %**	93.17%	94.52%	92.60%
Combine eyes data	95.00%	94.30%	94.26%	93.30%
Right eye data	94.82%	**95.60 %**	93.85%	93.23%

traveled in screen during saccade), variance in horizontal velocity (along x axis), saccade amplitude, number of saccades, mean saccade duration, total saccade duration. These features were chosen heuristically by calculating the highest mean and lowest variance of the difference between the male and female participants for all the features. Table 5.16 shows the accuracy achieved by four different classifiers that we used (RBFN, RF, SVM, and NB) in the classification of dyslexic and healthy participants for different eye configurations with the above mentioned 10 features. A maximum accuracy of $95.71 \pm 4.39\,\%$ was achieved by the RBFN classifier when the features from both eyes were averaged together. This classifier performed almost as good as RF classifier which has accuracy $95.6 \pm 5.60\,\%$ in the case when features from right eye.

Comparing with the results reported in [Benfatto et al., 2016], (accuracy of 95.6 % $\pm$ 4.5 % used a 90:10 ratio of training to testing data of all the 185 participants), we achieved slightly better accuracy with less number of features than their work. From a qualitative perspective, our features (e.g. minimum vertical acceleration, variance in vertical velocity for fixations and variance in horizontal velocity for saccade) seem to capture the left to right reading task quite well. With changing the training ratio from 90 % to 80 %, the accuracy got reduced by 1 %. It was also noticed that increasing the number of features (greater than 10) does not affect the accuracy much i.e. it changes (increases or decreases) only in 1st decimal values.

5.1.5 Quantitative Observations on the Dyslexia Dataset

In this section certain quantitative differences are investigated in eye movements of males and females that are known from the literature [Emam and Youssef, 2012, Huang and Chen, 2016, Sargezeh et al., 2019]. In this study, these differences were confirmed to be also present in the used dataset of prepubescent children (the previous works study older participants). All available data from the used dataset were considered in four groups: non-dyslexic and dyslexic participants both separated again in male and female. The data suggested that non-dyslexic female children read slower with longer fixations in compari-

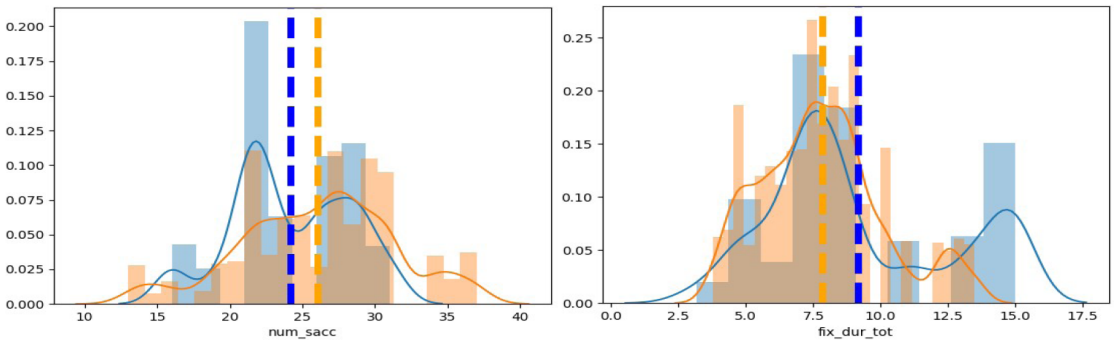
son to non-dyslexic male children (see Table 5.17). The same was previously observed in adult participants [Emam and Youssef, 2012, Miyahira et al., 2000c] and [Huang and Chen, 2016] in high school students aged 14–15. This depicts that this kind of difference in gender develops quite early. Also, dyslexic males explored a little slower than females, i.e., the mean saccadic velocities are lower with higher fixation duration which was also noted in [Suroya and Al-Samarraie, 2016]. In both non-dyslexic and dyslexic participants were observed that females have

Table 5.17: Mean and standard derivation of total fixation duration, saccade velocity, and number of saccades

	Mean total fix. duration (ms)	Mean saccade vel. (mm/s)	Mean number of saccades.
NDM	7.8 ± 2.2	1704 ± 680	24.1 ± 4.0
NDF	9.2 ± 3.4	1520 ± 497	26.1 ± 5.2
DM	14.2 ± 1.9	1039 ± 417	28.1 ± 5.9
DF	13.6 ± 1.5	1152 ± 293	31.4 ± 4.8

a slightly higher number of saccades than males. It could indicate that females demonstrate a more exploratory behavior than males. The same is observed in [Sargezeh et al., 2019, Heisz et al., 2013] for adult participants.

As shown in Figure 5.6a for the saacade number and Figure 5.6b total of fixation duration for the male and females of the non-dyslexic group, the data was not normally distributed. Therefore, in addition to the calculations in Table 5.17, Wilcoxon signed-rank test was implemented. We can see from the same plots that there were differences between males and females in both both saccade number (P-value of wilcoxon-test = 0.004) and total of fixation duration (P-value of wilcoxon-test = 0.033).



(a) Saccade number (b) Total of fixation duration

Figure 5.6: Saccade number and total of fixation duration for males and females in the non-dyslexic group (not normal distributed data)

5.2 Gender Prediction with the VST Dataset

The results in Section 5.1 published as [Zaidawi et al., 2020] showed that, gender prediction is possible in prepubsecent children with approximately **64 %** accuracy using machine learning. As a follow up of this work, another data set was used for this task that consisted of 58 non-dyslexic adult participants i.e. the VST dataset (see Section 3.1.2). The IVT algorithm with the parameters of VT = 150 and MFD = 100 ms were used for gender prediction in the VST dataset. The ANOVA method was used to rank the top features as described in Section 3.3.2. An increasing number of top features up to twelve were used with all the gender prediction experiments. A maximum of twelve features were considered since the accuracy started to drop beyond twelve (e.g. of top 12 features with using 12 seconds trajectory length Table 3.7). All the five ML classifiers (RBFN, SVM, RF, NB, and logReg) were used for this experiment. In all the experiments, cross-validation with 300 runs was performed. In each run, 48 participants (24 males and 24 females) were randomly chosen from all available participants in order to have a balanced number of males and females. Those selected participants were randomly split into a training set ($80\,\%$) and a test set ($20\,\%$).

For the experiments of gender prediction, the trajectory was further divided into segments, e.g., 12 seconds of the trajectory (less than 12 seconds was not possible since some participants does not have saccade), 20 seconds, 40 seconds, 1 minute, 2 minutes. These experiments were performed by taking gaze trajectories from the beginning to end and from end to beginning of the trajectory to study the fatigue or the attention span. Again the ANOVA method was used in each segment of the data and the top features were considered in the experiments based on these segments. Tables 5.19, 5.20 and 5.21 are examples of the top 12 features in Session 1 with a segment of 12 seconds from the end of the trajectory, Session 4 with a segment of 12 seconds from the beginning of the trajectory, and Session 4 with a segment of 20 seconds from the beginning of the trajectory respectively.

Finally, as mentioned in Section 3.4, some of the classifiers' parameters were retuned by performing grid search method on the train data. The hyper-parameter values that were used in the gender prediction experiments in the VST dataset are listed below. See Figure 5.7 which show an example of the RBFN K parameter tuning in gender prediction using grid search. The best value for the RBFN K parameter in Figure 5.7 was ten. As can be seen this value gave the highest test accuracy where the smallest difference between the test accuracy and its train accuracy was achieved.

1. RBFN: $K = 3$ was used with 12, 20, and 40 seconds of the trajectory length, while $K = 10$ was used with 60 and 120 seconds.

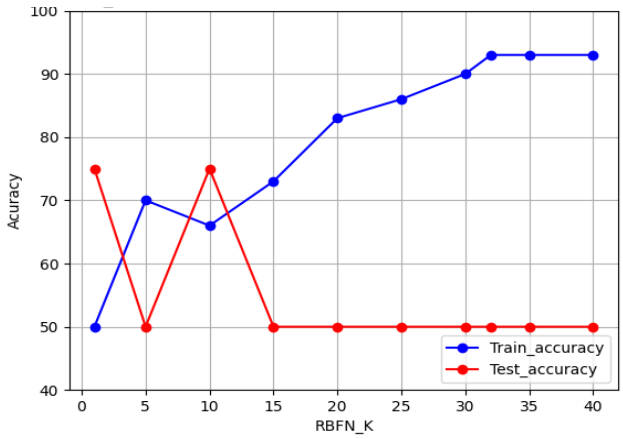


Figure 5.7: RBFN K parameter against the accuracy as example of tune K parameter (60 sec trajectory length of VST data)

2. SVM: "gamma" value of 0.2 and "C" value of 1 were used with 12 seconds, "gamma" value of 0.01 and "C" value of 4 were used with 20 seconds, "gamma" value of 1 and "C" value of 0.01 were used with 40 seconds, "gamma" value of 1.2 and "C" value of 5 were used with 60 seconds, and "gamma" value of 1.3 and "C" value of 0.2 were used with 120 seconds.

3. RDF: "Max-depth" value of 3 was used with 12, and 20 seconds, while "Max-depth" value of 5 was used with 40, 60, and 120 seconds. "N-estimators" value of 300 was used with all the trajectory segments.

These hyperparameters were selected based on a grid search using the train data as shown in Figure 5.7 for RBFN as an example.

The following subsections describe different experiments that were implemented with the VST dataset:

5.2.1 Effect of Length of the Trajectory

As mentioned in Section 5.2, the VST dataset provides long gaze trajectories of more than 4 minutes. Therefore, the effect of the trajectory length towards the gender prediction was investigated. We would like to know the minimum trajectory length that is sufficient for gender prediction with a high accuracy. Hence, variations in lengths of trajectories for gender prediction were considered. Also, these sub-trajectories were taken from the start and the end of the recording to additionally study the effect of possible fatigue on the accuracy. The VT = 150 and MFD = 100 ms parameters were used for the IVT algorithm in all the gender prediction experiments of this data.

Let us first consider the experiments performed on the initial segments of the trajectories. The best accuracies were achieved on Session 1 and Session 4 by using

only the first 12 or 20 seconds, respectively. The best accuracy achieved for Session 1 was **62.4** $\pm$ **0.80** % using the LogReg classifier in the 12 second segment. For Session 4 the best accuracy was **64.2** $\pm$ **0.80** % using the NB classifier in the 20 second segment (the results are shown in Table 5.18). All the achieved accuracies were obtained using the four top features of these individual cases. For example, the feature ranking for 12 seconds of Session 1 and for 20 seconds of Session 4 which corresponds to the best reported accuracies of these two Sessions are reported in Table 3.7 and Table 5.21 respectively from which the top four features were selected. Remaining such tables are computed for 12s, 20s, 40s, 60s and 120s trajectory lengths and are skipped here for brevity.

Table 5.18: Accuracies with standard error of the mean using the top four fullR features over 300 runs using 48 participants in VST (using Session 1 and 4, top 4 features, VT = 150) from the start of the trajectory lengths

Duration	Se. No.	LogReg %	SVM %	RF %	NB %	RBFN %
12 second	1	**62.4** $\pm$ **0.8**	53.5 $\pm$ 0.7	55.6 $\pm$ 0.8	60.9 $\pm$ 0.8	57.4 $\pm$ 0.8
	4	53.9 $\pm$ 0.9	50.4 $\pm$ 0.9	**56.5** $\pm$ **0.9**	55.2 $\pm$ 0.9	51.5 $\pm$ 0.8
20 second	1	55.2 $\pm$ 0.9	48.0 $\pm$ 0.5	51.7 $\pm$ 0.8	**59.0** $\pm$ **0.8**	50.8 $\pm$ 0.7
	4	62.3 $\pm$ 0.8	55.9 $\pm$ 0.7	61.2 $\pm$ 0.8	**64.2** $\pm$ **0.8**	59.0 $\pm$ 0.8
40 second	1	47.9 $\pm$ 0.9	51.9 $\pm$ 0.5	**58.2** $\pm$ **0.8**	54.3 $\pm$ 0.7	54.3 $\pm$ 0.7
	4	54.0 $\pm$ 0.8	51.4 $\pm$ 0.5	**56.0** $\pm$ **0.8**	51.3 $\pm$ 0.6	52.5 $\pm$ 0.6
60 second	1	**60.6** $\pm$ **0.9**	53.3 $\pm$ 0.7	56.8 $\pm$ 0.7	54.7 $\pm$ 0.6	55.0 $\pm$ 0.8
	4	55.0 $\pm$ 0.8	51.9 $\pm$ 0.5	**56.1** $\pm$**0.8**	50.3 $\pm$ 0.6	54.2 $\pm$ 0.5
120 second	1	60.3 $\pm$ 0.9	55.5 $\pm$ 0.7	58.3 $\pm$ 0.7	51.2 $\pm$ 0.4	**61.0** $\pm$ **0.8**
	4	**55.6** $\pm$ **0.8**	51.7 $\pm$ 0.4	50.1 $\pm$ 0.5	52.1 $\pm$ 0.6	51.5 $\pm$ 0.5

Table 5.19: Features ranked by ANOVA Score for VST dataset using session 1 and 12 seconds trajectory length (end)

Fixation features	ANOVA Score	Saccade features	ANOVA Score
1) Kurtosis of angular acceleration	3.85	1) Mean of acceleration Y	8.89
2) Distance with previous Fix	3.78	2) SD of Y	6.73
3) Kurtosis of angular velocity	3.66	3) Mean of velocity Y	3.68
4) Skewness of velocity Y	3.18	4) Maximum of velocity X	3.58
5) Minimum of velocity X	2.69	5) Dispersion	3.08
6) Skewness of angular acceleration	2.66	6) Maximum of acceleration X	2.62

Table 5.20: Features ranked by ANOVA Score for VST dataset using session 4 and 12 seconds trajectory length (start)

Fixation features	ANOVA Score	Saccade features	ANOVA Score
1) Distance with previous Fix	7.35	1) SD of angular acceleration	14.88
2) Mean duration of Fix	4.36	2) SD acceleration Y	12.79
3) Dispersion	4.19	3) Variance of acceleration Y	12.71
4) Fixation ratio	4.17	4) Maximum of acceleration Y	11.70
5) Variance of acceleration Y	4.02	5)SD of velocity Y	11.69
6) SD of acceleration Y	3.87	6) Minimum of angular acceleration	11.49

Table 5.21: Features ranked by ANOVA Score for VST dataset using Session 4 and 20 seconds trajectory length (start)

Fixation features	ANOVA Score	Saccade features	ANOVA Score
1) Distance with previous Fix	5.75	1) SD of acceleration Y	7.43
2) Fixation ratio	4.97	2) SD of angular acceleration	7.39
3) Variance of acceleration Y	4.48	3) Maximum of velocity Y	7.25
4) Maximum of angular velocity	4.47	4) Maximum of acceleration Y	6.06
5) Variance of velocity Y	4.44	5) SD of velocity Y	5.92
6) Variance of angular velocity	4.22	6) Minimum of acceleration Y	5.87

5.2.2 Effect of Fatigue or Attention span

The hypothesis of this section is that participants at the end of the trajectory get tired then behave more "natural". This can increase gender prediction accuracy. Therefore, the following experiments were performed to study the effect of fatigue or losing attention on gender prediction. First by repeating the same gender prediction experiments of Section 5.2.1 using Session 1 with a variety of trajectory lengths i.e. 12s, 20s, 40s, 60s, and 120s but taking them from the end of the trajectory.

Before that, their ANOVA top features for each trajectory segment were calculated (For example, the feature ranking for 12 seconds of Session 1 which corresponds to the best reported accuracy of this Session is reported in Table 5.19 from which the top four features were selected). Again the remaining such tables are computed for 20s, 40s, 60s and 120s trajectory lengths and are skipped here for brevity. As men-

tioned above, the best accuracies were achieved again with using the top four features of these different trajectories and the results showed that mostly the accuracies were higher when the data segment was taken from the end of the trajectory. As can be seen in Table 5.22, the achieved accuracies were better than the results reported in Table 5.18 where the segments were taken from the start of the trajectory. If we consider the 12s segments which give the highest accuracies for Session 1, then the end segment gives 69.9 $\pm$ 0.8 % whereas the beginning segment gives 62.4 $\pm$ 0.8 %. This is a surprisingly large difference of 7.5 %. Thus, fatigue as loss of attention has a large positive impact on gender prediction via eye movements.

Table 5.22: Accuracies with standard error of the mean using the top four features over 300 runs using 48 participants in VST (using Session 1 , top 4 features, VT = 150) from the end of the trajectory lengths

Duration	LogReg %	SVM %	RF %	NB %	RBFN %
12 second	**69.9 $\pm$ 0.8**	58.3$\pm$0.7	62.2$\pm$0.8	56.4 $\pm$ 0.6	58.3 $\pm$ 0.8
20 second	**68.1 $\pm$ 0.8**	57.2 $\pm$ 0.7	61.8 $\pm$ 0.8	58.0 $\pm$ 0.6	58.5 $\pm$ 0.8
40 second	**59.0 $\pm$ 0.9**	50.7 $\pm$ 0.5	58.9 $\pm$ 0.8	58.3 $\pm$ 0.7	54.5 $\pm$ 0.6
60 second	56.0 $\pm$ 0.9	51.5 $\pm$ 0.5	**57.7 $\pm$ 0.8**	54.8 $\pm$ 0.6	53.3 $\pm$ 0.6
120 second	56.7 $\pm$ 0.9	56.9 $\pm$ 0.7	60.1 $\pm$ 0.8	57.9 $\pm$ 0.7	**62.6 $\pm$ 0.7**

A second experiment was conducted to study this increase in accuracy, by segmenting the full trajectory of Session 1 into 12 segment sequences of 20 seconds each. Then the features were ranked by their ANOVA scores for each of the 12 trajectory segments. Next, in this experiment the top six features were used to perform gender prediction over 300 runs in each segment. The results (see Table 5.23 and Figure 5.8) showed again that the accuracies were mostly higher when the segment was taken from the end of the trajectory. The average accuracy of the first six trajectory segments is 55.3 %, while the average accuracy of the last six trajectory segments is 60.5 %

This increase in accuracy can be due to the blinking increasing at the end of the trajectory. One of the visual behaviors that can be easily observed is blinking when a participant is fatigue [Benedetto et al., 2011]. Many studies showed that increased fatigue results in longer and more frequent blinking. Some studies reported other measures related to increased mental fatigue, such as blink interval [Di Stasi et al., 2012, Schleicher et al., 2008]. Most studies found that blinking was the best indicator of fatigue compared to other eye-tracking measures including oculomotor-based metrics such as saccadic and fixation duration [Schleicher et al., 2008, Yamada and Kobayashi, 2018]. Another

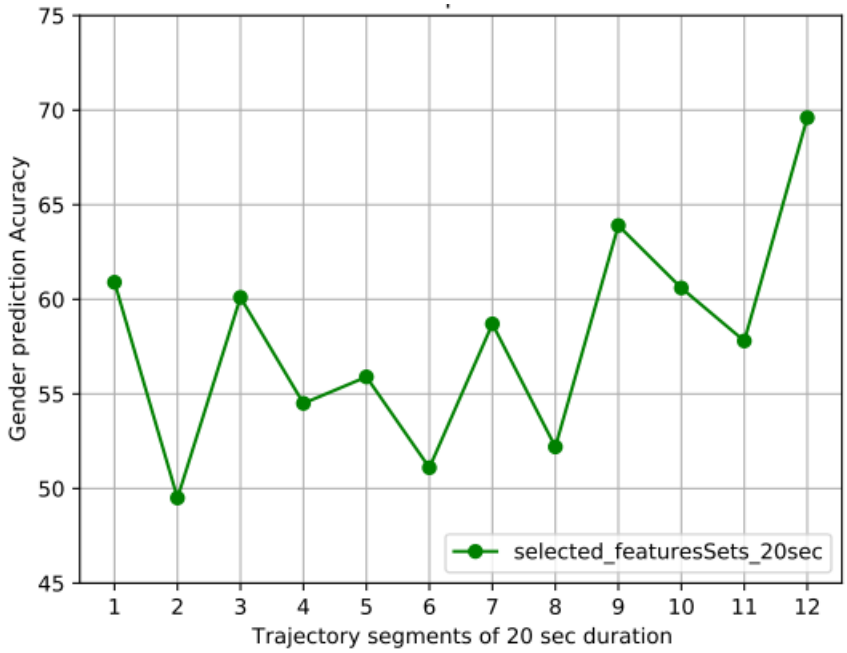


Figure 5.8: Study the effect of increasing the blink on gender prediction accuracy (by segmenting the trajectory into 12 sequences segments of 20 seconds, with using selected the top six features for each data segment).

study [Schleicher et al., 2008] suggested that fixation duration showed no relation to fatigue. Therefore, an analysis on the blink information was done. The blink segments were extracted from the NaN information using a duration threshold between 80–500 ms, since the duration of more than 500 ms are considered as micro-sleeping [Schleicher et al., 2008, Wang et al., 2011] and less than 80 ms can be device faults or other unknown reasons. Then, the blink duration mean and the standard deviation for each user was calculated in each trajectory segment. As shown in Table 5.24 the blink duration are different among the participants and were always significantly more in females than in males.

The results from the above experiments and the analysis may suggest that gender prediction from eye tracking data is working better when the participants have fatigue as they switch from their natural behavior. Given the nature of the visual search task, the gender difference in eye movements may not be evident when the participants are focused on solving the task and are not tired.

Finally, removing the invalid data (NaNs) from the data was showing a decrease in the accuracy which confirmed that these parts of the data (blinking) can encode some information about the participants as reported in [Kröger et al., 2020]. The results of removing the invalid data of using trajectory length of 12 sec from the end of the trajectory: 67.0 $\pm$ 0.9 for logReg, 53.3 $\pm$ 0.6 for SVM, 60.8 $\pm$ 0.8 for RF, 59.4 $\pm$ 0.7 for NB, and 60.2 $\pm$ 0.7 for RBFN. The best accuracy from the end was decreased

Table 5.23: Accuracies with standard error of the mean using 48 participants in VST (using Session 1 and VT = 150) to study the effect of fatigue (increase of the blink) on gender prediction accuracy (by segmenting the trajectory into 12 segment sequences of 20 seconds, with using ANOVA with top six features over 300 runs for each data segment).

Duration	Best Classifier	Accuracy %
1 (20 sec)	NB	60.9 ± 0.7
2 (20 sec)	NB	49.5 ± 0.5
3 (20 sec)	logReg	60.1 ± 0.8
4 (20 sec)	RBFN	54.5 ± 0.8
5 (20 sec)	logReg	55.9 ± 0.9
6 (20 sec)	SVM	51.1 ± 0.6
7 (20 sec)	RF	58.7 ± 0.8
8(20 sec)	RF	52.2 ± 0.9
9 (20 sec)	RF	63.9 ± 0.7
10 (20 sec)	logReg	60.6 ± 0.8
11 (20 sec)	logReg	57.8 ± 0.8
12 (20 sec)	logReg	69.6 ± 0.8

by $2.9\,\%$ (from $69.9\,\%$ to $67.0\,\%$), to compare see the results of 12 seconds from the end of the trajectory in Table 5.22.

5.2.3 Comparison with Related Works

In addition to the results of gender prediction in Section 5.1.4, the VST dataset was used also with such study by using the six state-of-the-art features reported in [Sargezeh et al., 2019]. As can be seen in Figure 5.9 our original choice of four top features (fullR-beginning and fullR-end respectively, fullR the features computed from the beginning or the end of the full sub-trajectory region of the stimuli) ranking based on ANOVA produced better accuracies than the features used in the study of Sargezeh [Sargezeh et al., 2019] with our best classifier which is logReg.

Comparing the best accuracy of using 12 seconds trajectory length from the beginning of the trajectory (see Table 5.18), from end of the trajectory (see Table 5.22), and with using the state of art features that were again applied on 12 seconds trajectory length from both beginning and the end of the trajectory (see Table 5.25), one can see from Figure 5.9 that the best case accuracy is found with using logReg and with ANOVA based four top features was $10.5\,\%$ higher (with 12 s of data from the end of

Table 5.24: The mean and standard deviation of the blink duration in seconds from the start and the end of the trajectory lengths for both males and females

Duration	Trajectory	Males	Females	All Participants
12 second	beginning	0.16 ± 0.03	0.32 ± 0.06	0.26 ± 0.03
	end	0.26 ± 0.05	0.48 ± 0.08	0.39 ± 0.05
20 second	beginning	0.26 ± 0.05	0.54 ± 0.09	0.42 ± 0.06
	end	0.40 ± 0.08	0.70 ± 0.11	0.58 ± 0.07
40 second	beginning	0.81 ± 0.16	1.38 ± 0.20	1.14 ± 0.15
	end	0.77 ± 0.16	1.30 ± 0.20	1.07 ± 0.14
60 second	beginning	1.13 ± 0.23	1.89 ± 0.32	1.57 ± 0.20
	end	1.20 ± 0.24	2.00 ± 0.34	1.66 ± 0.22
120 second	beginning	2.31 ± 0.47	3.45 ± 0.60	2.97 ± 0.51
	end	2.40 ± 0.48	3.80 ± 0.76	3.18 ± 0.41
180 second	beginning	3.50 ± 0.71	5.21 ± 0.89	4.49 ± 0.59
	end	3.60 ± 0.72	5.33 ± 0.91	4.58 ± 0.94

the trajectory) than the accuracy of using the state of art features which was using RBFN classifier (compare $69.9\,\%$ using logReg with $59.4\,\%$ using RBFN). While with taking the 12 seconds from the beginning of the trajectory, still using our four features were $1.5\,\%$ higher than the state of art features (Compare $62.4\,\%$ using logReg with $60.9\,\%$ using RBFN). Comparing across the two data sets, our best case accuracy is almost similar ($69\,\%$ vs $70\,\%$ reported in [Sargezeh et al., 2019]).

This study agrees with [Sargezeh et al., 2019] that there exists a gender difference in the eye movement behavior of humans. These results can not be directly compared with [Sargezeh et al., 2019] as the used recorded time in our dataset is only 12 seconds and in their data set was 240 seconds which is a significant difference in the amount of data that were available to these studies. Further, the stimuli used in the two works were different. In [Sargezeh et al., 2019], they used images as stimuli, while visual search stimulus was used in this study. Further, the total recording time per participant was 240 seconds, on the other hand only 12 seconds of data were used in this study.

5.2.4 Statistics and Quantitative Observations

The best accuracy from the beginning of the trajectory that was considered with only 20 seconds for Session 4. The data (see Table 5.26) suggests that path lengths are longer for females as compared to males which was also noted in [Sargezeh et al., 2019]. Females have shorter RFDSD (ratio fixation duration to

Table 5.25: Accuracies with standard error of the mean using the six state of art features over 300 runs using 48 participants in VST (using Session 1, VT = 150) from the beginning and the end of the trajectory lengths

Duration	Trajectory	LogReg %	SVM %	RF %	NB %	RBFN %
12 second	beginning	58.7 ± 0.8	60.8 ± 0.8	59.5 ± 0.8	58.2 ± 0.9	60.9 ± 0.8
	end	44.8 ± 0.9	55.9 ± 0.9	52.2 ± 0.9	51.1 ± 1.0	59.4 ± 0.9

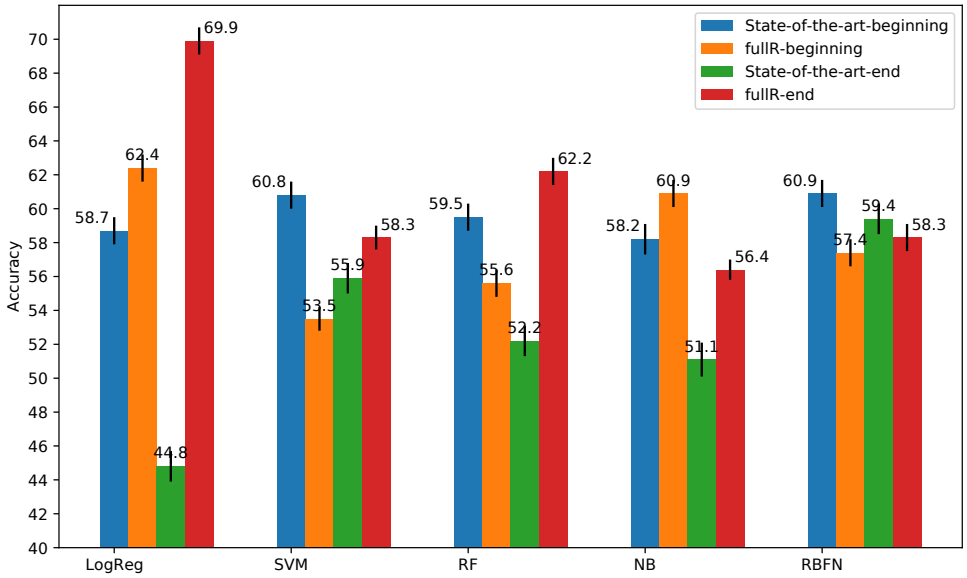


Figure 5.9: Comparison with using the features of the state-of-art and behavior of five classifiers

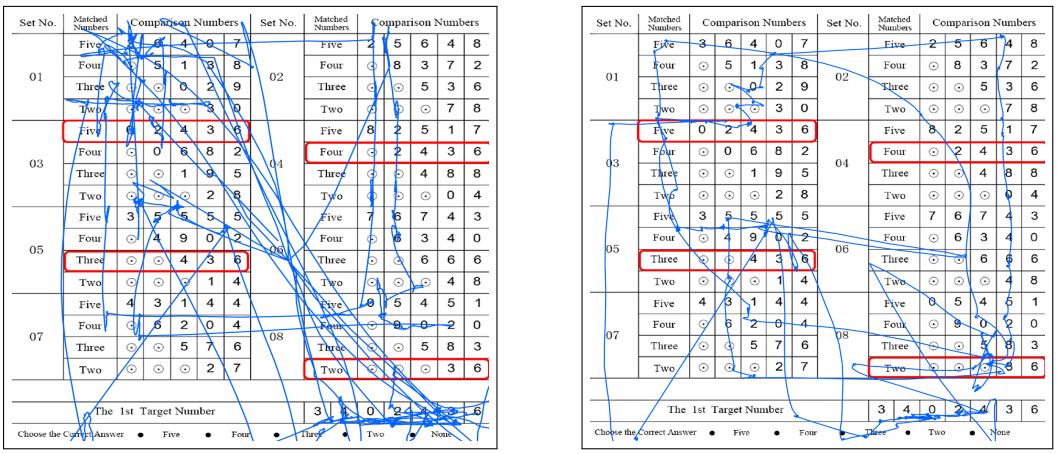
saccade duration) as compared to males. Distance to previous fixation is the highest ranked fixation feature by ANOVA, the finding was that males have more mean distance to previous fixation when compared with females by 1.2 times. According to these observations, a speculation was that females have more exploratory gaze behavior. The same has been previously observed in [Sargezeh et al., 2019, Zaidawi et al., 2020]. An example of the stimuli used and the eye trajectory path of one male and one female participant can be seen in Figure 5.10. For this pair of participants, clearly the female shows more exploratory gaze behavior than the male.

5.3 Gender Prediction with the GOF Dataset

In addition to the Dyslxia and VST datasets, another stimuli was considered for gender prediction. The GOF dataset consists of a huge number of 378 partici-

Table 5.26: Mean and standard error of the mean values for males and females - VST dataset

Gender	Path length	RDFSD	Distance previous fixation
M	657.6 ± 283.6	8.5 ± 1.5	211.8 ± 40.4
F	1055.8 ± 475.9	6.6 ± 1.0	177.4 ± 29.6



(a) Female participant

(b) Male participant

Figure 5.10: VST dataset stimuli and their gaze points trajectory for 20 seconds from the start. (a) Participant "sb" searching for a number and (b) Participant "yh" searching a for number

pants from different age groups. This motivates us to use this data to study the effect of different factors on gender prediction accuracy e.g. stimuli, age, length of the trajectory, and different sets of features. In this study, two sets of features were used. The first set of features was computed for each region of interest (ROI) individually. As described in Section 3.3.2 and inspired by [Coutrot et al., 2016, Sammaknejad et al., 2017, Baron-Cohen, 2002] four regions were used: left eye, right eye, nose, and mouth. These ROI were defined by hand as rectangles. When computing a feature for the ROI e.g. nose, only the trajectories inside the nose region were considered. Figure 3.8 shows ROIs defined by hand on two stimuli. The second set of features was computed using the full trajectory region of the stimuli (fullR) as described previously in Section 3.3.2.

The ANOVA method was used to rank the top features in both sets of features and incremental numbers of top features up to twelve were used in all the experiments (e.g. of fullR top features Table 3.8 and ROI Table 3.9). Again a maximum of twelve features were considered since the accuracy started to drop when we increased the number of features beyond twelve. All the five ML classifier (RBFN, SVM, RF, NB, and logReg) were used. Parameters tuning for RF, SVM, and RBFN was performed

using grid search. The ratio of females to males was always balanced choosing them equally and randomly at each run. In all the experiments, a cross validation with 50 runs was performed as the standard deviation in the accuracy metric becomes stagnant with this number of runs. In each run the participants were randomly chosen. Those selected participants were randomly split into a training-set ($80\,\%$) and a test-set ($20\,\%$) for each new run. For the experiments, the trajectory was further divided into segments, i.e. 12 seconds of the trajectory (less than 12 seconds was not possible because some participants do not have a single fixation). Similarly, 20 seconds, 40 seconds, 60 seconds, and 120 seconds segments were used to study the effect of length of gaze trajectories on gender prediction.

Finally, as mentioned in Section 3.4, some of the parameters of the classifiers were re-tuned by performing a grid search on the train data; this was done in different age groups and with different features sets (fullR and ROI features). Below is an example of the tuned hyper-parameter values that were used in two age groups (20–72 and 20–30 years) and with using the ROI features (all segments were taken from the beginning of the trajectory).

1. RBFN: "K" value of 1 was used with 12, 20, 40, and 60 seconds of the trajectory length, while RBFN "K" value of 4 was used with 120 and 120 seconds.

2. SVM: "gamma" value of 0.01 and "C" value of 0.01 were used for 12 seconds, and "gamma" value of 0.001 and "C" value of 0.01 were used with 20, 40, 60 and 120 seconds

3. RDF: "Max-depth" value of 3 was used for 12, while "Max-depth" value of 2 was used with 20, 40, 60, and 120 seconds. "N-estimators" value of 200 was used with 12, 20, and 60 seconds while "N-estimators" value of 100 was used with 40 and 120 seconds of the trajectory length.

The following sections describe different experiments that were implemented with the GOF dataset.

5.3.1 The Effect of Different Feature Sets and Trajectory Lengths

As mentioned in Section 5.3, the trajectory was divided into segments in all the experiments inorder to study the effect of trajectory length on gender prediction. Furthermore, the mentioned two types of the features were used in all the experiments (fullR and ROI). The 378 participants were divided into different age groups of 20–72, 20–30 and 31–50 for all the experiments. The best accuracies were achieved with using only the top two features determined by their ANOVA score as shown in Table 5.27. The gender prediction results in the three age groups are shown in Table 5.28. In

the age group 20–72, the fullR features achieve the best accuracy of **63.7** ± **0.60** % using the LogReg classifier in the 120 seconds segment and ROI features achieve an accuracy of **60.9** ± **0.70** % using the SVM classifier in the 120 seconds segment. In the age group 20–30, the fullR features achieve the best accuracy of **66.2** ± **0.90** % using the LogReg classifier in the 120 seconds segment and ROI features achieve an accuracy of **61.0** ± **0.90** % using the RBFN classifier in the 120 seconds segment. In the age group 31–50, the ROI features achieve the best accuracy of **64.5** ± **1.60** % using the LogReg classifier in the 12 second segment and fullR features achieve an accuracy of **54.5** ± **1.50** % using the RF classifier in the 60 seconds segment. These results show that gender prediction in this dataset is biased towards age group when working with the same set of features. With different sets of features in different age groups, it is possible to achieve similar prediction accuracy. For instance, fullR features yield better accuracy in the age group 20–30 and ROI features yield better accuracy in the older age group of 31–50.

Table 5.27: Top features for the GOF dataset giving the best accuracies with their ANOVA scores; fullR features in 120 seconds segment, ROI features in 12 seconds segment.

120 seconds	fullR features	ANOVA
Fix	Maximum angular velocity	24.40
Sac	Saccade ratio	39.29

12 seconds	ROI features	ANOVA
Fix	Number of fixations in nose region	7.75
Sac	Saccadic amplitude in nose region	14.03

Table 5.28: Best accuracies with standard error of the mean using two top fullR and ROI features for different age groups of the GOF dataset.

Age Group	Accuracy fullR	Classifier fullR	Duration fullR	Accuracy ROI	Classifier ROI	Duration ROI	Participant numbers
20–72	**63.7** ± **0.60** %	LogReg	120 s	60.9 ± 0.70 %	SVM	120 s	370 (185 F, 185 M)
20–30	**66.2** ± **0.90** %	LogReg	120 s	61.0 ± 0.90 %	RBFN	120 s	200 (100 F, 100 M)
31–50	54.5 ± 1.50 %	RF	60 s	**64.5** ± **1.60** %	LogReg	12 s	80 (40 F, 40 M)

5.3.2 Effect of Age on Gender Prediction

In order to compare the two age groups, other experiments were implemented by taking an equal number of participants in the age groups: 40 females and 40 males. The achieved accuracies are reported in Table 5.29. In the age group 20–30, we

achieve the best accuracy using the fullR features which is **64.2** $\pm$ **1.40** % using RF classifier. In the age group 31–50 years, we found out that ROI features gave us the best accuracy of **61.5** $\pm$ **1.8** % using the SVM classifier. The results of the age group 51–72 has not been included since there are only 20 participants in the experiments and the results cannot be relied upon. All the accuracies have been achieved using only the top two features. The features used for the best accuracy experiments are shown in Table 5.27. From these results, we see that the fullR features predicted the gender better in younger adults aged 20–30 years and ROI features predicted the gender better in older adults aged 31–50 years. This might suggest that participants in younger age group are more explorative than participants in the older age group which tends to focus on specific region of interests.

Table 5.29: Accuracies with standard error of the mean in different age groups with different feature type in 120 seconds segment

Age Group	Accuracy fullR	Accuracy ROI	Participants number
20–30	**64.2** $\pm$ **1.40** %	57.08 $\pm$ 1.60 %	80 (40 F, 40 M)
31–50	54.1 $\pm$ 1.60 %	**61.5** $\pm$ **1.80** %	80 (40 F, 40 M)

5.3.3 Comparison with Related Work

As mentioned earlier in this chapter, the study of [Sargezeh et al., 2019] has the highest accuracy which uses machine learning classifiers to predict gender using eye movement data. They reported an accuracy of **70** $\pm$ **13.22** % using 5 runs. The demographics of the study were as follows: 25 males and 20 females took part in the study aged 25–34 years. Indoor images were used as stimuli from the Change Blindness database[3].

In order to compare the features that introduced in this thesis with the state-of-the-art features, a similar experiments were performed using the same number of participants. Three types of features were used (the fullR features, ROI features, and the features of [Sargezeh et al., 2019]) for the experiments.

The state-of-the-art features included six features in total. These features are fixation duration, spatial density feature, RFDSD (ratio fixation duration to saccade duration), number of saccades, saccadic amplitude, and path length. The training to testing ratio was 80 % to 20 %. We ran the experiments using five Machine Learning classifiers. The experiments were run for five different runs (as in [Sargezeh et al., 2019]).

[3] https://search.bwh.harvard.edu/new/CBDatabase.html

The results are shown in Figure 5.11, the best accuracy of **72.5** ± **9.35** % was achieved with using the fullR-SOTA features (6 state of the art features based on the complete fixation and saccade trajectories) and the RF classifier. Similarly, the ROI features and the RF classifier were produced best accuracy of **62.5** ± **5.00** %. Finally, the best accuracy with using the state-of-the-art features and the RBFN classifier were **55.0** ± **9.10** %.

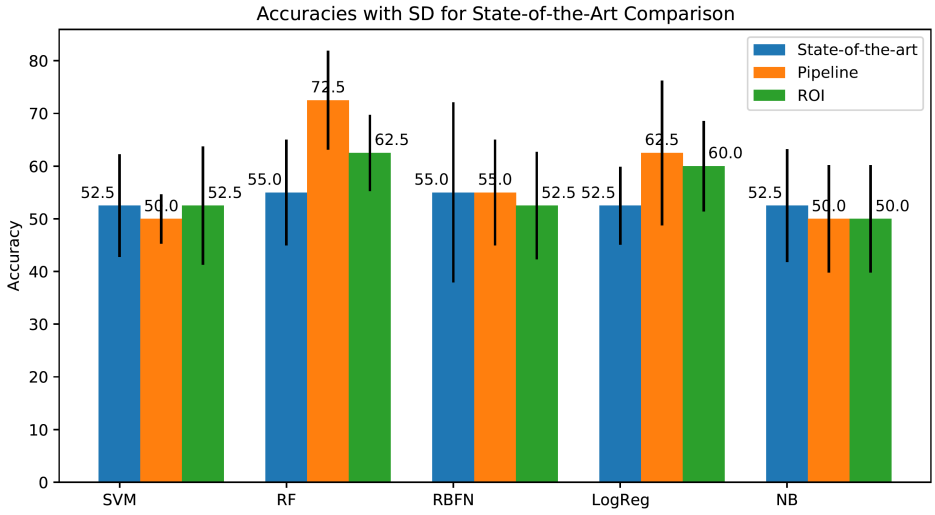


Figure 5.11: Comparison with state-of-the-art and behavior of five ML classifiers using three feature types and with using the GOF dataset

5.3.4 Optimizing Weights for Fixation and Saccade Classifiers

The gazing strategies of males and females are different as well as each person has unique gazing strategies [Emam and Youssef, 2012, Hall et al., 2010, Kaspar and König, 2012]. In the paper [Garza et al., 2016], they report that males and females differ in choosing their focus points on the stimuli. Hence, we can see a possibility that based on the fixation points the accuracy of gender prediction may increase. Also, men make shorter fixations and women tend to be more exploratory such that they have longer fixation movements [Kaspar and König, 2012]. In all the experiment, so far equal weights were used for both the fixation and the saccade classifiers. Therefore, experiments to optimize the weights were conducted to investigate whether that improves the gender prediction accuracy. In order to find the optimal weights for the two classifiers, the Nelder-Mead Method [Nelder and Mead, 1965] was used again.

It can be observed in Table 5.30, that higher fixation weights lead to better gender prediction accuracies in this dataset. The accuracy was increased by 1.1 % (from

66.2 % to 67.3 %). The optimizing of weights was also done for the state-of-the-art best accuracy experiment in the GOF dataset. The results showed that increasing the fixation weights to 0.781 from 0.5 increases the accuracy from 72.5 % to 77.5 % using the fullR features. It can be seen that the SD in these experiments is quite high.

Table 5.30: Our best accuracy of using 120 seconds and the top two fullR features with standard error of the mean using Nelder-Mead Method to optimize the fixation and saccade classifiers with using the GOF dataset.

Features type	Features Number	CLF weights Sac/Fix	Accuracy	No. of Runs
fullR	2	0.5/0.5	66.2 ± 0.90 %	50
	2	0.330/0.670	$\mathbf{67.3 \pm 0.98}$ %	50
fullR-SOTA	6	0.5/0.5	72.5 ± 4.10 %	5
in Section 5.3.3	6	0.219/0.781	$\mathbf{77.5 \pm 4.10}$ %	5

5.3.5 Statistics and Quantitative Observations

In this section some statistics and quantitative differences between the eye movements of females and males are discussed based on existing literature [Li et al., 2018b, Sargezeh et al., 2019, Zaidawi et al., 2020].

The first 120 seconds of the trajectory were used for the analysis in this section which was conducted to check whether the statistics of the data as mentioned in [Coutrot et al., 2016] was comparable after removing corrupt trajectories and capping the trajectory length at 120 seconds. They discovered that men had lower saccadic amplitudes and longer fixation duration than females. The analysis findings (see Table 5.33) confirm that the path lengths for females were longer and the saccadic amplitudes were also higher in females. It is important to be noted that the standard deviation for these features are high since the data is not normally distributed. Therefore, the Wilcoxon signed-rank test [Woolson, 2007] was performed to prove that these differences are significant. As shown in Table 5.31, the p-values for path length and saccade amplitude features are less than 0.005, the smaller the p-value, the more likely it is to reject the null hypothesis and that indicates there are significant differences in these features. Furthermore, the median of path lengths and saccade amplitudes are calculated. As shown in Table 5.34) the median is also confirm that the path lengths for females were longer and the saccadic amplitudes were higher in females. . This indicates that females were more explorative than males.

The Wilcoxon was performed again on the top two ROI and fullR features (see

Tables 5.32 and 5.31). The tests were carried out on various age groups between females and males to find out any significant difference. For the fullR features, the p-values for the two used features (for all the age groups) were always less than 3×10^{-7}. The p-values is less than the threshold of 0.005. The smaller the p-value, more likely it is to reject the null hypothesis. That indicates there are significant differences between the age groups, and it is more obvious in the age group of 20–30. All the p-values were smaller than 0.005 for the ROI features, except for the feature "Number of fixations in nose region" in the age group 20–30 (p-value = 0.084) and the feature "Saccadic amplitude in nose region" in the age group of 31–50 (p-value = 0.096). These results indicate that significant differences exist in number of fixation feature in the age group 20–72 and are more obvious in the age group 31–50 as the p-value in this age group is smaller than the age group 20–30.

Table 5.31: fullR features (p-value) - GOF dataset

Age Group	Path length	Saccadic amplitude	Maximum Angular velocity	Saccade ratio
20 - 72	**1.19E-69**	**6.29E-227**	**1.41E-10**	**0**
20 - 30	**1.85E-86**	**1.37E-193**	**8.31E-37**	**3.70E-240**
31 - 50	**4.03E-19**	**1.14E-26**	**2.90E-07**	**1.44E-57**

Table 5.32: ROI features (p-value) for the GOF dataset

Age group	No. of fixations (Nose)	Saccadic amplitude (Nose)
20 – 72	**0.0013**	**1.09E-05**
20 – 30	0.0839	**4.09E-05**
31 – 50	**0.0058**	0.0961

The mean number of fixations in females and males in different age groups is shown in Table 5.33. It can be seen that females, regardless of any age group had more fixations on the left eye of the stimuli. On the other hand, males fixate more on the right eye and mouth being independent of any age group. Previous studies of this kind showed very strong left eye bias during the first 250 ms of exploration in women [Leonards and Scott-Samuel, 2005]. In the research work of [Sammaknejad et al., 2017] they reported the probabilities for fixations in left eye for male were 0.1732 and for female were 0.201 and it was the highest probability difference between males and females in this study. The work of [Sæther et al., 2009] suggests that females attend more towards the eye region compared to males. All these previous works back our quantitative findings that women had stronger left eye bias compared to males.

Table 5.33: Mean on path length, saccade amplitude, and number of fixations in four ROIs for males and females in different age groups with standard error of the mean using GOF dataset.

Age Group	Path length	Sac. amplitude	Left eye	Right eye	Nose	Mouth
20–72 M	991 ± 1605	308 ± 1567	66 ± 5	96 ± 8	27 ± 3	18 ± 3
20–72 F	2997 ± 14749	2065 ± 14741	90 ± 6	66 ± 7	27 ± 3	13 ± 2
20–30 M	944 ± 591	215 ± 269	69 ± 7	98 ± 11	27 ± 4	16 ± 4
20–30 F	2024 ± 3774	1125 ± 3693	84 ± 6	72 ± 9	26 ± 3	13 ± 3
31–50 M	1208 ± 5098	534 ± 5213	65 ± 11	97 ± 15	27 ± 5	20 ± 5
31–50 F	2521 ± 10225	1382 ± 10192	105 ± 14	43 ± 14	38 ± 8	14 ± 4

Table 5.34: Median on path length and saccade amplitude for males and females in different age groups with standard error of the mean using GOF dataset.

Age Group	Path length	Saccade amplitude
20–72 M	84.08	40.79
20–72 F	101.28	58.16
20–30 M	80.51	41.24
20–30 F	101.07	62.14
31–50 M	82.63	37.50
31–50 F	95.24	47.56

5.4 Conclusion

This chapter explained various gender prediction experiments using three different datasets (Dyslexia, VST, and GOF). The proposed approach for gender prediction performed consistently well for the different datasets. The best achieved accuracy with Dyslexia data was $63.8 \pm 0.5\,\%$, with VST dataset was $69.9 \pm 0.8\,\%$, and with GOF dataset was $77.5 \pm 4.10\,\%$.

Furthermore, the Dyslexia dataset provided an opportunity to study gender prediction in isolated groups of dyslexic and non-dyslexic groups. The results showed that gender prediction based on eye movements is possible for prepubescent children aged 9–10. This is contrary to previous studies which indicated that differences in eye movement of prepubescent children were not significant. The best accuracies of $63.8 \pm 0.5\,\%$ matched previous studies that were performed with adults. It was found

that the gender prediction was better in the isolated groups than in the mixed group. Also, the results of gender prediction in non-dyslexic children were better than in the dyslexic group. A hierarchical classifier was built to improve the accuracy of the mixed group. The hierarchical classifier predicts dyslexia first with (high accuracy) and performed almost equally well as in [Benfatto et al., 2016] with a smaller number of features. Finally, the results show that non-dyslexic females read slower with longer fixations in comparison to non-dyslexic males.

In addition, in the case of the VST dataset, the effect of the trajectory length on the accuracy was studied. The results show that the best accuracy is achieved for short segments 12 seconds from 240 seconds in the visual search task stimuli. Furthermore, the effect of fatigue (existence of blinking) was studied. The better accuracies were achieved when the trajectory was taken from the end of the recording. This hints that predicting the gender can be easier when the participant was losing attention or started to have fatigue especially when engaged in a cognitively demanding task like visual searching of numbers. Given the nature of the visual search task, the gender difference in eye movements may not be evident when the participants were focused on solving the task and were not tired. In comparison to the state of art accuracy, the best case accuracy was almost similar to the state of the art accuracy ($69.9 \pm 0.8\,\%$ vs $70.0 \pm 13.22\,\%$ that was reported in [Sargezeh et al., 2019]). Lastly for the VST dataset, from the quantitative observation, the results show that women compared to men were more explorative in their gaze behavior.

In the GOF dataset, the demographics of this data played an important role to study different factors on gender prediction. To the best of our knowledge, there were no previous studies that have conducted gender prediction using a total of 378 users comprising of 193 males and 185 females. To study gender prediction in different age groups (20–72, 20–30, and 31–50) and using two types of features (fullR and ROI features) were implemented in this data. Reflecting on the proposed objective, accuracies of up to $66.2 \pm 0.90\,\%$ were achieved using the fullR features in the age group of 20–30 years. Using the state-of-the-art features with this dataset was improved by $2.5\,\%$ using the GOF dataset. Different weights for fixation and saccade classifiers were found that affected the accuracies. The Nelder-Mean method was used to optimize the weights in our best accuracy experiments. The results showed that increasing the fixation weights increases the best accuracy of the GOF dataset by $1.1\,\%$ (from $66.2\,\%$ to $67.3\,\%$) and state-of-the-art accuracy by $5\,\%$ (from $72.5\,\%$ to $77.5\,\%$). Fixations were more important when predicting gender as compared to saccades. Finally, the statistics confirmed that the path lengths for females were longer and the saccadic amplitudes were also higher in females. This indicated that females were more explorative than males. Additionally, the previous studies [Leonards and Scott-Samuel, 2005, Sammaknejad et al., 2017], also showed

that the women have a stronger left eye bias compared to males when viewing faces as stimuli being independent of any age group.

Chapter 6

Conclusion and Outlook

The synopsis of this thesis and the main scientific contributions of the work are presented in this chapter. As an outlook, some possible future work is described at the end.

6.1 Thesis Summary

With the eyes functioning as an interface between the outside world and the human brain there is no surprise that the human visual system is physically and neurologically complex. Measuring where a person looks (gaze point) or eye movement is known as eye tracking. Researchers have formulated various algorithms and approaches for tracking location and gaze autonomously, used in various applications. Eye tracking research is gaining attractions due to its capacity to facilitate various tasks, particularly for authentication purposes such as user identification or gender prediction also, it has been used in other fields, such as disease diagnosis (e.g. dyslexia detection), marketing, and gaming. The first dedicated conference to eye tracking research and applications (ETRA) was established in 2000 and has been taking place every two years until 2018 and every year since 2018 which shows the growing significance of this research area. Figure 6.1 shows the growth of annual number of publications in DBLP database which contain "Eye Tracking" in their title since the year 2001. In this thesis, the general motivation is to improve the state of the art results in biometrics and gender prediction based on eye movement data. This has been done by developing a reliable approach for both of these classification tasks using different datasets of varying stimuli.

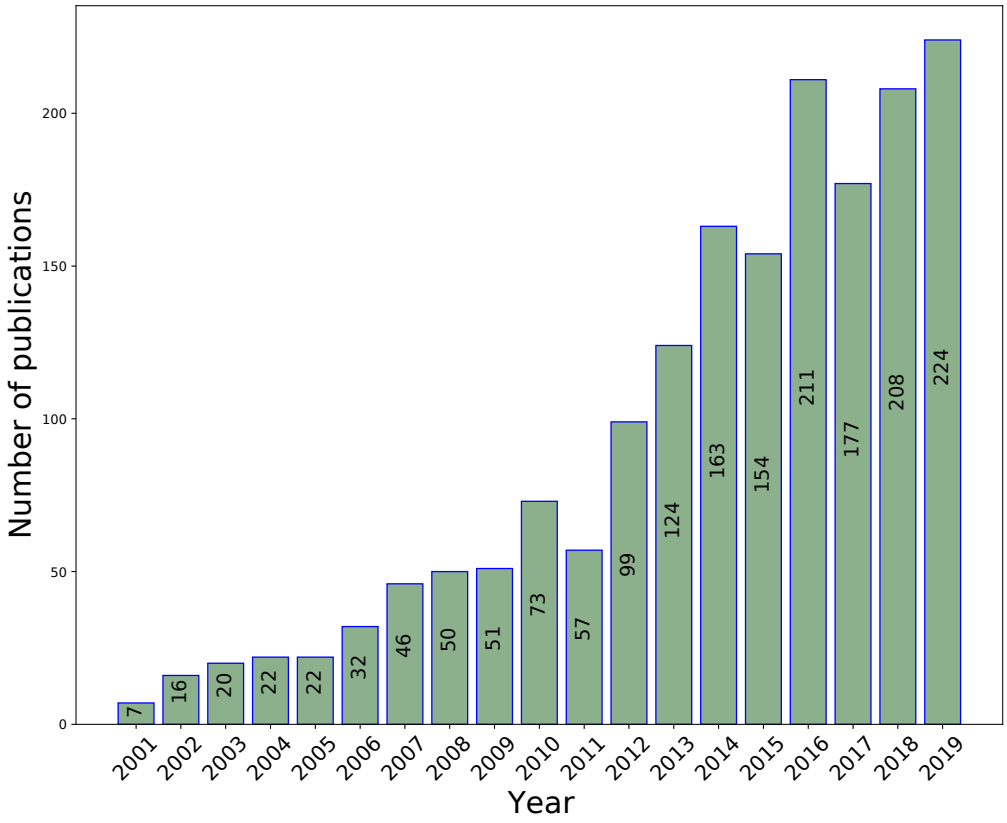


Figure 6.1: Number of eye tracking publications within computer science over 19 years

6.2 Scientific Contributions

This section describes the scientific contributions within the two topics of this thesis.

6.2.1 User Identification

This thesis presents an approach for user identification which works consistently and robustly across different datasets (RAN, TEX, MIT, VST, and GOF) of varying stimuli. The best achieved accuracies were $96.64\,\%$ with RAN, $93.53\,\%$ with TEX, $94.72\,\%$ with VST, and $84.48\,\%$ with GOF datasets over 50 runs. The best results of the winning pipeline in the 2015 BioEye competition [Rigas and Komogortsev, 2017], i.e. [George and Routray, 2016] in state-of-the-art were $89.54\,\%$ with RAN, $85.62\,\%$ with TEX over only one run. Our method improves the state-of-the-art performance and provides more robust predictions. Additionally, an extensive study has been carried out to investigate different factors (for example, IVT parameters, higher order derivative features, effect of gender and age, template aging effect, trajectory length, blinks etc.) that can affect user identification performance. The results

suggest that selecting the best VT parameters and including higher-order derivative features have the greatest positive impact on accuracy. Other findings were that user identification works better in the solo female group than in solo male groups and hence is biased towards gender. Similarly, it works better in an older age group than in a younger age group of participants. Furthermore, to our knowledge, the classification performance degree depends on the degree of task-independence. Training with using eye tracking data obtained from participants when watching several images, and evaluation with using similar, but *different* images is called *weakly task-independent* classification. To study this, our approach has been carried out on a popular dataset (MIT) that has never been used for gaze biometrics before. The outcome of using this dataset showed that our approach achieved $86.7\,\%$ accuracy with only 30+30 samples (training + testing); each sample has a duration of three seconds. With 300 training samples, our methods achieve even $94.7\,\%$.

A summary of the results of the investigated factors and their impact on user identification are as follows:

1. The effect of stimuli was investigated by using four different stimuli. Using the default IVT parameters and all trajectory lengths available for each data showed that the RAN data was giving the highest accuracy ($92.62 \pm 0.13\,\%$ with RAN, $90.90 \pm 0.10\,\%$ with TEX, $85.69 \pm 0.16\,\%$ with VST, and $77.91 \pm 0.51\,\%$ with GOF) over 50 runs. While homogenizing the datasets by taking the same number of users and trajectory lengths in all the datasets showed that both RAN and TEX were giving the best accuracies in comparison with VST and GOF stimuli ($90.37 \pm 0.48\,\%$ with RAN, $90.90 \pm 0.10\,\%$ with TEX, $71.62 \pm 0.35\,\%$ with VST, and $84.48 \pm 0.79\,\%$ with GOF).

2. We study the effect of the IVT parameters towards accuracy of user identification. The results show that tuning the velocity threshold parameter and fixing the minimum fixation duration was giving the highest impact on the user identification accuracy. The accuracy was increased by $3\,\%$ for RAN, by $2\,\%$for TEX, by $9\,\%$ for VST, and by $5\,\%$ for GOF datasets.

3. We investigate the effect of adding higher-order derivatives (such as velocity, acceleration, jerk, etc.). We find that the accuracy is improved by $2\,\%$ for RAN, by $1\,\%$ for TEX and VST, and by $3\,\%$ for GOF datasets, by including until jounce level features in all the datasets except for the VST dataset, where the accuracy rises only until the jerk.

4. Some datasets contain data which can be used to deduce blink information. From the blink information we compute certain features. The inclusion of those

features increased the accuracy by $1\,\%$ for RAN and VST, by $0.5\,\%$ for TEX datasets.

5. Combining the following factors had the greatest impact on accuracy improvement: IVT parameters, higher-order derivative features, and blink classifier increased the accuracy 4 % for RAN, 3 % for TEX, and 9 % for VST datasets respectively. The final accuracy of $96.64 \pm 0.07\,\%$ for the RAN dataset, $93.53 \pm 0.11\,\%$ for the TEX dataset, and $94.79 \pm 0.03\,\%$ for the VST dataset respectively were obtained over 50 runs.

6. Further experiments were conducted and the outcome suggest that user identification works better in the solo female group (accuracy of $88.85 \pm 0.32\,\%$) than in the solo male group (accuracy of $77.37 \pm 0.50\,\%$) consisting of 150 females and 150 males with the GOF dataset over 50 runs and hence the user identification is biased towards gender. Similarly, it worked better in an older age group (accuracy of $91.43 \pm 0.47\,\%$ with age group 41–72) than in a younger age group (accuracy of $85.96 \pm 0.79\,\%$ with age group 20–40) consisting of 56 and 56 participants, respectively in the GOF dataset over 50 runs.

7. Finally, the accuracy of user identification decreased for all RAN $11.41\,\%$, TEX by $18.17\,\%$, and VST by $18.6\,\%$ datasets when there was a significant time gap between train and test sessions (as shown in Table 4.11). This may be attributed to changing physiological parameters of the participants, device characteristics, and some other inexplicable effects. However, in comparison with the state of the art results, our approach has improved the accuracy by $5.40\,\%$ with using RAN dataset and $2.70\,\%$ with using TEX dataset over one run.

6.2.2 Gender Prediction

For this task three different datasets (Dyslexia, VST, and GOF datasets) were used. The best achieved accuracy with Dyslexia data was $63.8 \pm 0.50\,\%$ over 1000 runs, with VST dataset was $69.9 \pm 0.8\,\%$ over 300 runs, and with GOF was $67.3 \pm 0.98 \pm 4.10\,\%$ over 50 runs (and $77.5\,\%$ over 5 runs when compared with the state of the art).

The Dyslexia dataset provided an opportunity to study gender prediction in isolated groups of dyslexic and non-dyslexic children. A novel method for gender prediction utilizing prepubescent children's eye movements aged 9–10 [Zaidawi et al., 2020] was proposed and the results show that gender prediction is possible. This is contrary to previous studies which indicated that differences in eye movement of prepubescent children were not significant [Miyahira et al., 2000b]. The best accuracies of 64 % matched previous studies [Moss et al., 2012] that were performed with adults. It was

found that gender prediction was better in the isolated groups than in the mixed group. Also, the results of gender prediction in the non-dyslexic group were better than in the dyslexic group.

In mixed groups, our classifiers' accuracy plummets drastically to $56.4\,\%$ with mixed group features. A hierarchical classifier was built, that uses dyslexia prediction to greatly enhance gender prediction accuracy in mixed populations to $62.1\,\%$. The hierarchical classifier predicts dyslexia first (which achieves high accuracy and performed almost equally well as in [Benfatto et al., 2016], but with lesser number features). Finally for dyslexia dataset, the results of quantitative observations showed that the non-dyslexic females read slower with longer fixations in comparison to the non-dyslexic males.

In the case of the VST dataset the effect of the trajectory length on the accuracy was studied. The results show that only 12 seconds from 240 seconds in VST were enough to predict gender. Furthermore, the effect of fatigue (existence of blinking) was studied. Better accuracies were achieved when the trajectory was taken from the end of the recording. This hints that predicting the gender can be easier when the participant was losing attention or started to have fatigue especially when engaged in a cognitively demanding task like visual searching of numbers task as they switch from their natural behavior. Given the nature of the visual search task, gender differences in eye movements may not be evident when the participants were focused on solving the task and were not tired. In comparison to the state of art accuracy, the best case accuracy was almost similar to the state of the art accuracy ($69.9\,\%$ vs $70.0\,\%$ reported in [Sargezeh et al., 2019]). Lastly for the VST dataset, from the quantitative observation, the results showed that women compared to men were more exploration in their gaze behavior.

In the GOF dataset the demographics of this data played an important role to study different factors affecting the gender prediction accuracy. To the best of our knowledge, there were no previous studies that have conducted gender prediction using a dataset with as many participants as in GOF (378 users comprised of 193 males and 185 females) except for a recent work in [Mohammad, 2021]. We study gender prediction in different age groups (20–72, 20–30, and 31–50) and using two types of features (fullR and ROI features). Reflecting on the proposed objective, accuracies of up to 66.2 % were achieved using fullR features in the age group of 20–30 years. Using the state-of-the-art features with this dataset was improved by 2.5 % using the GOF dataset. Different weights for fixation and saccade classifiers were found that affected the accuracies. The Nelder-Mead method was used to optimize the weights in our best accuracy experiments. The results

showed that increasing the fixation weights increased the best accuracy of the GOF dataset by 1.1 % (from 66.2 % to 67.3 %) and state-of-the-art accuracy by 5 % (from 72.5 % to 77.5 %). Fixations were more important when predicting gender as compared to saccades. The statistics confirmed that the path lengths for females were longer and the saccadic amplitudes were also higher in females. The total number of fixations was larger for males than for females. This indicated that females were more explorative than males. Additionally, the previous studies [Leonards and Scott-Samuel, 2005, Sammaknejad et al., 2017], also showed that women have a stronger left eye bias compared to males when viewing faces as stimuli being independent of any age group.

6.2.3 Limitations

We believe more work is needed on improving the accuracy of both user identification and gender prediction when there is a significant time gap between train and test sessions. This is crucial for feasibility of eye movement biometrics and gender classification in real-world applications.

Also, none of the tested methods is capable of strong task-independent user identification. Our initial user identification work [Schröder et al., 2020] results suggested that transfer learning is highly non-symmetric. Training on text and evaluating on random dots performs three times better than the other way around. Therefore, we believe in the future more work is needed on methods that perform well with strongly task-independent settings. Especially the difference in the performance of the different classifiers we observed seems to be a promising start. Task-independence could also be combined with multi task learning similar to the approach from Kaiser et al. [Kaiser et al., 2017].

It will be interesting also to use the eye movement data to predict the age group of people. Here it would also be interesting to study what is the appropriate size of the age groups e.g. 10 to 20, 20 to 30, 30 to 40 or 10 to 25, 25 to 40, and so on, and in which cases we get statistically significant results. Then this classifier of age prediction can be combined with the user and gender prediction classifiers to build an overall hierarchical classier that can predict the user ID, gender, and age of the participants and also use the cross dependencies between the classifiers to infer better accuracies.

Increasing the number of participants is expected to have a negative impact on the system's performance. Increasing the number of participants increases

the complexity of the feature space, and accordingly increases the difficulty of the identification task which can lead to a decrease in accuracy. So user identification must be carried out with a big number of participants in the future (e.g. the recent released dataset in [Lohr et al., 2021]).

In the scope of this thesis, we used supervised ML models to predict the gender in different datasets. However, some state of the art studies (e.g. [Sammaknejad et al., 2017]) have shown that statistical models of transitions between different ROIs while viewing a face stimulus can outperform ML models for gender prediction task. However these studies are done on datasets with few number of participants hence it will be interesting to test such an approach on a large dataset like GOF dataset were we have 378 participants available and check whether statistical models outperform the ML models. Also, there has been a recent work [Mohammad, 2021] which uses an unsupervised deep learning approach for gender prediction but the reported results showed a high mean absolute error. This could also be an avenue for future investigations.

Also, lighting conditions are different from day-time to night-time which affects the eye-tracking system, thus, affect the extracted features. In the future, this needs to be addressed by recording multiple sessions from the same participant under different human and lighting conditions.

Finally, the ethical implications of this kind of research must be carefully examined since this approach enables us to predict user attributes such as ID, gender etc. merely from the eye movements of the person. Some perpetrators might use such technology without the human consent and use their eye movements to deduce certain information about them or, persons may agree that their eye movements are recorded, but maybe unaware that this may reveal their gender or other attributes. This raises new questions and challenges in securing user privacy. Hence, the society must be made aware of the implications of such a technology and ethical regulations should be prepared and put in place in the future.

Bibliography

[Alghowinem et al., 2013] Alghowinem, S., Goecke, R., Wagner, M., Parker, G., and Breakspear, M. (2013). Eye movement analysis for depression detection. In *2013 IEEE International Conference on Image Processing*, pages 4220–4224. IEEE.

[Alkan and Cagiltay, 2007] Alkan, S. and Cagiltay, K. (2007). Studying computer game learning experience through eye tracking. *British Journal of Educational Technology*, 38(3):538–542.

[Altman and Bland, 2005] Altman, D. G. and Bland, J. M. (2005). Standard deviations and standard errors. *Bmj*, 331(7521):903.

[Aly, 2005] Aly, M. (2005). Survey on multiclass classification methods. *Neural Netw*, 19:1–9.

[Andersson et al., 2017] Andersson, R., Larsson, L., Holmqvist, K., Stridh, M., and Nyström, M. (2017). One algorithm to rule them all? an evaluation and discussion of ten eye movement event-detection algorithms. *Behavior Research Methods*, 49(2):616–637.

[Annetta et al., 2007] Annetta, L. A., Slykhuis, D., and Wiebe, E. (2007). Evaluating gender differences of attitudes and perceptions toward powerpoint for preservice science teachers. *Eurasia Journal of Mathematics, Science & Technology Education*, 3(4).

[Appadurai and Bhargavi, 2021] Appadurai, J. P. and Bhargavi, R. (2021). Eye movement feature set and predictive model for dyslexia: Feature set and predictive model for dyslexia. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 15(4):1–22.

[Armstrong and Olatunji, 2012] Armstrong, T. and Olatunji, B. O. (2012). Eye tracking of attention in the affective disorders: A meta-analytic review and synthesis. *Clinical psychology review*, 32(8):704–723.

[Babich, 2012] Babich, A. (2012). Biometric authentication. types of biometric identifiers.

[Babin et al., 2006] Babin, S. L., Hood, A. J., Jeter, C. B., and Sereno, A. B. (2006). Executive functions: eye movements and neuropsychiatric disorders. *Encyclopedia of Neuroscience (2009)*, 4:117–122.

[Banks et al., 2012] Banks, M. S., Read, J. C., Allison, R. S., and Watt, S. J. (2012). Stereoscopy and the human visual system. *SMPTE motion imaging journal*, 121(4):24–43.

[Baratloo et al., 2015] Baratloo, A., Hosseini, M., Negida, A., and El Ashal, G. (2015). Part 1: simple definition and calculation of accuracy, sensitivity and specificity.

[Baron-Cohen, 2002] Baron-Cohen, S. (2002). The extreme male brain theory of autism. *Trends in cognitive sciences*, 6(6):248–254.

[Bayes, 1968] Bayes, T. (1968). Naive bayes classifier. *Article Sources and Contributors*, pages 1–9.

[Benedetto et al., 2011] Benedetto, S., Pedrotti, M., Minin, L., Baccino, T., Re, A., and Montanari, R. (2011). Driver workload and eye blink duration. *Transportation Research Part F: Traffic Psychology and Behaviour*, 14(3):199–208.

[Benfatto et al., 2016] Benfatto, M. N., Seimyr, G. Ö., Ygge, J., Pansell, T., Rydberg, A., and Jacobson, C. (2016). Screening for dyslexia using eye tracking during reading. *PloS one*, 11(12):e0165508.

[Billeci et al., 2017] Billeci, L., Narzisi, A., Tonacci, A., Sbriscia-Fioretti, B., Serasini, L., Fulceri, F., Apicella, F., Sicca, F., Calderoni, S., and Muratori, F. (2017). An integrated eeg and eye-tracking approach for the study of responding and initiating joint attention in autism spectrum disorders. *Scientific Reports*, 7(1):1–13.

[Bishop, 2006] Bishop, C. M. (2006). Multiclass logistic regression. In *Pattern recognition and machine learning*, chapter 4.3.4, pages 209–210. Springer.

[Breiman, 1996] Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[Broomhead and Lowe, 1988] Broomhead, D. S. and Lowe, D. (1988). Radial basis functions, multi-variable functional interpolation and adaptive networks. Technical report, Royal Signals and Radar Establishment Malvern (United Kingdom).

[Büyük, 2021] Büyük, M. (2021). Classifying sex based on eye tracking data: A machine learning study. B.S. thesis.

[Cárdenas et al., 2013] Cárdenas, R. A., Harris, L. J., and Becker, M. W. (2013). Sex differences in visual attention toward infant faces. *Evolution and Human Behavior*, 34(4):280–287.

[Caselli et al., 2009] Caselli, M., Trizio, L., De Gennaro, G., and Ielpo, P. (2009). A simple feedforward neural network for the pm 10 forecasting: Comparison with a radial basis function network and a multivariate linear regression model. *Water, Air, and Soil Pollution*, 201(1):365–377.

[Cesa-Bianchi et al., 2006] Cesa-Bianchi, N., Gentile, C., and Zaniboni, L. (2006). Hierarchical classification: combining bayes with svm. In *Proceedings of the 23rd international conference on Machine learning*, pages 177–184.

[Chakraborty and Sundaram, 2020] Chakraborty, V. and Sundaram, M. (2020). Machine learning algorithms for prediction of dyslexia using eye movement. In *Journal of Physics: Conference Series*, volume 1427, page 012012. IOP Publishing.

[Clay et al., 2019] Clay, V., Koenig, P., and Koenig, S. (2019). Eye tracking in virtual reality. *Journal of Eye Movement Research*, 12(1).

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

[Coutrot et al., 2016] Coutrot, A., Binetti, N., Harrison, C., Mareschal, I., and Johnston, A. (2016). Face exploration dynamics differentiate men and women. *Journal of vision*, 16(14):16–16.

[Crundall and Underwood, 2011] Crundall, D. and Underwood, G. (2011). Visual attention while driving: measures of eye movements used in driving research. In *Handbook of traffic psychology*, pages 137–148. Elsevier.

[Cutler et al., 2012] Cutler, A., Cutler, D. R., and Stevens, J. R. (2012). Random forests. In *Ensemble machine learning*, pages 157–175. Springer.

[Dalrymple et al., 2019] Dalrymple, K. A., Jiang, M., Zhao, Q., and Elison, J. T. (2019). Machine learning accurately classifies age of toddlers based on eye tracking. *Scientific reports*, 9(1):1–10.

[Daniel, 1990] Daniel, W. W. (1990). Kruskal–wallis one-way analysis of variance by ranks. *Applied nonparametric statistics*, pages 226–234.

[Darwish, 2013] Darwish, A. A. (2013). Biometric identification based on eye movements and iris features using task-driven and task-independent stimuli. Master's thesis, American University of Sharjah.

[de Souza Jacomini et al., 2012] de Souza Jacomini, R., do Nascimento, M. Z., Dantas, R. D., and Ramos, R. P. (2012). Comparison of pca and anova for information selection of cc and mlo views in classification of mammograms. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 117–126. Springer.

[Derks et al., 1995] Derks, E. P. P. A., Pastor, M. S. S., and Buydens, L. M. C. (1995). Robustness analysis of radial base function and multi-layered feed-forward neural network models. *Chemometrics and Intelligent Laboratory Systems*, 28(1):49–60.

[Di Stasi et al., 2012] Di Stasi, L. L., Renner, R., Catena, A., Cañas, J. J., Velichkovsky, B. M., and Pannasch, S. (2012). Towards a driver fatigue test based on the saccadic main sequence: A partial validation by subjective report data. *Transportation research part C: emerging technologies*, 21(1):122–133.

[Domingos and Pazzani, 1997] Domingos, P. and Pazzani, M. (1997). On the optimality of the simple bayesian classifier under zero-one loss. *Machine learning*, 29(2):103–130.

[Dorey, 2010] Dorey, F. (2010). In brief: The p value: What is it and what does it tell you?

[Doughty, 2002] Doughty, M. J. (2002). Further assessment of gender-and blink pattern-related differences in the spontaneous eyeblink activity in primary gaze in young adult humans. *Optometry and Vision Science*, 79(7):439–447.

[Duchowski et al., 2016] Duchowski, A. T., Jörg, S., Allen, T. N., Giannopoulos, I., and Krejtz, K. (2016). Eye movement synthesis. In *Proceedings of the ninth biennial ACM symposium on eye tracking research & applications*, pages 147–154.

[Dudoit and Fridlyand, 2002] Dudoit, S. and Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome biology*, 3(7):1–21.

[Edgar et al., 2017] Edgar, T., Manz, D., and Manz, D. (2017). Exploratory study. *Research methods for cyber security*, 29:95–130.

[El Hmimdi et al., 2021] El Hmimdi, A. E., Ward, L. M., Palpanas, T., and Kapoula, Z. (2021). Predicting dyslexia and reading speed in adolescents from eye movements in reading and non-reading tasks: A machine learning approach. *Brain Sciences*, 11(10):1337.

[Emam and Youssef, 2012] Emam, A. and Youssef, A. E. (2012). Do females read faster than males? an empirical study using eye tracking systems. *International Journal of Computer Science Issues (IJCSI)*, 9(3):232.

[Eraslan et al., 2020] Eraslan, S., Yesilada, Y., Yaneva, V., and Harper, S. (2020). 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*, pages 1–10.

[Esfahani, 2016] Esfahani, N. M. (2016). A brief review of human identification using eye movement. *Journal of Pattern Recognition Research*, 11(1):15–24.

[Evans et al., 2012] Evans, K. M., Jacobs, R. A., Tarduno, J. A., and Pelz, J. B. (2012). Collecting and analyzing eye tracking data in outdoor environments. *Journal of Eye Movement Research*, 5(2):6.

[EyeLink, 2022] EyeLink, S. R. (2022). Eyelink 1000 plus: A highly accurate, precise, and versatile eye tracker. https://www.sr-research.com/eyelink-1000-plus/. [Online; accessed 28-January-2022].

[Faan et al., 2010] Faan, R. W. B. M., Dmsc, V. H. M., Kerber, K., and Kerber, K. A. (2010). Baloh and honrubia's clinical neurophysiology of the vestibular system.

[Fawcett, 2006] Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

[Fernández et al., 2013] Fernández, G., Mandolesi, P., Rotstein, N. P., Colombo, O., Agamennoni, O., and Politi, L. E. (2013). Eye movement alterations during reading in patients with early alzheimer disease. *Investigative ophthalmology & visual science*, 54(13):8345–8352.

[Fisher, 1936] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(7):179–188.

[Freitas and Carvalho, 2007] Freitas, A. and Carvalho, A. (2007). A tutorial on hierarchical classification with applications in bioinformatics. In *Research and trends in data mining technologies and applications*, pages 175–208. IGI Global.

[Freund et al., 1999] Freund, Y., Schapire, R., and Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612.

[Friedman et al., 2017] Friedman, L., Nixon, M. S., and Komogortsev, O. V. (2017). Method to assess the temporal persistence of potential biometric features: Application to oculomotor, gait, face and brain structure databases. *PloS one*, 12(6):e0178501.

[Friedman et al., 1997] Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning*, 29(2):131–163.

[Galdi et al., 2016] Galdi, C., Nappi, M., Riccio, D., and Wechsler, H. (2016). Eye movement analysis for human authentication: a critical survey. *Pattern Recognition Letters*, 84:272–283.

[Gao and Han, 2012] Gao, F. and Han, L. (2012). Implementing the nelder-mead simplex algorithm with adaptive parameters. *Computational Optimization and Applications*, 51(1):259–277.

[Garland and Martin, 2005] Garland, D. and Martin, B. N. (2005). Do gender and learning style play a role in how online courses should be designed. *Journal of Interactive Online Learning*, 4(2):67–81.

[Garza et al., 2016] Garza, R., Heredia, R. R., and Cieslicka, A. B. (2016). Male and female perception of physical attractiveness: An eye movement study. *Evolutionary Psychology*, 14(1):1474704916631614.

[George and Routray, 2016] George, A. and Routray, A. (2016). A score level fusion method for eye movement biometrics. *Pattern Recognition Letters*, 82:207–215.

[Gibaldi and Sabatini, 2021] Gibaldi, A. and Sabatini, S. P. (2021). The saccade main sequence revised: A fast and repeatable tool for oculomotor analysis. *Behavior Research Methods*, 53(1):167–187.

[Gibbons and Pratt, 1975] Gibbons, J. D. and Pratt, J. W. (1975). P-values: interpretation and methodology. *The American Statistician*, 29(1):20–25.

[Girden, 1992] Girden, E. R. (1992). *ANOVA: Repeated measures*. Number 84. Sage.

[Griffith et al., 2020] Griffith, H. K., Lohr, D. J., Abdulin, E., and Komogortsev, O. (2020). Gazebase: A large-scale, multi-stimulus, longitudinal eye movement dataset. *CoRR*, abs/2009.06171.

[Guyon et al., 2002] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1):389–422.

[Haldar, 2013] Haldar, S. (2013). Statistical and geostatistical applications in geology. *In: mineral exploration: principles and applications. Waltham: Elsevier*, pages 157–82.

[Hall et al., 2010] Hall, J. K., Hutton, S. B., and Morgan, M. J. (2010). Sex differences in scanning faces: Does attention to the eyes explain female superiority in facial expression recognition? *Cognition & Emotion*, 24(4):629–637.

[Hao et al., 2007] Hao, P.-Y., Chiang, J.-H., and Tu, Y.-K. (2007). Hierarchically svm classification based on support vector clustering method and its application to document categorization. *Expert Systems with applications*, 33(3):627–635.

[Harezlak et al., 2021] Harezlak, K., Blasiak, M., and Kasprowski, P. (2021). Biometric identification based on eye movement dynamic features. *Sensors*, 21(18):6020.

[Haria et al., 2022] Haria, R. V. V., Zaidawi, S. A., and Maneth, S. (2022). Eye movement analysis to predict gender using different sets of features. Computer Vision, Imaging and Computer Graphics Theory and Applications VISAPP 2022 conference. submitted in November 2021.

[Heisz et al., 2013] Heisz, J. J., Pottruff, M. M., and Shore, D. I. (2013). Females scan more than males: A potential mechanism for sex differences in recognition memory. *Psychological science*, 24(7):1157–1163.

[Hessels et al., 2017] Hessels, R. S., Niehorster, D. C., Kemner, C., and Hooge, I. T. (2017). Noise-robust fixation detection in eye movement data: Identification by two-means clustering (i2mc). *Behavior research methods*, 49(5):1802–1823.

[Hilden, 1984] Hilden, J. (1984). Statistical diagnosis based on conditional independence does not require it. *Computers in biology and medicine*, 14(4):429–435.

[Holland and Komogortsev, 2011] Holland, C. and Komogortsev, O. V. (2011). Biometric identification via eye movement scanpaths in reading. In *2011 IEEE International Joint Conference on Biometrics, IJCB 2011, Washington, DC, USA, October 11-13, 2011*, pages 1–8.

[Holland and Komogortsev, 2012] Holland, C. D. and Komogortsev, O. V. (2012). Biometric verification via complex eye movements: The effects of environment and stimulus. In *IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems, BTAS 2012, Arlington, VA, USA, September 23-27, 2012*, pages 39–46.

[Horsley et al., 2013] Horsley, M., Eliot, M., Knight, B. A., and Reilly, R. (2013). *Current trends in eye tracking research*. Springer.

[Huang and Chen, 2016] Huang, P.-S. and Chen, H.-C. (2016). Gender differences in eye movements in solving text-and-diagram science problems. *International Journal of Science and Mathematics Education*, 14(2):327–346.

[Hwang and Lee, 2018] Hwang, Y. M. and Lee, K. C. (2018). Using an eye-tracking approach to explore gender differences in visual attention and shopping attitudes in an online shopping environment. *International Journal of Human–Computer Interaction*, 34(1):15–24.

[Iqbal, 2012] Iqbal, T. (2012). *A robust real time eye tracking and gaze estimation system using particle filters*. The University of Texas at El Paso.

[Isokoski et al., 2009] Isokoski, P., Joos, M., Spakov, O., and Martin, B. (2009). Gaze controlled games. *Universal Access in the Information Society*, 8(4):323.

[Jacob and Karn, 2003] Jacob, R. J. and Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *The mind's eye*, pages 573–605. Elsevier.

[Jäger et al., 2019] Jäger, L. A., Makowski, S., Prasse, P., Liehr, S., Seidler, M., and Scheffer, T. (2019). Deep eyedentification: Biometric identification using micro-movements of the eye. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 299–314. Springer.

[Jenkin, 2012] Jenkin, R. (2012). The human visual system. In *The Manual of Photography and Digital Imaging*, pages 59–76. Routledge.

[Jothi Prabha and Bhargavi, 2019] Jothi Prabha, A. and Bhargavi, R. (2019). Prediction of dyslexia from eye movements using machine learning. *IETE Journal of Research*, pages 1–10.

[Juan, 2006] Juan, S. (30 Jun 2006). Why do babies blink less often than adults? `https://www.theregister.com/2006/06/30/the_odd_body_blinking/`. [Online; accessed 11-August-2021].

[Judd et al., 2009] Judd, T., Ehinger, K. A., Durand, F., and Torralba, A. (2009). Learning to predict where humans look. In *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*, pages 2106–2113.

[Kaiser et al., 2017] Kaiser, L., Gomez, A. N., Shazeer, N., Vaswani, A., Parmar, N., Jones, L., and Uszkoreit, J. (2017). One model to learn them all. *arXiv:1706.05137*.

[Kasneci et al., 2021] Kasneci, E., Kasneci, G., Trautwein, U., Appel, T., Tibus, M., Jaeggi, S. M., and Gerjets, P. (2021). Do your eye movements reveal your performance on an iq test? a study linking eye movements and socio-demographic information to fluid intelligence. *PsyArXiv*.

[Kaspar and König, 2012] Kaspar, K. and König, P. (2012). Emotions and personality traits as high-level factors in visual attention: a review. *Frontiers in Human Neuroscience*, 6:321.

[Kasprowski et al., 2012] Kasprowski, P., Komogortsev, O. V., and Karpov, A. (2012). First eye movement verification and identification competition at BTAS 2012. In *IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems, BTAS 2012, Arlington, VA, USA, September 23-27, 2012*, pages 195–202.

[Kasprowski and Ober, 2004] Kasprowski, P. and Ober, J. (2004). Eye movements in biometrics. In *Biometric Authentication, ECCV 2004 International Workshop, BioAW 2004, Prague, Czech Republic, May 15, 2004, Proceedings*, pages 248–258.

[Katsini et al., 2020] Katsini, C., Abdrabou, Y., Raptis, G. E., Khamis, M., and Alt, F. (2020). The role of eye gaze in security and privacy applications: Survey and future hci research directions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

[Kinsman et al., 2010] Kinsman, T., Bajorski, P., and Pelz, J. B. (2010). Hierarchical image clustering for analyzing eye tracking videos. In *2010 Western New York Image Processing Workshop*, pages 58–61. IEEE.

[Komogortsev et al., 2010] Komogortsev, O. V., Jayarathna, S., Aragon, C. R., and Mechehoul, M. (2010). Biometric identification via an oculomotor plant mathematical model. In Morimoto, C. H., Istance, H. O., Hyrskykari, A., and Ji, Q., editors, *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications, ETRA 2010, Austin, Texas, USA, March 22-24, 2010*, pages 57–60. ACM.

[Kotsiantis et al., 2006] Kotsiantis, S. B., Zaharakis, I. D., and Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26(3):159–190.

[Krishna et al., 2019] Krishna, V., Ding, Y., Xu, A., and Höllerer, T. (2019). Multimodal biometric authentication for vr/ar using eeg and eye tracking. In *Adjunct of the 2019 International Conference on Multimodal Interaction*, pages 1–5.

[Kröger et al., 2019] Kröger, J. L., Lutz, O. H., and Müller, F. (2019). What does your gaze reveal about you? on the privacy implications of eye tracking. In Friedewald,

M., Önen, M., Lievens, E., Krenn, S., and Fricker, S., editors, *Privacy and Identity Management. Data for Better Living: AI and Privacy - 14th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2.2 International Summer School, Windisch, Switzerland, August 19-23, 2019, Revised Selected Papers*, volume 576 of *IFIP Advances in Information and Communication Technology*, pages 226–241. Springer.

[Kröger et al., 2020] Kröger, J. L., Lutz, O. H.-M., and Müller, F. (2020). What does your gaze reveal about you? on the privacy implications of eye tracking. *IFIP Advances in Information and Communication Technology*, 576:226–241.

[Kruskal and Wallis, 1952] Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621.

[Lagun et al., 2011] Lagun, D., Manzanares, C., Zola, S. M., Buffalo, E. A., and Agichtein, E. (2011). Detecting cognitive impairment by eye movement analysis using automatic classification algorithms. *Journal of Neuroscience Methods*, 201(1):196–203.

[Langley et al., 1992] Langley, P., Iba, W., Thompson, K., et al. (1992). An analysis of bayesian classifiers. In *Aaai*, volume 90, pages 223–228. Citeseer.

[Lankes and Stoeckl, 2020] Lankes, M. and Stoeckl, A. (2020). Gazing at pac-man: Lessons learned from a eye-tracking study focusing on game difficulty. In *ACM Symposium on Eye Tracking Research and Applications*, ETRA '20 Short Papers, New York, NY, USA. Association for Computing Machinery.

[Lee, 1991] Lee, Y. (1991). Handwritten digit recognition using k nearest-neighbor, radial-basis function, and backpropagation neural networks. *Neural computation*, 3(3):440–449.

[Leigh and Zee, 2015] Leigh, R. J. and Zee, D. S. (2015). *The neurology of eye movements*. OUP USA.

[Leonards and Scott-Samuel, 2005] Leonards, U. and Scott-Samuel, N. E. (2005). Idiosyncratic initiation of saccadic face exploration in humans. *Vision research*, 45(20):2677–2684.

[Li et al., 2018a] Li, C., Xue, J., Quan, C., Yue, J., and Zhang, C. (2018a). Biometric recognition via texture features of eye movement trajectories in a visual searching task. *PloS one*, 13(4):e0194475.

[Li et al., 2018b] Li, C., Xue, J., Quan, C., Yue, J., and Zhang, C. (2018b). Biometric recognition via texture features of eye movement trajectories in a visual searching task. *PloS one*, 13(4):e0194475.

[Lin et al., 2004] Lin, C.-S., Huan, C.-C., Chan, C.-N., Yeh, M.-S., and Chiu, C.-C. (2004). Design of a computer game using an eye-tracking device for eye's activity rehabilitation. *Optics and lasers in engineering*, 42(1):91–108.

[Liu et al., 2015] Liu, W., Yu, X., Raj, B., Yi, L., Zou, X., and Li, M. (2015). Efficient autism spectrum disorder prediction with eye movement: A machine learning framework. In *2015 International conference on affective computing and intelligent interaction (ACII)*, pages 649–655. IEEE.

[Lohr et al., 2021] Lohr, D., Griffith, H., and Komogortsev, O. V. (2021). Eye know you: Metric learning for end-to-end biometric authentication using eye movements from a longitudinal dataset. *arXiv preprint arXiv:2104.10489*.

[Lohr et al., 2020] Lohr, D. J., Aziz, S., and Komogortsev, O. (2020). Eye movement biometrics using a new dataset collected in virtual reality. In *ACM Symposium on Eye Tracking Research and Applications*, ETRA '20 Adjunct, New York, NY, USA. Association for Computing Machinery.

[Lorena et al., 2008] Lorena, A. C., De Carvalho, A. C. P. L. F., and Gama, J. M. P. (2008). A review on the combination of binary classifiers in multiclass problems. *Artificial Intelligence Review*, 30(1-4):19.

[Lorigo et al., 2006] Lorigo, L., Pan, B., Hembrooke, H., Joachims, T., Granka, L., and Gay, G. (2006). The influence of task and gender on search and evaluation behavior using google. *Information processing & management*, 42(4):1123–1131.

[Majaranta and Bulling, 2014] Majaranta, P. and Bulling, A. (2014). Eye tracking and eye-based human–computer interaction. In *Advances in physiological computing*, pages 39–65. Springer.

[Makowski et al., 2021] Makowski, S., Prasse, P., Reich, D. R., Krakowczyk, D., Jäger, L. A., and Scheffer, T. (2021). Deepeyedentificationlive: Oculomotoric biometric identification and presentation-attack detection using deep neural networks. *IEEE Trans. Biom. Behav. Identity Sci.*, 3(4):506–518.

[Mercer Moss et al., 2012] Mercer Moss, F. J., Baddeley, R., and Canagarajah, N. (2012). Eye movements to natural images as a function of sex and personality. *PLoS One*, 7(11):e47870.

[Miyahira et al., 2000a] Miyahira, A., Morita, K., Yamaguchi, H., Morita, Y., and Maeda, H. (2000a). Gender differences and reproducibility in exploratory eye movements of normal subjects. *Psychiatry and clinical neurosciences*, 54(1):31–36.

[Miyahira et al., 2000b] Miyahira, A., Morita, K., Yamaguchi, H., Nonaka, K., and Maeda, H. (2000b). Gender differences of exploratory eye movements: A life span study. *Life Sciences*, 68(5):569 – 577.

[Miyahira et al., 2000c] Miyahira, A., Morita, K., Yamaguchi, H., Nonaka, K., and Maeda, H. (2000c). Gender differences of exploratory eye movements: a life span study. *Life sciences*, 68(5):569–577.

[Mohammad, 2021] Mohammad, F. (2021). Artificial intelligence for the predication of demographics from unsupervised eye tracking data. Master's thesis.

[Molitor et al., 2015] Molitor, R. J., Ko, P. C., and Ally, B. A. (2015). Eye movements in alzheimer's disease. *Journal of Alzheimer's disease: JAD*, 44(1):1.

[Moss et al., 2012] Moss, F. J. M., Baddeley, R., and Canagarajah, N. (2012). Eye movements to natural images as a function of sex and personality. *PLoS One*, 7(11):e47870.

[Nam et al., 2020] Nam, U., Lee, K., Ko, H., Lee, J.-Y., and Lee, E. C. (2020). Analyzing facial and eye movements to screen for alzheimer's disease. *Sensors*, 20(18).

[Nelder and Mead, 1965] Nelder, J. A. and Mead, R. (1965). A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313.

[Nieto, 2015] Nieto, M. P. (2015). File: human visual pathway. svg. Accessed: 2021-12-10.

[Noble, 2006] Noble, W. S. (2006). What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567.

[Obaidellah and Haek, 2018] Obaidellah, U. and Haek, M. A. (2018). Evaluating gender difference on algorithmic problems using eye-tracker. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 1–8.

[Olsen, 2012] Olsen, A. (2012). The tobii i-vt fixation filter. *Tobii Technology*, 21.

[Olsen and Matos, 2012] Olsen, A. and Matos, R. (2012). Identifying parameter values for an I-VT fixation filter suitable for handling data sampled with various sampling frequencies. In *Proceedings of the 2012 Symposium on Eye-Tracking*

Research and Applications, ETRA 2012, Santa Barbara, CA, USA, March 28-30, 2012, pages 317–320.

[Pan et al., 2004] Pan, B., Hembrooke, H. A., Gay, G. K., Granka, L. A., Feusner, M. K., and Newman, J. K. (2004). The determinants of web page viewing behavior: an eye-tracking study. In *Proceedings of the 2004 symposium on Eye tracking research & applications*, pages 147–154.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Pérez-Moreno et al., 2016] Pérez-Moreno, E., Romero-Ferreiro, V., and García-Gutiérrez, A. (2016). Where to look when looking at faces: visual scanning is determined by gender, expression and tasks demands. *Psicológica*, 37(2):127–150.

[Philbin et al., 1995] Philbin, M., Meier, E., Huffman, S., and Boverie, P. (1995). A survey of gender and learning styles. *Sex roles*, 32(7-8):485–494.

[Pieters and Wedel, 2017] Pieters, R. and Wedel, M. (2017). A review of eye-tracking research in marketing. In *Review of marketing research*, pages 143–167. Routledge.

[Porta et al., 2021] Porta, M., Dondi, P., Zangrandi, N., and Lombardi, L. (2021). Gaze-based biometrics from free observation of moving elements. *IEEE Transactions on Biometrics, Behavior, and Identity Science*.

[Prinzler et al., 2021] Prinzler, M. H. U., Schröder, C., Zaidawi, S. M. K. A., Zachmann, G., and Maneth, S. (2021). Visualizing prediction correctness of eye tracking classifiers. In Bulling, A., Huckauf, A., Gellersen, H., Weiskopf, D., Bace, M., Hirzle, T., Alt, F., Pfeiffer, T., Bednarik, R., Krejtz, K., Blascheck, T., Burch, M., Kiefer, P., Dodd, M. D., and Sharif, B., editors, *2021 Symposium on Eye Tracking Research and Applications, ETRA 2020, Virtual Event, Germany, May 25-27, 2021, Short Papers*, pages 10:1–10:7. ACM.

[Purves et al., 2001] Purves, D., Augustine, G. J., Fitzpatrick, D., Katz, L. C., LaMantia, A.-S., McNamara, J. O., Williams, S. M., et al. (2001). Types of eye movements and their functions. *Neuroscience*, pages 361–390.

[Raatikainen, 2019] Raatikainen, P. (2019). Automatic detection of developmental dyslexia from eye movement data.

[Raatikainen et al., 2021] Raatikainen, P., Hautala, J., Loberg, O., Kärkkäinen, T., Leppänen, P., and Nieminen, P. (2021). Detection of developmental dyslexia with machine learning using eye movement data. *Array*, 12:100087.

[Rakoczi et al., 2013] Rakoczi, G., Duchowski, A., Casas-Tost, H., and Pohl, M. (2013). Visual perception of international traffic signs: influence of e-learning and culture on eye movements. In *Proceedings of the 2013 Conference on Eye Tracking South Africa*, pages 8–16.

[Ramot, 2001] Ramot, D. (2001). Average duration of a single eye blink. https://bionumbers.hms.harvard.edu/bionumber.aspx?s=y&id=100706&ver=0. [Online; accessed 11-August-2021].

[Rayner, 1998] Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372.

[Readman et al., 2021] Readman, M. R., Polden, M., Gibbs, M. C., Wareing, L., and Crawford, T. J. (2021). The potential of naturalistic eye movement tasks in the diagnosis of alzheimer's disease: A review. *Brain Sciences*, 11(11).

[Rello and Ballesteros, 2015] Rello, L. and Ballesteros, M. (2015). Detecting readers with dyslexia using machine learning with eye tracking measures. In Carriço, L., Mirri, S., Guerreiro, T. J., and Thiessen, P., editors, *Proceedings of the 12th Web for All Conference, W4A '15, Florence, Italy, May 18-20, 2015*, pages 16:1–16:8. ACM.

[Research, 2020] Research, E. (2020). Virtual Reality Market Research. https://explorerresearch.com/learn/consumer-research-techniques/virtual-reality-market-research/?fbclid=IwAR2g2FKQpFgGTfP4FoFfpFEb1UEmerBVD2AJZ7SMqe6mqCvo_iEJMZyV1dc. Accessed: 2020-03-30.

[Rigas et al., 2012] Rigas, I., Economou, G., and Fotopoulos, S. (2012). Human eye movements as a trait for biometrical identification. In *IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems, BTAS 2012, Arlington, VA, USA, September 23-27, 2012*, pages 217–222. IEEE.

[Rigas and Komogortsev, 2017] Rigas, I. and Komogortsev, O. V. (2017). Current research in eye movement biometrics: An analysis based on bioeye 2015 competition. *Image Vision Computing*, 58:129–141.

[Rish et al., 2001] Rish, I. et al. (2001). An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46.

[Romrell, 2014] Romrell, D. (2014). Gender and gaming: A literature review. In *annual meeting of the AECT International Convention, Hyatt Regency Orange County, Anaheim, CA*, pages 11–22.

[Rupp and Wallen, 2007] Rupp, H. A. and Wallen, K. (2007). Sex differences in viewing sexual stimuli: An eye-tracking study in men and women. *Hormones and behavior*, 51(4):524–533.

[Sæther et al., 2009] Sæther, L., Van Belle, W., Laeng, B., Brennen, T., and Øvervoll, M. (2009). Anchoring gaze when categorizing faces' sex: evidence from eye-tracking data. *Vision research*, 49(23):2870–2880.

[Salvucci and Goldberg, 2000] Salvucci, D. D. and Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the Eye Tracking Research & Application Symposium, ETRA 2000, Palm Beach Gardens, Florida, USA, November 6-8, 2000*, pages 71–78.

[Sammaknejad et al., 2017] Sammaknejad, N., Pouretemad, H., Eslahchi, C., Salahirad, A., and Alinejad, A. (2017). Gender classification based on eye movements: A processing effect during passive face viewing. *Advances in cognitive psychology*, 13(3):232.

[Sargezeh et al., 2019] Sargezeh, B. A., Tavakoli, N., and Daliri, M. R. (2019). Gender-based eye movement differences in passive indoor picture viewing: An eye-tracking study. *Physiology & behavior*, 206:43–50.

[Sargolzaei et al., 2016] Sargolzaei, A., Abdelghani, M., Yen, K. K., and Sargolzaei, S. (2016). Sensorimotor control: computing the immediate future from the delayed present. *BMC bioinformatics*, 17(7):501–509.

[Savitzky and Golay, 1964] Savitzky, A. and Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639.

[Schafer, 2011a] Schafer, R. W. (2011a). What is a savitzky-golay filter? [lecture notes]. *IEEE Signal Process. Mag.*, 28(4):111–117.

[Schafer, 2011b] Schafer, R. W. (2011b). What is a savitzky-golay filter?[lecture notes]. *IEEE Signal processing magazine*, 28(4):111–117.

[Schleicher et al., 2008] Schleicher, R., Galley, N., Briest, S., and Galley, L. (2008). Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired? *Ergonomics*, 51(7):982–1010.

[Scholkopf et al., 1999] Scholkopf, B., Mika, S., Burges, C. J., Knirsch, P., Muller, K.-R., Ratsch, G., and Smola, A. J. (1999). Input space versus feature space in kernel-based methods. *IEEE transactions on neural networks*, 10(5):1000–1017.

[Schröder et al., 2020] Schröder, C., Zaidawi, S. M. K. A., Prinzler, M. H. U., Maneth, S., and Zachmann, G. (2020). Robustness of eye movement biometrics against varying stimuli and varying trajectory length. In Bernhaupt, R., Mueller, F. F., Verweij, D., Andres, J., McGrenere, J., Cockburn, A., Avellino, I., Goguey, A., Bjøn, P., Zhao, S., Samson, B. P., and Kocielnik, R., editors, *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–7. ACM.

[Schuckers et al., 2015] Schuckers, S., Cannon, G., and Tekampe, N. (2015). Fido biometrics requirements. Accessed: 2021-04-04.

[Schwenker, 2000] Schwenker, F. (2000). Hierarchical support vector machines for multi-class pattern recognition. In *KES'2000. Fourth International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies. Proceedings (Cat. No. 00TH8516)*, volume 2, pages 561–565. IEEE.

[Seha et al., 2021] Seha, S. N. A., Hatzinakos, D., Zandi, A. S., and Comeau, F. J. E. (2021). Improving eye movement biometrics in low frame rate eye-tracking devices using periocular and eye blinking features. *Image and Vision Computing*, 108:104124.

[Sen and Megaw, 1984] Sen, T. and Megaw, T. (1984). The effects of task variables and prolonged performance on saccadic eye movement parameters. In *Advances in Psychology*, volume 22, pages 103–111. Elsevier.

[Shen et al., 2021] Shen, R., Zhan, Q., Wang, Y., and Ma, H. (2021). Depression detection by analysing eye movements on emotional images. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7973–7977. IEEE.

[Singh et al.,] Singh, H., Bhatia, J., and Kaur, J. Eye tracking based driver fatigue monitoring and warning system. inpower electronics (iicpe). In *2010 India International Conference*, pages 1–6.

[Sokolova et al., 2006] Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006). Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence*, pages 1015–1021. Springer.

[Sperry, 1950] Sperry, R. W. (1950). Neural basis of the spontaneous optokinetic response produced by visual inversion. *Journal of comparative and physiological psychology*, 43(6):482.

[Stein, 2014] Stein, J. (2014). Dyslexia: the role of vision and visual attention. *Current developmental disorders reports*, 1(4):267–280.

[Steyerberg et al., 2010] Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., and Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128.

[Streiner, 1996] Streiner, D. L. (1996). Maintaining standards: differences between the standard deviation and standard error, and when to use each. *The Canadian journal of psychiatry*, 41(8):498–502.

[Student, 1908] Student (1908). The probable error of a mean. *Biometrika*, pages 1–25.

[Sun et al., 2016] Sun, Q., Xia, J., Nadarajah, N., Falkmer, T., Foster, J., and Lee, H. (2016). Assessing drivers' visual-motor coordination using eye tracking, gnss and gis: a spatial turn in driving psychology. *Journal of spatial science*, 61(2):299–316.

[Suroya and Al-Samarraie, 2016] Suroya, S. H. and Al-Samarraie, H. (2016). Gender differences in the visual prediction of dyslexia. In *Proceedings of the 2nd IEEE International Conference on Human Computer Interactions*. Saveetha University Chennai.

[Taschenbuch Verlag Schiffman, 2001] Taschenbuch Verlag Schiffman, H. (2001). Sensation and perception: An integrated approach.

[Tharwat, 2020] Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*.

[Togaçar et al.,] Togaçar, M., Ergen, B., and Cömert, Z. A deep feature learning model for pneumonia detection applying a combination of mrmr feature selection and machine learning models. irbm 1, 1–11 (2019).

[Vanderah and Gould, 2020] Vanderah, T. and Gould, D. (2020). *Nolte's The Human Brain E-Book: An Introduction to its Functional Anatomy*. Elsevier Health Sciences.

[Vanitha and Kasthuri, 2021] Vanitha, G. and Kasthuri, M. (2021). Dyslexia prediction using machine learning algorithms–a review. *International Journal of Aquatic Science*, 12(02).

[Vanston and Strother, 2017] Vanston, J. E. and Strother, L. (2017). Sex differences in the human visual system. *Journal of neuroscience research*, 95(1-2):617–625.

[Vassallo et al., 2009] Vassallo, S., Cooper, S. L., and Douglas, J. M. (2009). Visual scanning in the recognition of facial affect: Is there an observer sex difference? *Journal of Vision*, 9(3):11–11.

[von Helmholtz, 1925] von Helmholtz, H. (1925). 31. handbuch der physiologischen optik southall. *JPC*, 3:455.

[Wade et al., 2003] Wade, N. J., Tatler, B. W., and Heller, D. (2003). Dodge-ing the issue: Dodge, javal, hering, and the measurement of saccades in eye-movement research. *Perception*, 32(7):793–804.

[Wahba et al., 1993] Wahba, G., Wang, Y., Gu, C., Klein, M., et al. (1993). Structured machine learning for'soft'classification with smoothing spline anova and stacked tuning, testing and evaluation. *Advances in Neural Information Processing Systems*, 6.

[Wang et al., 2017] Wang, S., Woods, R. L., Costela, F. M., and Luo, G. (2017). Dynamic gaze-position prediction of saccadic eye movements using a taylor series. *Journal of vision*, 17(14):1–11.

[Wang et al., 2011] Wang, Y., Toor, S. S., Gautam, R., and Henson, D. B. (2011). Blink frequency and duration during perimetry and their relationship to test–retest threshold variability. *Investigative ophthalmology & visual science*, 52(7):4546–4550.

[Weerahandi, 1995] Weerahandi, S. (1995). Anova under unequal error variances. *Biometrics*, pages 589–599.

[Wei et al., 2017] Wei, L., Wan, S., Guo, J., and Wong, K. K. L. (2017). A novel hierarchical selective ensemble classifier with bioinformatics application. *Artificial intelligence in medicine*, 83:82–90.

[Wikipedia, 2022] Wikipedia (2022). Visual perception — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Visual%20perception&oldid=1063147901. [Online; accessed 28-January-2022].

[Witkovsky, 2013] Witkovsky, V. (2013). A note on computing extreme tail probabilities of the noncentral t distribution with large noncentrality parameter. *arXiv preprint arXiv:1306.5294*.

[Woolson, 2007] Woolson, R. (2007). Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials*, pages 1–3.

[Yamada and Kobayashi, 2018] Yamada, Y. and Kobayashi, M. (2018). Detecting mental fatigue from eye-tracking data gathered while watching video: Evaluation in younger and older adults. *Artificial intelligence in medicine*, 91:39–48.

[Yao et al., 2018] Yao, D., Calhoun, V. D., Fu, Z., Du, Y., and Sui, J. (2018). An ensemble learning system for a 4-way classification of alzheimer's disease and mild cognitive impairment. *Journal of neuroscience methods*, 302:75–81.

[Zaidawi et al., 2022] Zaidawi, S. A., Prinzler, M. H., Lührs, J., and Maneth, S. (2022). An extensive study of user identification via eye movements across multiple datasets. submitted in 2021.

[Zaidawi et al., 2020] Zaidawi, S. M. K. A., Prinzler, M. H. U., Schröder, C., Zachmann, G., and Maneth, S. (2020). Gender classification of prepubescent children via eye movements with reading stimuli. In Truong, K. P., Heylen, D., Czerwinski, M., Berthouze, N., Chetouani, M., and Nakano, M., editors, *Companion Publication of the 2020 International Conference on Multimodal Interaction, ICMI Companion 2020, Virtual Event, The Netherlands, October, 2020*, pages 1–6. ACM.

[Zandi et al., 2019] Zandi, A. S., Quddus, A., Prest, L., and Comeau, F. J. (2019). Non-intrusive detection of drowsy driving based on eye tracking data. *Transportation research record*, 2673(6):247–257.

[Zhu et al., 2020] Zhu, J., Wang, Z., Gong, T., Zeng, S., Li, X., Hu, B., Li, J., Sun, S., and Zhang, L. (2020). An improved classification model for depression detection using eeg and eye tracking data. *IEEE transactions on nanobioscience*, 19(3):527–537.