universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Eye-Movement Similarity: Scanpath Analysis for Recognizing Learning Disorders.

verfasst von | submitted by

Dott. Nicolò Filippucci

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 921
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Informatik
Betreut von | Supervisor:	Univ.-Prof. Dipl.-Ing. Dr. -Ing. Moritz Grosse-Wentrup

Acknowledgements

I would like to express my gratitude to everyone who contributed to the accomplishment of this thesis and supported me throughout my Master's studies.

First, this project would not have been possible without the supervision of Prof. Jozef Arato, who guided me through every stage of developing and writing the thesis. Over the years we worked together, he was always extremely available and helpful. His expertise and insights were fundamental to the success of this project. I am deeply grateful to him for believing in me and my ideas.

Many thanks also go to Prof. Ferenc Kemeny from the University of Graz, for sharing their data with us and for all the time and attention he dedicated to our collaboration.

I would also like to thank Prof. Moritz Grosse-Wentrup for giving me the opportunity to collaborate with the Neuroinformatics Group and with Prof. Arato. His valuable feedback helped me improve my work, and thanks to him, this experience has been one of the most formative and important of my entire academic journey.

To my family, friends, and relatives, thank you for always being there for me, despite the distance. In particular, a great thanks goes to my parents. If I was able to do all this, I owe it mostly to them.

Finally, a special thanks goes to my girlfriend, who supported me every day, especially during the writing of this thesis.

Abstract

Diagnosing learning disorders such as dyslexia is challenging due to subtle and often non specific symptoms, making early identification difficult for non-experts. Given their impact on educational attainment and personal development, early detection methods are crucial. Prior eye-tracking studies suggest that ocular behaviour reflects reading difficulties, yet many rely on single datasets and inconsistent techniques, raising concerns about stability and generalizability.

This project utilizes *PyEyeSim*, an open-source library offering comprehensive eye-movement analysis tools, including Gaussian Hidden Markov Models (GHMM) and novel Saccade-Angle Similarity measures. We systematically evaluate these methods in conjunction with state-of-the-art classifiers, specifically, a Multi-Layer Perceptron (MLP) and a Convolutional Neural Network (CNN) to identify the most effective approach for a generalizable diagnostic framework. Our analysis spans two distinct datasets, differing in language and stimulus characteristics: the first comprises 70 subjects (35 control, 35 dyslexic) reading three long textual stimuli; while the second, initially containing 200 subjects, focuses on 36 short sentences and yields two cleaned subgroups: 88 subjects from Munich (49 control, 39 dyslexic) and 83 from Graz (60 control, 23 dyslexic).

Training/testing splits followed 5-fold cross-validation (80/20% in the first dataset; 90/10% in the second).

Integrating the saccade-angle similarity measure as an input feature significantly enhanced the MLP model's performance. In the first dataset, this approach yielded notable improvements: average accuracy for Task 1 increased from $91.43\% \pm 5.35\%$ to $92.86\% \pm 4.52\%$; Task 2 maintained a similar result from $90.00\% \pm 7.28\%$ to $91.43\% \pm 8.33\%$ or even $90.00\% \pm 5.71\%$ (the same accuracy level but with reduced standard deviation); Task 3 instead improved from $88.57\% \pm 5.71\%$ to $90.00\% \pm 3.5\%$.

For the second Dataset, the MLP model, augmented with fixation features and the novel similarity score, achieved accuracies of 97.77% across the 5 folds, and remained stable above 90% considering 100 folds: $91.44\% \pm 9.54\%$ for the Munich subjects and $93.50\% \pm 7.39\%$ the Graz subjects. Even in this case outperformed results obtained without the similarity measure. While the CNN's performance consistently remained below that achieved using MLP, the GHMM demonstrated strong accuracy, ranging from 91.11% to 95.55% over the 5 folds, closely approaching the MLP's efficacy.

These results show that combining advanced eye-movement metrics with a tailored MLP architecture yields a robust and generalizable framework for early dyslexia detection across both long and short textual stimuli. For short stimuli, the GHMM offers a competitive alternative with added benefits of simplicity, interpretability, and explainability.

Kurzfassung

Die Diagnose von Lernstörungen wie Dyslexie ist aufgrund subtiler und oft unspezifischer Symptome anspruchsvoll, wodurch eine frühzeitige Erkennung durch Nicht-Expertinnen und -Experten erschwert wird. Angesichts der erheblichen Auswirkungen auf Bildungswege und persönliche Entwicklung sind Früherkennungsmethoden von zentraler Bedeutung. Frühere Eye-Tracking-Studien weisen darauf hin, dass das okulare Verhalten Rückschlüsse auf Leseschwierigkeiten erlaubt, beruhen jedoch häufig auf einzelnen Datensätzen und uneinheitlichen Analyseverfahren, was Fragen nach Stabilität und Verallgemeinerbarkeit aufwirft.

Dieses Projekt verwendet *PyEyeSim*, eine Open-Source-Bibliothek zur umfassenden Analyse von Augenbewegungen, die sowohl Gaussian Hidden Markov Models (GHMM) als auch ein neu entwickeltes Sakkadenwinkel-ÄhnlichkeitsmaSS umfasst. Diese Methoden werden in Kombination mit modernen Klassifikatoren, einem Multi-Layer Perceptron (MLP) und einem Convolutional Neural Network (CNN), systematisch untersucht um den leistungsfähigsten Ansatz für ein generalisierbares diagnostisches Rahmenwerk zu identifizieren. Grundlage bilden zwei unterschiedliche Datensätze: Der erste umfasst 70 Teilnehmende (35 Kontrollpersonen, 35 Dyslexie-Betroffene), die drei längere Textstimuli lesen. Der zweite Datensatz enthält ursprünglich 200 Teilnehmende mit 36 kurzen Sätzen und wurde nach Datenbereinigung in zwei Untergruppen aufgeteilt: 88 Personen aus München (49 Kontroll-, 39 Dyslexie-Gruppe) und 83 aus Graz (60 Kontroll-, 23 Dyslexie-Gruppe), die sich in Sprache und Stimuluscharakteristik unterscheiden. Die Datenaufteilung erfolgte mittels 5-facher Kreuzvalidierung (80/20 % im ersten Datensatz, 90/10 % im zweiten). Die Einbeziehung des Sakkadenwinkel-ÄhnlichkeitsmaSSes als zusätzliches Eingangsmerkmal führte zu einer signifikanten Leistungssteigerung des MLP-Modells. Im ersten Datensatz verbesserte sich die durchschnittliche Genauigkeit bei Aufgabe 1 von $91.43\% \pm 5.35\%$ auf $92.86\% \pm 4.52\%$, bei Aufgabe 2 blieb sie stabil ($90.00\% \pm 7.28\%$ auf $91.43\% \pm 8.33\%$) und bei Aufgabe 3 stieg sie von $88.57\% \pm 5.71\%$ auf $90.00\% \pm 3.5\%$. Im zweiten Datensatz erreichte das erweiterte MLP eine Genauigkeit von 97.77% über 5 Folds und blieb auch über 100 Folds hinweg robust über 90%: $91.44\% \pm 9.54\%$ (München) und $93.50\% \pm 7.39\%$ (Graz). In allen Fällen übertraf das Modell Varianten ohne das ÄhnlichkeitsmaSS. Während das CNN konstant niedrigere Werte erzielte, zeigte das GHMM eine starke Genauigkeit zwischen 91.11% und 95.55% über die 5 Folds und näherte sich damit dem MLP an. Die Ergebnisse belegen, dass die Kombination fortgeschrittener Augenbewegungsmetriken mit einer angepassten MLP-Architektur ein robustes und generalisierbares Verfahren zur Früherkennung von Dyslexie über lange und kurze Textstimuli hinweg bietet. Für kurze Stimuli stellt das GHMM eine interpretierbare, einfache und zugleich leistungsfähige Alternative dar, die zusätzlich Transparenz und Erklärbarkeit in diagnostischen Anwendungen fördert.

Contents

List of Tables

List of Figures

List of Algorithms

1. Introduction

The eyes are often described as 'windows to the mind' revealing subtle clues about cognition and emotion through gaze patterns[4]. This connection between vision and cognition has fascinated researchers since the late 19th century, when pioneers like Émile Javal first observed and documented systematic patterns in reading eye movements[5]. From these early observations using direct visual inspection, the field has evolved through mechanical recording devices in the 1950s to today's sophisticated digital tracking systems, capable of capturing microsecond-level temporal resolution and fraction-of-degree spatial accuracy[6]. Within these increasingly precise patterns, lies the potential to understand how our brains process information and, how these operations may differ between individuals with unique cognitive characteristics.

Research on eye movement analysis has already yielded significant insights into distinguishing between healthy individuals and those with cognitive or learning challenges. Studies consistently show that specific eye movement patterns can reliably indicate differences in cognitive processing, aiding in identifying learning disorders and other neurological conditions[4, 7, 8].

However, despite these promising outcomes, the accessibility of such tools remains limited. Advanced eye-tracking systems and their specialized setups often come at a high cost, making them impractical for widespread use outside well-funded research environments. This financial barrier restricts broader application, limiting the potential of eye movement analysis to help a wide range of people in real-world settings.

This study aims to use Scanpath analysis, as a diagnostic tool to detect subtle, atypical visual processing patterns that may indicate learning challenges within neurotypical populations. The primary objective is to develop an analytic framework capable of differentiating between typical and atypical eye movement behaviours. To achieve this, we employed two state-of-the-art models and two innovative approaches developed in the context of the open-source library PyEyeSim[1][2], which is used throughout this project to analyse and process all scanpath data: Saccade-Angle Similarity (a novel measure introduced by Sancarlo, R., Dare, Z., Arato, J., & Rosenberg, R.[9]), Hidden Markov Model (HMM)-based pattern recognition[10], Multi-Layer Perceptron (MLP) based on the work of Sedmidubsky et al. [11] and Convolutional Neural Network (CNN) following the architecture proposed by Neruil et al. [8]. These techniques will then be evaluated on two different data sets to assess their performance and generalisability.

The ultimate goal of this project is to develop a standardized, efficient, and adaptable pipeline for scanpath-based diagnostics, which will enable the application of this approach

[1] PyEyeSim official repository: `https://github.com/jozsarato/PyEyeSim`
[2] PyEyeSim fork used for this project: `https://github.com/nicofilippucci/PyEyeSim`

in clinical and educational contexts. We intend to improve this tool for wider, more affordable accessibility by simulating how well these diagnostic methods could work on different setup. This approach holds promise for standardize how learning disorders can be identified, making advanced diagnostic capabilities more accessible.

1.1. Motivation and Background

Learning disorders represent a critical challenge in contemporary educational and clinical environments. These conditions are often present with unclear symptoms that can remain undetected until significant educational complications appear, highlighting the need of a more accessible way for early and rapid detection.

My academic journey in computer science has provided a pivotal perspective on learning disorders. As one of the few students with a diagnosed Specific Learning Disorder (SLD) in my master's program, I found a stark contrast to my previous educational experiences, where instead I had encountered a greater prevalence of students with SLDs. This has stimulated critical reflection on the problems that still afflict people with learning disorder, particularly in the field of education. Then this has become a motivation for a deeper exploration of the identification and management of these disorder.

Eye-tracking technology has opened new pathways for exploring the unique visual and cognitive patterns associated with learning disorders. Research demonstrates that eye-movement analysis can provide profound insights into how individuals process visual information, revealing significant differences between those with learning disorders and neurotypical individuals. Specifically, advanced metrics such as fixation patterns, saccades, and scanpaths can uncover subtle variations in cognitive processing that may indicate underlying learning difficulties.[12, 13].

This study aims to conduct a comprehensive analysis of current challenges in learning disorder identification. The primary research objectives include:

1. Examining existing diagnostic methodologies
2. Exploring technological innovations that can enhance early detection
3. Developing strategies to improve diagnostic accessibility

The core of this research is to address critical gaps in learning disorder diagnosis methodologies. By developing more accessible identification strategies, the study seeks to accelerate identification processes by fostering more inclusive educational environments that support diverse cognitive profiles.

1.1.1. Eye Movement

Eye movements encompass the voluntary and involuntary movements of the eyes, which are essential for visual perception and the accurate interpretation of our environment. These movements are controlled by a complex interaction of neurological processes, including brain areas such as the frontal eye fields, the superior colliculus, and the

cerebellum. Eye movements are integral to activities such as reading, scanning the environment, and tracking moving objects.

Eye movement analysis can be used to diagnose and monitor neurological and psychiatric disorders, such as Parkinson's disease, schizophrenia, and concussion-related impairments[14]. Recorded scanpaths data are utilized to enhance our understanding of cognitive processes, visual attention, and to develop advanced human-computer interaction systems.

Components of Eye Movement

Eye movements can be categorized into several components, each with distinct characteristics and functions[15, 16]:

- **Saccades**: rapid, jerky movements that shift the focal point of the eye from one position to another, lasting only 20 to 50 milliseconds. They occur when we read, scan a scene, or follow a moving object. Saccades are crucial for redirecting our line of sight and are controlled by a sophisticated neural network involving the brainstem and cortical areas[15].
- **Fixations**: the moments when the eyes remain stationary and focused on a single point. During a fixation, visual information is gathered and processed. Fixations typically last for about 100-600 milliseconds and are interspersed with saccades[17, 18].
- **Smooth pursuit**: movements that allow the eyes to smoothly follow a moving target. Unlike saccades, these movements are continuous and are used to maintain a steady image of the moving object on the fovea, the central part of the retina with the highest visual acuity[19, 20].
- **Vergence movements**: disjunctive movements where the eyes move in opposite directions to obtain or maintain single binocular vision. These movements are essential for depth perception and focus adjustment when viewing objects at different distances[15, 21].
- **Vestibulo-Ocular Reflex (VOR)**: stabilizes gaze during head movements by producing eye movements in the opposite direction of head motion. This reflex is crucial for maintaining a stable visual environment, especially during rapid or dynamic activities[15, 22].

In Figure 1.1, we present an example of a scanpath. Using specialized hardware, such as an eye tracker, it is possible to capture a subject's gaze and construct a temporal structure comprising fixations and saccades.

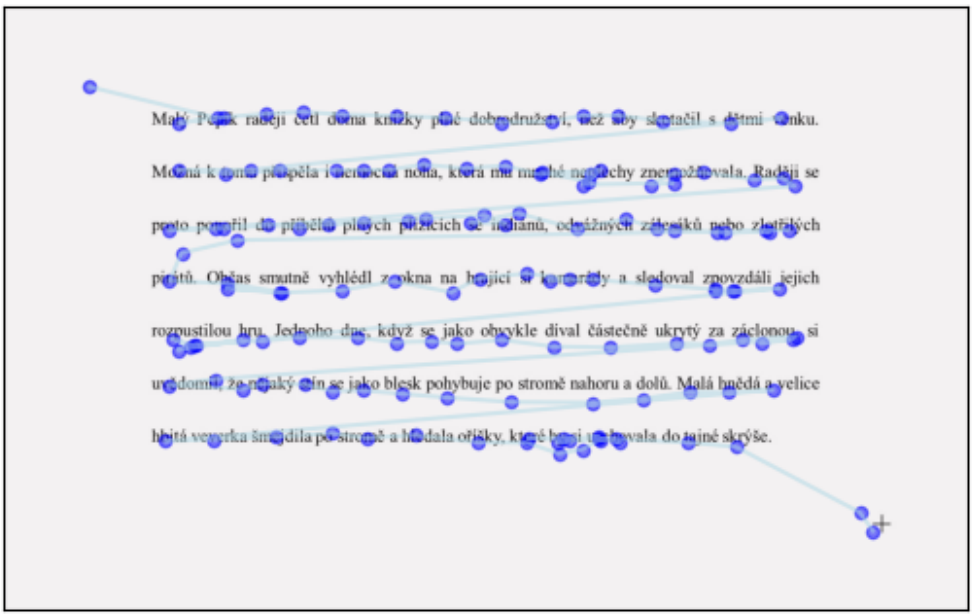


Figure 1.1.: Fixation and Saccades example from the ETDD70 Dataset [1]

.

1.1.2. Learning Disorders

Learning Disorders (LDs) encompass a range of neurological conditions that impair an individual's ability to process, store, or respond to information, despite possessing average or above-average intelligence. In the literature, terms like *learning disorder*, *learning disability*, and *learning difficulty* are often used interchangeably due to linguistic variations and differing organizational definitions, rather than representing distinct clinical entities.

These lifelong conditions manifest in diverse ways, affecting skills such as reading, writing, spelling, mathematics, and reasoning. While LDs are persistent, early diagnosis and tailored interventions can enable individuals to develop strategies that mitigate their challenges and promote academic and professional success.

The diagnostic criteria for LDs are primarily based on the guidelines in the Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5) [23] and the International Classification of Diseases, 11th Edition (ICD-11) [24]. Accurate diagnosis is critical not only for identifying affected individuals but also for ensuring access to necessary support and insurance coverage for diagnostic and therapeutic services. However, diagnostic criteria may vary across countries, potentially leading to inconsistencies in prevalence rates and treatment approaches.

LDs are generally classified into four main subtypes [23, 24, 25]:

- *Dyslexia*: difficulties with reading and decoding text.
- *Dysgraphia*: challenges related to writing, including spelling and organizing ideas.

- *Dysorthographia*: difficulty in correctly translating the sounds that make up words into graphic symbols.

- *Dyscalculia*: problems with mathematical concepts and calculations.

Each subtype presents unique challenges requiring individualized assessments to accurately identify specific difficulties.

Globally, LDs affect approximately 20% of the population [26], with significant implications in educational and professional contexts. Beyond academic performance, LDs can impact how individuals acquire and process knowledge, especially when taught using conventional methods. For example, individuals with dyslexia may improve reading comprehension over time but still struggle with reading fluency and spelling. These challenges highlight the need for personalized educational approaches that leverage strengths to maximize potential.

The social and emotional impact of LDs can also be profound. For instance, individuals with reading difficulties, may experience low self-esteem, social isolation, and increased emotional stress and depression, particularly if support systems are inadequate[27, 28]. Positive outcomes are linked to early interventions, supportive environments, and adaptive strategies that focus on individual strengths. Timely diagnosis and customized educational plans are essential in minimizing adverse effects and helping individuals achieve their long-term goals.

Recognizing Learning Disorders

Early recognition of LDs is crucial for timely support and intervention. While symptoms vary based on the type of disorder, there are common indicators that may suggest the presence of LDs:

- **Reading Difficulties:** Struggles with reading fluency, word decoding, and comprehension.

- **Writing Challenges:** Issues with spelling, grammar, and organizing thoughts on paper.

- **Mathematics Problems:** Difficulty with arithmetic, problem-solving, and grasping mathematical concepts.

- **Memory Issues:** Trouble remembering information, instructions, or sequential tasks.

- **Attention Difficulties:** Challenges in sustaining focus, organizing tasks, or completing assignments.

- **Language Delays:** Delays in speech or language development, or difficulty expressing ideas verbally.

It is important to note that symptoms can manifest differently among individuals, and not everyone with LDs will exhibit the same signs. Additionally, symptoms may overlap with other conditions, such as Attention-Deficit/Hyperactivity Disorder (ADHD) or Autism Spectrum Disorder (ASD), complicating the diagnostic process. Thus, a comprehensive evaluation by qualified professionals, such as psychologists or educational specialists, is necessary for accurate diagnosis and effective intervention planning.

By identifying LDs early and implementing appropriate support strategies, families, educators, and healthcare providers can collaborate to create a nurturing environment tailored to the individual's unique needs, thereby fostering positive outcomes.

Moreover, the use of eye-tracking technology can provide a more objective and quantitative assessment of cognitive processes, providing support to the diagnosis process.

1.1.3. Eye-Tracking Hardware

Eye-tracking technology allows the measurement of eye position, gaze direction, and eye movement patterns. This hardware typically combines optical sensors, infrared illumination, and specialized software to interpret eye movements and gaze positions accurately. Various techniques are utilized in these device, each with unique characteristics and associated costs.

Eye-tracking systems varies widely in terms of approach and application. Here are some of the primary techniques used:

- **Video-Oculography (VOG)**: VOG, or video-based eye-tracking, is the most common and versatile technique used today. It involves a camera, often with infrared (IR) capability, which captures high-resolution images of the eyes. By using IR light to illuminate the eyes, VOG systems can detect the reflection of light on the cornea (corneal reflection) and the position of the pupil to calculate gaze direction and movement. VOG hardware is known for its accuracy and is widely used in research and consumer applications.
  - **Cost Range:** Prices for VOG-based hardware vary significantly. Entry-level VOG systems can be low-cost but usually handcrafted and not available on the market[29], while High-end research-grade VOG systems can range from $ 12,000 to $ 40,000[30], depending on accuracy and features.
- **Electro-Oculography (EOG)**: EOG is an older method of eye-tracking that relies on measuring the electrical potentials between electrodes placed near the eyes. Eye movements generate small electrical changes in the surrounding muscles, and these variations in electrical signals can be used to determine the position and orientation of the eyes. EOG is less dependent on environmental lighting and can be used in low-visibility settings. However, it is not as precise as VOG and has largely been replaced by optical methods.
  - **Cost Range:** EOG-based systems are generally more affordable than VOG

systems, ranging from $ 300 to $ 1,500[3] [4]. However, they are less common in modern consumer and research applications due to lower resolution and accuracy.

- **Infrared-based Remote Eye Trackers**: Remote eye-tracking uses infrared light to capture eye movements from a distance, eliminating the need for wearable hardware. Remote systems are convenient for applications where physical contact with the participant or user is not ideal. They are frequently used in marketing research, web usability studies, and human-computer interaction research. The lack of direct contact can reduce accuracy compared to head-mounted solutions, but remote systems offer ease of use and flexibility.
  - **Cost Range:** Low-range remote eye trackers typically cost around $ 1,000 [5]. Instead for High-end remote systems (for example EyeLink 1000 Plus [6]) it is difficult to find an exact price since it is necessary to formally request a quote, anyway these device can cost $ 20,000 or more.

- **Embedded and Wearable Eye Trackers**: Wearable eye trackers are integrated into headsets, glasses, or virtual reality (VR) devices, allowing for mobility and ecological validity in real-world studies. These trackers are often used in research fields where users need to be mobile, such as sports science, vehicle research, or consumer studies. They combine VOG technology in a compact form, making them highly versatile for various real-world applications.
  - **Cost Range:** The price of wearable eye trackers varies depending on the quality and sophistication. Entry-level wearable devices may cost between $ 500 and $ 3,000[7], while research-grade may cost from $ 5,000 to $ 15,000[8].

Problem in devices accuracy

The accuracy of eye-tracking devices is a critical factor in determining their suitability for different applications. Typically, it is measured in terms of spatial resolution, temporal resolution, precision, and drift.

The accuracy can be influenced by various factors, including the type of technology used, the quality of the hardware, and the calibration process. In the following, we discuss in more detail the most important challenges that may be encountered when using these devices:

[3] pluxbiosignals (EOG) Sensor: `https://www.pluxbiosignals.com/products/electrooculography-eog-sensor-1?pr_prod_strat=e5_desc&pr_rec_id=d6e70b4c7&pr_rec_pid=7029144322239&pr_ref_pid=7023032795327&pr_seq=uniform`

[4] BlueGain EOG Biosignal Amplifier: `https://www.crsltd.com/tools-for-vision-science/eye-tracking/bluegain-eog-biosignal-amplifier/`

[5] GP3 Eye Tracker:`https://www.gazept.com/product/gazepoint-gp3-eye-tracker/?srsltid=AfmBOoqWGBu3ZyzPSb0_EKRkAayN7HMIZOfmwdwq2uWC-Kz2oDErKRz7&v=7d0db380a5b9`

[6] EyeLink 1000 Plus:`https://www.sr-research.com/eyelink-1000-plus/`

[7] Kexxe Eye Tracking Glasses: `https://kexxu.com/products/kexxe-eye-tracking-glasses`

[8] NEON: `https://pupil-labs.com/products/neon`

- **Lighting Conditions**: Changes in lighting can affect the performance of eye-tracking devices, particularly those that rely on infrared illumination. Bright or uneven lighting can interfere with the detection of eye movements and reduce tracking accuracy.
- **Head Movements**: Rapid head movements or changes in head position can disrupt the calibration of eye-tracking devices and lead to tracking errors. Some eye trackers are more sensitive to head movements than others, affecting their accuracy in dynamic environments.
- **Calibration Quality**: The accuracy of eye-tracking devices depends on the quality of the calibration process. Proper calibration is essential for accurate gaze estimation and eye movement tracking. Calibration errors can lead to inaccuracies in gaze data and affect the reliability of eye-tracking measurements.

All the eye-tracking devices presented suffer from high costs, custom or invasive hardware, or inaccuracy under real-world conditions[31]. In controlled research settings with high-quality, expensive hardware and meticulous calibration processes, these devices can achieve exceptional precision. However, practical applications present substantial challenges that can compromise measurement accuracy. The primary limitations in real-world scenarios stem from hardware constraints. Unlike research environments with elaborate and expensive equipment, typical applications struggle to replicate the same level of technical sophistication. These variations can introduce meaningful deviations in eye-tracking performance and reliability.

This analysis conclusively demonstrates the current impracticality of developing an affordable eye-tracking assessment that maintains high-quality technical standards. The prohibitive hardware costs create a significant barrier to widespread implementation, particularly for individuals and organizations with limited financial resources.

1.2. State of the Art

Studies have demonstrated that patterns in eye movement such as fixation duration, saccadic velocity, and scanpath variability are significantly correlated with specific learning challenges. For instance, children with dyslexia often exhibit longer fixation times and more frequent regressions while reading compared to their neurotypical peers. Using this information, researchers can extract and analyse important features. Various classification methods, including machine learning algorithms (particularly deep learning models), were then employed to identify these nuanced patterns, offering potential for early and automated diagnosis.

Despite the promise, the performance of these models and in particular, the evaluation and comparison of the most innovative methods and their applicability in real-world contexts, remains an area that requires deeper analysis. There are many methods today that present very positive classification results; however, there is a lack of sufficient tests regarding their generalizability. It will therefore be the aim of our project to

investigate this issue by considering the most relevant methods and in comparison of multiple datasets, each one with different characteristics.

Moreover, approaches like infrared eye-trackers, typically require specialized laboratory equipment to capture precise gaze data, which limits the accessibility and scalability of these methods. Recent results have shown that it is possible to obtain gaze data with a good level of quality (compared to laboratory hardware) while still maintaining low costs. This suggests that eye-tracking could become a low-cost, scalable diagnostic tool suitable for widespread use. This aspect will be another important point to consider in our work.

1.2.1. Distinctive features in individuals with Learning Disorders

Individuals with learning disorders, particularly dyslexia, exhibit a range of distinctive eye-movement characteristics when engaged in reading and related tasks. Compared to neurotypical readers, those with dyslexia show a higher number of fixations, longer fixation durations, shorter and more variable saccades, and elevated regression rates, reflecting a laborious and effortful visual sampling strategy[12, 13, 32, 33]. In addition, global scanpath dynamics indicate that the sequences of dyslexic readers are less consistent in space and time than those of typical readers.[34]

We will now present the most important and distinctive features that have been identified.

Fixation Metrics

Firstly, we can see important differences in the number of fixations and their duration, which are generally higher for individuals with learning disorders. These subjects make significantly more fixations per word and per sentence, which suggests increased effort in processing orthographic information[12]. Furthermore, dyslexic readers demonstrate prolonged mean fixation durations and greater total fixation time over texts, indicating a slower processing speed and a diminished benefit from parafoveal preview[33]. Additionally, the variability in their fixation durations is elevated, pointing to a more inconsistent allocation of visual attention across words. [13]

Saccade Metrics

Dyslexic readers produce shorter saccades (reduced jump length between fixations), implying a narrower perceptual span and reliance on serial, rather than parallel, decoding strategies. In addition, it can be seen that these subjects show a greater variance in saccade amplitude, both in comparison with Control subjects and with other subjects with the same disorders. [12, 13]

Regression Metrics

Dyslexic individuals regress more frequently to earlier text positions, often multiple times on complex words or sentences, indicating difficulty in decoding and integrating linguistic information.[12, 13]

Global Dynamics and Temporal Patterns

Dyslexic reading time series exhibit lower recurrence rates and determinism, indicating less patterned and more unstable temporal dynamics in fixations and saccades. As a consequence, it is possible to observe higher Entropy and shorter mean line lengths in dyslexic eye-movement data, signalling greater irregularity and reduced temporal coherence in gaze behaviour.[34]

1.2.2. Learning Disorder Detection using Eye-Tracking

Eye-tracking technology is a promising technique for identifying healthy subjects and those who have certain neurological or cognitive impairments, according to recent research [14]. Numerous studies have utilized this technique to explore and analyse the differences in visual processing behaviours associated with learning difficulties, such as dyslexia.

To better understand and explore these differences, researchers have employed a variety of analytical methods on eye-tracking data collected from both neurotypical participants and those diagnosed with learning disorders. Techniques range from traditional statistical analysis to more sophisticated machine learning algorithms aimed at identifying patterns indicative of cognitive impairments. By examining the visual scanning pathways and gaze metrics discussed in the previous section, these studies have been able to identify atypical eye movement behaviours that may serve as markers for dyslexia.

Statistical Analysis

Many works in research have reported significant statistical differences in the visual scanpaths of neurotypical readers and individuals with specific learning disorders, including dyslexia. These differences have been observed across a range of eye-tracking metrics, such as fixation count, saccade length, and regression frequency. Studies have further shown that such effects can vary depending on the characteristics of the presented stimulifor example, the use of different fonts, including dyslexia-friendly designs such as OpenDyslexic [35, 36, 37]. Analyses consistently highlight atypical gaze behaviours in dyslexic readers, reinforcing the role of eye-tracking as a reliable tool for distinguishing between typical and impaired visual processing [38, 39, 40].

In particular, in the work of Franzen, L., Stark, Z. & Johnson, A.P.[13] grouplevel differences between readers with and without dyslexia were assessed using a combination of generalized linear mixedeffects models (GLMMs), independentsamples ttests, and standardized effect size with 95% confidence intervals. Reading duration was analysed trialwise in a GLMM with fixed effects of Group (Dyslexia vs. Control), Font (Times New Roman vs. OpenDyslexic), and their interaction, and random intercepts for Participant and Text. The main effect of Group on reading duration was highly significant (Xš = 13.431, df = 1, $p < 0.001$), confirming that adults with dyslexia read more slowly than controls. Wordsperminute reading rates were compared using twosided independentsamples ttests ($t(65) = 20.51$, $p < 0.0001$), yielding a large effect size (Hedges g = 1.67, 95% CI [1.486, 1.858]) indicative of substantially slower reading in the dyslexia

group. For eyemovement metrics, Hedges g effect size was computed, finding differences between the two groups for almost all of the metrics considered.

Machine Learning

The application of machine learning and artificial intelligence has significantly enhanced the potential of eye-tracking systems for automated LD detection. By leveraging large datasets, researchers can build models capable of identifying patterns in eye-movement that are indicative of learning disorders.

Below we explain in detail the experimental setups, key results, and strengths and limitations of relevant studies.

The first notable study to mention, is that of Nilsson Benfatto et al. [12], performed on 185 Swedish secondgraders (97 highrisk, 88 lowrisk), reading gradeappropriate text on screen while eye movements were recorded for under one minute per subject. Then, over 170 metrics were extracted, and predictive models (e.g., random forests, logistic regression) were trained and evaluated via repeated crossvalidation.

The best classifier achieved 96% overall accuracy with balanced sensitivity and specificity, demonstrating that brief recordings suffice for reliable dyslexia screening.

The analysis and results allow us to consider some pros and cons of this method:

- *Pros:* Highly interpretable metrics linked to known oculomotor behaviours;
- *Cons:* Relies on handcrafted features which may omit subtle patterns; performance may degrade on different languages or orthographies without retuning feature extraction parameters.

Neruil et al. [8] conducted an experiment on the same dataset. Raw timeseries signals were interpolated to a fixed length and transformed into magnitudespectrum representations. A CNN then processed either the timedomain signal or its frequencydomain spectrum, using only minimal preprocessing and 100fold crossvalidation to guard against overfitting. The best modeltrained on the interpolated magnitude spectrumachieved 96.6% mean accuracy, outperforming the previous benchmark.

Processing of raw signals captures non-linear relationships without manual feature engineering and rigorous crossvalidation ensures reliable performance estimates. Moreover, since the magnitudespectrum is symmetric only the first half is required, this representation reduced the amount of process data by a factor of two.

However, this method also presents some problems:

- CNNs are black boxes, offering limited interpretability
- Requires substantial computational resources and a relatively large dataset for training
- Frequencydomain focus may obscure finegrained temporal dynamics

Vajs et al. [41] presented another approach using a CNN. 30 Serbian children (15 dyslexic, 15 controls; ages 713) read the same text under 13 background/overlay colour configurations. Raw gaze coordinates were encoded as coloured images (each pixel representing spatial position and time), then fed into a VGG16based CNN.

The best pipeline attained 87% classification accuracy, demonstrating that visually encoded gaze trajectories can effectively drive deep learning models with minimal preprocessing.

- *Pros:* Novel visual encoding makes full use of spatialtemporal patterns; colour variations help assess robustness across display conditions.
- *Cons:* Small sample size limits statistical power; colouredimage encoding may introduce artifacts unrelated to actual reading behaviours.

Rello & Ballesteros [42] examined 97 Spanish speakers (48 dyslexic, 49 controls), each read 12 texttypeface passages while gaze data were captured. From the raw eyetracking logs, 22 summary measures (e.g., total reading time, mean fixation duration, regression count) plus age were selected via recursive feature elimination. A support vector machine (SVM) classifier was trained with 10fold crossvalidation and achieved 80.18% accuracy in distinguishing dyslexic from typical readers.

- *Pros:* Uses established machine learning pipeline with transparent feature selection.
- *Cons:* Limited dataset size and language specificity may constrain generalizability.

Then, continuing with the work proposed by Neruil et al and Rello & Ballesteros, there are many other studies that have applied machine learning techniques to analyse this type of data, including: logistic regression, SVM, and random forest[43, 44, 45]. All these studies have yielded very promising results, even though they share certain problems or weaknesses:

- Often, the features extracted from eye-tracking systems differ from study to study, making it difficult to compare the same methods on different datasets.
- The studies present only one database, so the evaluation of the various models is limited and lacks generalizability.

A recent study by Sedmidubsky et al. [11] utilized the newly released ETDD70 Dataset [1]. 70 Czech fourth-graders (35 dyslexic, 35 non-dyslexic; ages 910) completed three reading taskssyllable matrix, meaningful text, and pseudo-text. Binocular gaze was recorded at 500 Hz and processed with the i2mc fixation detector (min. duration = 40 ms) to extract fixations, saccades, regressions, and a suite of whole-task and Region of Interest (ROI) features.

Among all classifiers, the Multi Layer Perceptron model achieved the highest dyslexia classification accuracy, approximately 90 % on unseen participants, outperforming the visualization-based representations classified by ResNet18/50.

- *Pros:* Training the MLP model for 20 epochs takes less than 1 second and the model size occupies only around 1 MB.
- *Cons:* The selected features are based on combination of wholetask and ROI features only work for stimuli with long texts, extending over several lines.

In the same way, there are new studies who have integrated the use of neural networks and Eye-Tracking Technology into the process of classifying dyslexic subjects [46, 47, 48]. By achieving good performance, ranging from 86.65% and 97.7% of accuracy, they suggest that these methods represent a meaningfu; advancement in automated dyslexia classification. Demonstrating that neural network approaches can achieve high accuracy while maintaining computational efficiency suitable for real-world applications.

Considerations

Despite the promise of eye-tracking technology and the machine learning models, several challenges remain:

- **Data Quality and Availability**: High-quality eye-tracking data is essential for accurate analysis, but it is currently very difficult to find a large dataset specifically for learning disorder detection.
- **Generalization of Models**: Machine learning models trained on specific datasets may not generalize well to other populations, leading to reduced reliability in real-world applications.
- **Hardware Limitations**: Eye-tracking systems require specialized hardware, which can be costly and may not be accessible to all individuals.
- **Real-Time Processing**: To be effective in real settings, eye-tracking systems must be capable of real-time processing to provide rapid feedback at the end of the test.
- **Interpretability**: While machine learning models can achieve high accuracy, their decision-making processes are often opaque, making it difficult to understand the underlying factors contributing to a diagnosis.

The results consistently suggest that deviations in eye movement patterns can be indicative of underlying cognitive challenges, making eye-tracking a valuable tool for early detection and intervention. In particular, the identification of irregular gaze patterns in both children and adults with dyslexia highlights the potential of eye-tracking analysis to support more accurate and efficient diagnosis of learning disorders.

Given the large number of studies on the subject and the very promising results that have emerged from each of them, it is necessary to define a model that is balanced in terms of accuracy and speed of execution so that this system can be integrated into a real-time executable application.

In the following chapters, we will study in more detail two new approaches presented in the introductory chapter and compare them with existing methods.

2. Datasets

In this chapter, we will present and analyse the two datasets used for the project, ETDD70[1] and a set of data collected by Gangl, M. et al. in their study of Lexical Reading in Dysfluent Readers of German[32].

Firstly, it is important to highlight that publicly available datasets in this field of research are quite limited. Among the datasets under review, ETDD70 is the only one that is openly accessible online. In contrast, access to the second dataset required direct communication with the University of Graz.

Considering that the two datasets come from different studies and countries, there are many differences, including the language and type of stimuli used. This will allow us to evaluate all the algorithms under different setups, in order to test and improve their generalizability.

Another important difference is that, as we have seen in the subsection 1.2.2, the ETDD70 Dataset has already been tested with some methods of automatic classification, while for the second dataset there are no studies with such investigations.

In the following sections we will focus on the explanation and analysis of the datasets' subjects, tasks, stimuli, and other characteristics: how it is structured (i.e., length, number of lines, and any other interesting features), how the experiment was conducted, and the type of words used.

The type of words that are considered in this context are Non Words, Pseudo Words and Normal Words. Below, we explain their differences in more detail:

- A Non Word is a string of letters that follows the phonotactic rules of a language (i.e., it could be a word in that language) but has no meaning.

- Pseudo Words or Pseudohomophones are Nonwords that sound like words, the pronunciation matches a real word, but spelling does not.

For example, in English, the word *'Phone'* can be transformed in the Pseudohomophone *'Fone'* and the corresponding Non Word *'Phote'*. Another example from one of the dataset used in this study, is the German word *'Mittwoch'* (Wednesday) that can be changed in the Pseudohomophone *'Mitwoch'* and the Non Word *'Littmoch'*.

These type of words require non lexical route processing and it has been proven in previous studies[49][50] that individuals with dyslexia exhibit specific deficits when decoding non-words, beyond what is observed in reading-level or age-matched control groups.

2.1. ETDD70: EyeTracking Dyslexia Dataset

The **ETDD70** Dataset [1] comprises highfrequency eyetracking recordings from a cohort of 70 Czech elementary school children (aged 910), equally divided into dyslexic and neurotypical readers. The dataset records eye movements while doing three Czech text-reading tasks: reading syllables (Task 1), reading meaningful texts (Task 2), and reading pseudo-texts (Task 3).

2.1.1. Participants

The data were collected from 70 participants (35 dyslexic, 35 typical readers) in 4^{th} grade.

Suitable participants were recruited in cooperation with a psychological counselling centre, which facilitated the recruitment of pupils diagnosed with dyslexia. Non-dyslexic readers were recruited in collaboration with the counselling services of selected primary schools.

2.1.2. Stimuli

Three reading tasks were digitized from Czech standardized dyslexia diagnostics, with careful control of font, spacing, and background:

- **Syllables (Task 1):** Ninety Czech syllables organized in 9 columns and 10 rows, black *Times New Roman* on a gray background. Rows were linguistically stratified into open (consonant + vowel), closed (consonant + vowel + consonant), and various consonantal clustersmeaningful and meaninglessensuring balanced coverage of syllable types. Participants read aloud rowwise, terminating by fixating a final cross.
- **Meaningful Text (Task 2):** A sevenline narrative (six sentences). The text was set with a black font, doublespaced on gray background, with the same fixationcross closure. Participants read the entire passage aloud.
- **PseudoText (Task 3):** Seven lines of 15 artificially generated, Czechlike non-words. Formatting matched the meaningful text to isolate lexical processing. The goal was again to read the text aloud and smoothly.

The section A.1 contains images of the three stimuli, along with two example scanpaths for each stimulus and group.

2.1.3. Data Collection

Eye movements were recorded with a remote eyetracker at sampling rates of 250 Hz. Then, raw gaze data were postprocessed with the I2MC algorithm [51] for fixation detection (minimum duration 40 ms). This data is then used to extract event-based features such as fixations, saccades, regressions, and other statistical features.

2.1.4. Descriptive Data Analysis

We conducted a descriptive analysis of the scanpath data to obtain an overview of the number and distribution of fixations and saccades across the different stimuli, and to determine whether differences exist between the two groups. This will then be analysed in more detail with statistical tests in section 4.1.

In Figure 2.1 we can see a general overview of number of fixations, fixation duration and the length of the saccades for the three stimuli.

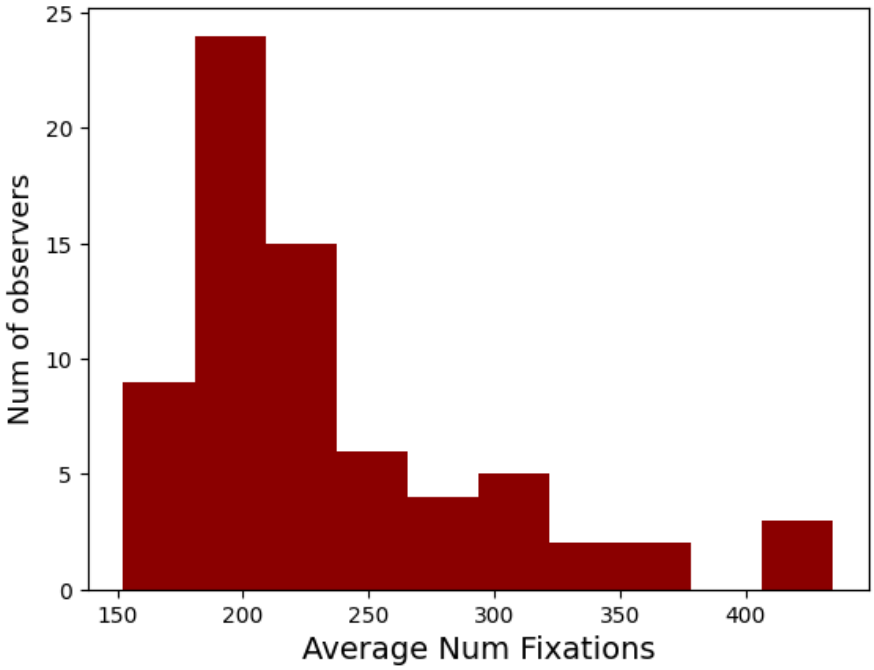
From the images it is possible to get some first insights on the data and the different stimuli. Firstly we can observe that in Figure 2.1a we have a normal distribution between 200 and 250 fixation with some outliers that have high number of fixation (>400). To be more precise there are 9 subjects that present very high fixations, all of them belong to the dyslexic group.

Comparing the stimuli, we observe that the first image (Task 1) shows a relatively low number of fixations but the greatest scanpath length. This aligns with the characteristics of the task, which contains more rows with words spaced farther apart. The second stimulus shows shorter fixation durations and a higher number of fixations, likely due to the increased number of words compared with the first stimulus.

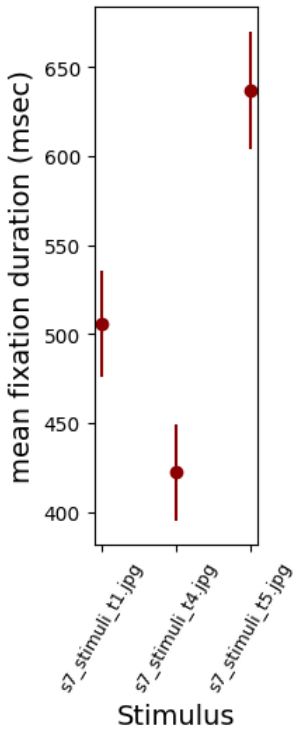
Finally, the third stimulus (containing nonwords) exhibits the longest fixation durations overall. Additionally, the standard deviation of the total scanpath length is noticeably larger than in the other two stimuli, suggesting greater variability in participants reading process.

Below we report also a summary of the global mean and standard deviation value obtained from the descriptive analysis:

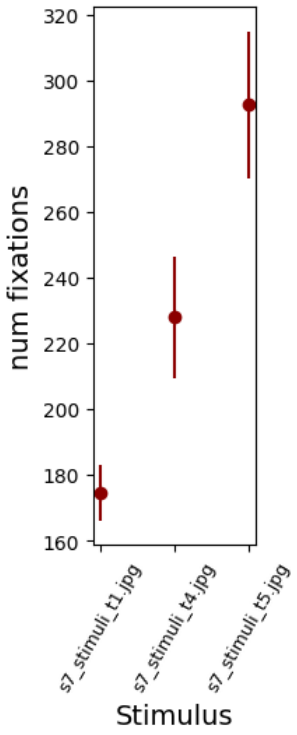
- Mean fixation number: 231.79 $\pm$ 64.12
- Mean saccade amplitude: 104.0 $\pm$ 13.3 pixels
- Mean scanpath length: 21945.6 $\pm$ 3936.5 pixels
- Mean fixation duration: 521.7 $\pm$ 112.6 msec



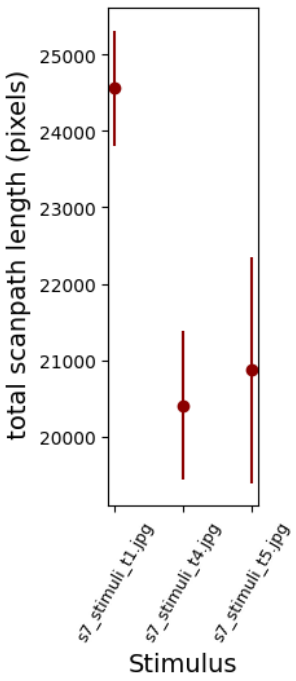
(a) Number of subjects by Average Number of Fixations



(b) Fixations Duration, Mean and Standard deviation



(c) Number of fixations, Mean and Standard deviation



(d) Scanpath Length, Mean and Standard deviation

Figure 2.1.: ETDD70: Descriptive Analysis

In Figure 2.2 we present a comprehensive comparison of eye-tracking performance metrics between Controlled Subjects (Group 0) and Dyslexic Subjects (Group 1). The analysis reveals systematic and consistent differences between the two groups across all measured parameters.

The dyslexic group, compared to the control group, demonstrates consistently higher fixation numbers and entropy, longer scanpath lengths, and lower saccade amplitudes across all three stimulus conditions.

The data reveals that Task 1 appears to be the most challenging for both groups, eliciting the highest number of fixations, longest scanpaths, and largest saccade amplitudes. Conversely, Task 3 seems to require more focused attention with shorter saccades but maintains the pattern of increased fixations in the dyslexic group. Another interesting observation is that Task 2 and 3, which are stimuli structured in a very similar way, while in some cases having similar standard deviations with only small variations in the mean, some important differences in values can be noted:

- Fixation Entropy of the two groups in Task 3 has close values to each other (the same patterns are visible in the Task 1).
- Mean and standard deviation of Saccade Amplitude seem quite lower in Task 3 compared to Task 2.

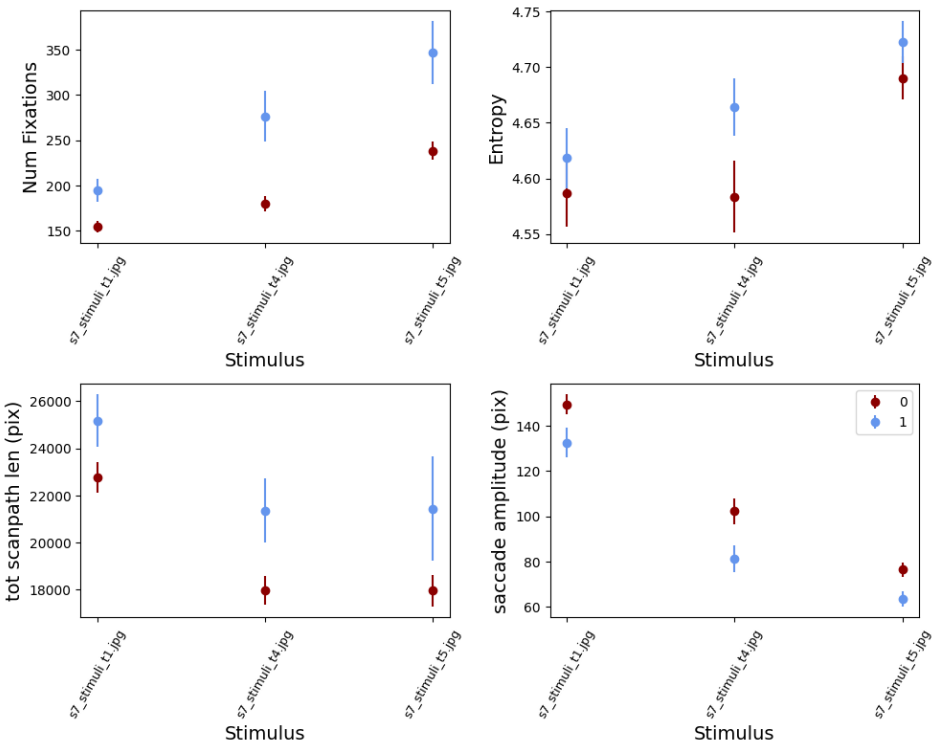


Figure 2.2.: ETDD70: Descriptive Analysis by Group

2.2. Dysfluent Readers of German

Gangl et al. [32, 52] investigated the reading strategies of German-speaking children in third and fourth grade who have different reading and spelling abilities. The dataset used in our study originates from their research.

With access to the data recorded by the eye-tracker and the subjects' information, it was possible to extract the fixations and organize them in a CSV file. In subsection 2.2.4 we will explore in more detail how we prepared the raw data and then used it in our study.

Data were collected at two collaborating sites, Graz (Austria) and Munich (Germany). The Dataset consist of 200 children who were grouped based on standardized assessments into the following categories: Typical Readers (TD); Isolated Spelling Deficit (SD); Isolated Reading Deficit (RD); and Combined Reading and Spelling Deficit (RSD).

Eye-tracking data were recorded while the children read aloud words, *pseudohomophones*, and *non-words*.

The study aimed to explore how orthographic knowledge, reflected by spelling performance, influences the application of lexical and sublexical reading strategies. This dataset provides a valuable foundation for examining the interplay between orthographic representations and reading fluency in a transparent orthography such as German.

In this project, after a preliminary analysis of the data, we will determine how to subdivide the group of subjects and how the various classification models perform. We will also investigate the advantages and disadvantages of each model, especially comparing the two groups of subjects.

2.2.1. Participants

There are 117 registered subjects from Munich and 83 from Graz. All children had German as first language, a nonverbal IQ at or above 85, and normal or corrected-to-normal vision. Children with a clinical diagnosis of attention deficit hyperactivity disorder or an increased score on a parental questionnaire on attention deficits were not admitted to the study.

In Table 2.1 we can see the exact number of subjects subdivided for Munich and Graz and then for the respective groups.

Group	**Munich**	**Graz**
Typical Readers	35	32
Isolated Spelling Deficit	21	24
Isolated Reading Deficit	21	8
Combined Reading and Spelling Deficit	37	15
No Group	3	4

Table 2.1.: Dysfluent Readers of German - Distribution of groups for Munich and Graz subjects.

2.2.2. Stimuli

In total, there are 36 stimuli containing text displayed in single lines in black on a white background in Arial font. Of these, 6 stimuli contain only normal words, 20 with pseudowords and nonwords, and 10 containing only nonwords.

Pseudohomophones were created by replacing one phonetically identical grapheme in a word, while nonwords were created by replacing one grapheme in each syllable of a word. The total set of items was divided into three blocks and assembled in four pseudo-randomized orders, which were randomly assigned to participants. Two of the blocks contained words and pseudohomophones (10 stimuli each), while the third block contained only nonwords (10 stimuli).

The stimuli containing only real words are labelled as practice tasks, since they were been used to let the participants familiarize themselves with the environment and perform the proposed reading tasks. We expect to see strong differences from these stimuli and the ones used in the actual test. This will be considered with particular attention during our descriptive analysis and especially during the classification test.

Currently, we do not have direct access to the stimuli images containing the text; the text was probably displayed automatically by the software used for collecting the data (RSResearch). However, it is possible to have a clearer explanation of each stimulus by referring to Table B.1 in the appendix section B.1, where it is also possible to visualize some fixation and saccade distribution.

2.2.3. Data Collection

Eye-movements of the dominant eye were recorded with an EyeLink 1000 Tower Mount eye-tracker in Graz and an EyeLink 1000 Plus Desktop Mount eye-tracker in Munich. Stimulus presentation was similar at both collaborating sites with an uppercase letter height of about 0.62ř of visual angle. Children put their forehead up against a forehead rest to minimize head movements. A 9-point calibration cycle at the beginning and after each break was used to ensure a spatial resolution of less than 0.5ř of visual angle. Children were instructed to read each item aloud at their own speed without making mistakes while their eye-movements were recorded. Reading aloud was chosen in order to control for reading accuracy. Reading errors were noted by the experimenter and corrected only during practice.

2.2.4. Data Preparation

We extracted the fixation data directly from the raw data recorded by the eye tracker, considering all trails with positive read time as valid.

Firstly, we converted the EyeLink Data File (EDF) into a more readable ASCII file (ASC) using the SR Research script EDF2ASC. Then from the ASC we analysed each subject and each trial creating a CSV file with all the details of the recorded fixations (starting [ms], end [ms], duration [ms] and average x and y coordinates of the fixation).

In the end we created a final CSV, containing this information and the groups to which each subject belongs. It is possible to load this file as a `pandas DataFrame` and use it in the *PyEyeSim* library for our tests.

Of all the recorded sessions, 141 had problems that invalidated and restarted the task. In our study, this can be considered problematic, since each time the subject restarts the session, they will improve at the reading task. We therefore considered this in our analysis and removed all the repeated trials.

The classification process involves multiple stimuli and missing subjectstimulus data pairs can create significant issues for the tests.

Regarding the Munich subjects, since we had enough subjects to test, we decided to do not include the child that had problems during at least one task; reducing the number of subjects from 117 to 88 (as reported in the abstract). Meanwhile, the Graz subjects with no valid data in at least one task where 57, and clearly remove them was not a valid option. In chapter 3 we will explain better how we handled this cases and prepared the input data for each classification model.

Differences between Munich and Graz Subjects

Although both the Munich and Graz data originate from the same set of stimuli, a careful inspection of Figure 2.3 reveals a pronounced visual distinction between the two groups of scanpaths. The trajectories corresponding to the Graz subjects and those from Munich occupy different spatial regions of the stimulus area and differ substantially in overall spread and alignment. Specifically, the Graz scanpaths appear more compact and slightly shifted relative to those from Munich, whose trajectories extend over a wider area. This indicates a systematic offset in fixation positions and an overall difference in scanpath magnitude.

Such spatial discrepancies are most likely attributable to variations in the experimental setup, particularly differences in screen size, resolution, or coordinate calibration across recording sites. Even minor inconsistencies in display parameters or the way gaze coordinates are mapped to screen space can produce global translation and scaling effects in the resulting fixation data. These effects are clearly visible in the plot, where the entire spatial configuration of one groups scanpaths appears displaced relative to the others.

This observation has important methodological implications. If data from the two recording sites were analysed jointly without accounting for these systematic differences, the resulting analyses such as statistical comparisons of fixation metrics, or machine learning classification could be biased by these recording specific artifacts. In such cases, a classifier might erroneously learn to distinguish between Graz and Munich subjects based on spatial characteristics induced by hardware or calibration differences, rather than on underlying cognitive or behavioural features of interest.

To mitigate this risk, the two groups were treated separately in all classification and statistical analyses. This separation ensures that the models focus on the relevant behavioural distinctions (e.g., between Dyslexic and Control readers). Additionally, analysing the datasets independently provides an opportunity to evaluate the robustness of the proposed classification methods under conditions of class imbalance, since the Graz

subset contains fewer Dyslexic participants than Controls. Overall, this careful handling of cross-site variability increases both the reliability and interpretability of the results.

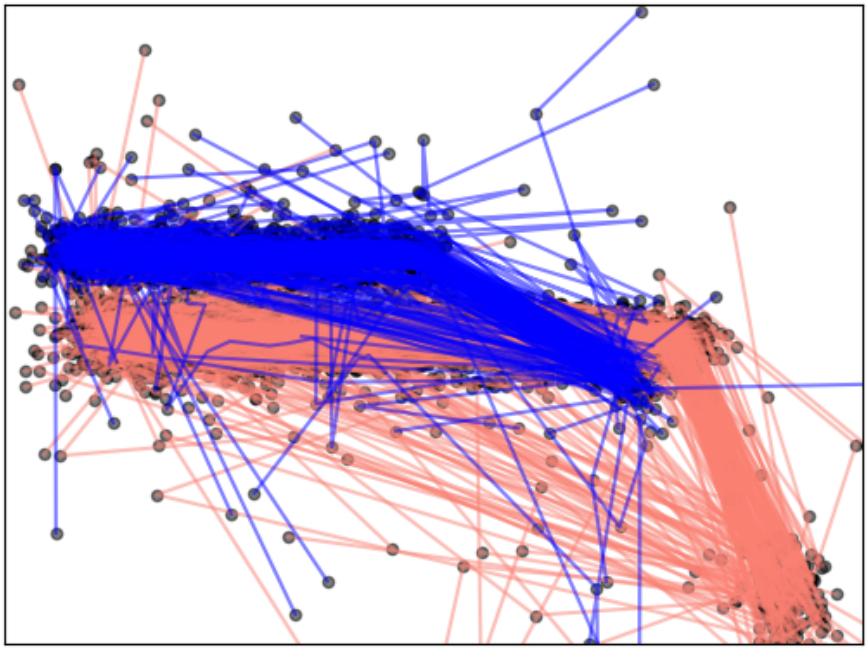


Figure 2.3.: Dysfluent Readers of German: Scanpath visualization of all Graz and Munich Subjects

2.2.5. Descriptive Data Analysis

As with the previous dataset, we performed descriptive analysis on the scanpath data. We also repeated the analysis on the two set of subjects separately, to see if it is possible to identify interesting differences.

In Figure 2.4, we can observe that the distribution of fixations is similar in both datasets. In this case, the stimuli are shorter, resulting in a generally lower number of fixations. From the descriptive analysis in Figure 2.5, it is possible to visually observe differences in the subgroup of stimuli. In Figure 2.5b and especially in Figure 2.5a, the stimuli with pseudowords and nonwords (form 1 to 10 and form 21 to 30) have similar mean and standard deviations. Moreover, the stimuli with only nonwords seem to be the most demanding tasks, having on average a higher number of fixations and very high fixation durations. In Figure 2.5c, all stimuli (excluding the last 6, which contain only words)present similar total scanpath lengths, which is reasonable as all stimuli have almost the same length. Finally, the only-word stimuli present strange patterns with a low number of fixations and high variability in fixation duration. This is probably due to the nature of the last 6 stimuli which, as mentioned above, were used to familiarize the participant with the environment and the tasks to be performed.

The exact values derived from the the analysis are reported below:

- Mean fixation number: 37.13 ± 10.07
- Mean saccade amplitude: 81.1 ± 20.0 pixels
- Mean scanpath length: 2549.3 ± 649.7 pixels
- Mean fixation duration: 382.9 ± 67.3 msec

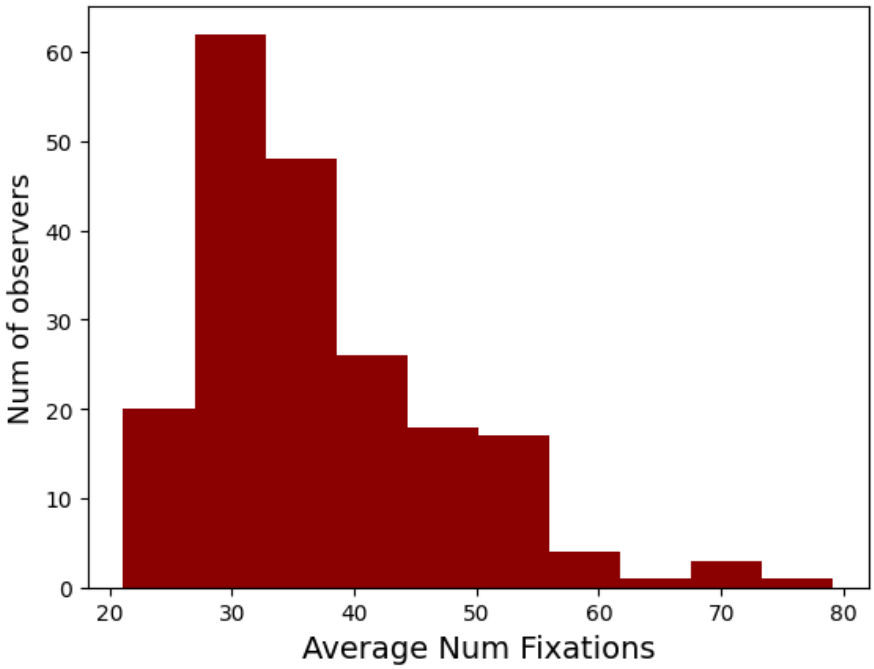
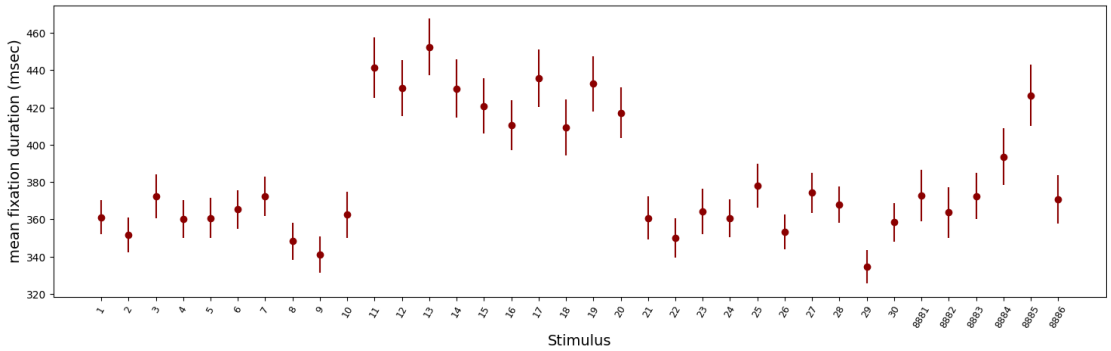
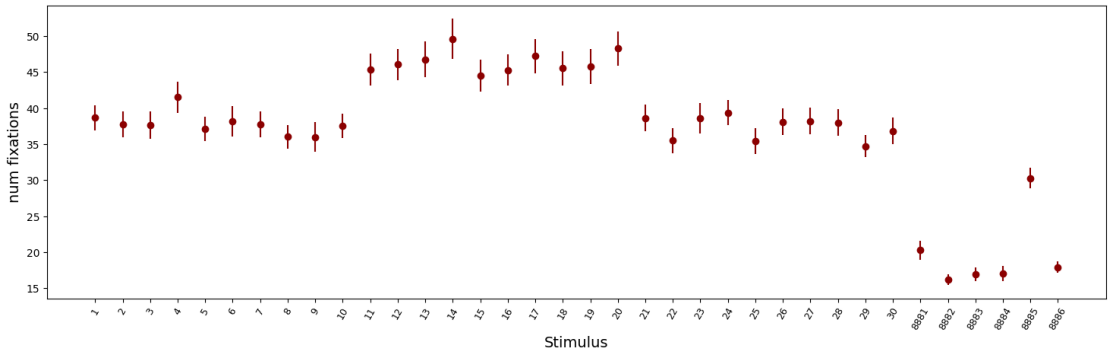


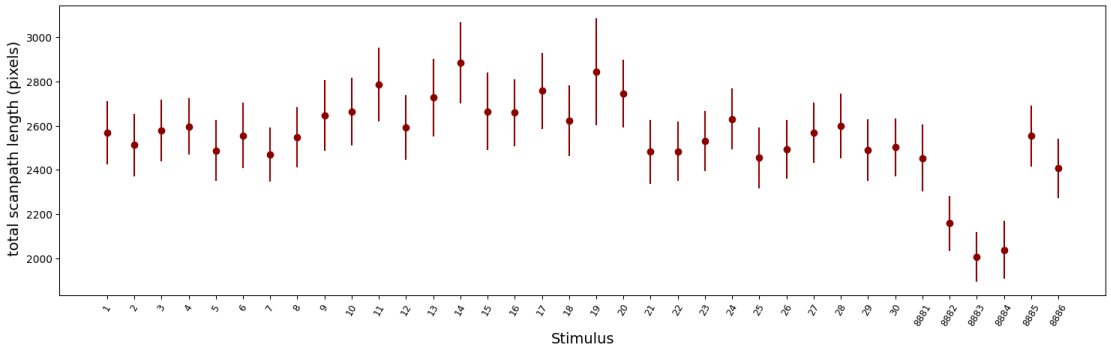
Figure 2.4.: Dysfluent Readers of German: Number of subjects by Average Number of Fixations



(a) Fixations Duration, Mean and Standard deviation



(b) Number of fixations, Mean and Standard deviation



(c) Scanpath Length, Mean and Standard deviation

Figure 2.5.: Dysfluent Readers of German: Descriptive Analysis

In Figure 2.6 we compare different metrics between Typical Reader (Group 4), Isolated Spelling Deficit (Group 3), Isolated Reading Deficit (Group 2), Reading and Spelling Deficit (Group 1), No Group (NaN).

As also reported in the paper, we can notice that subjects with Isolated Spelling Deficit do not show great differences compared to typical readers. In almost every stimulus the mean and standard deviation of these two groups are very close to each other and well separated from the other two groups (1 and 2).

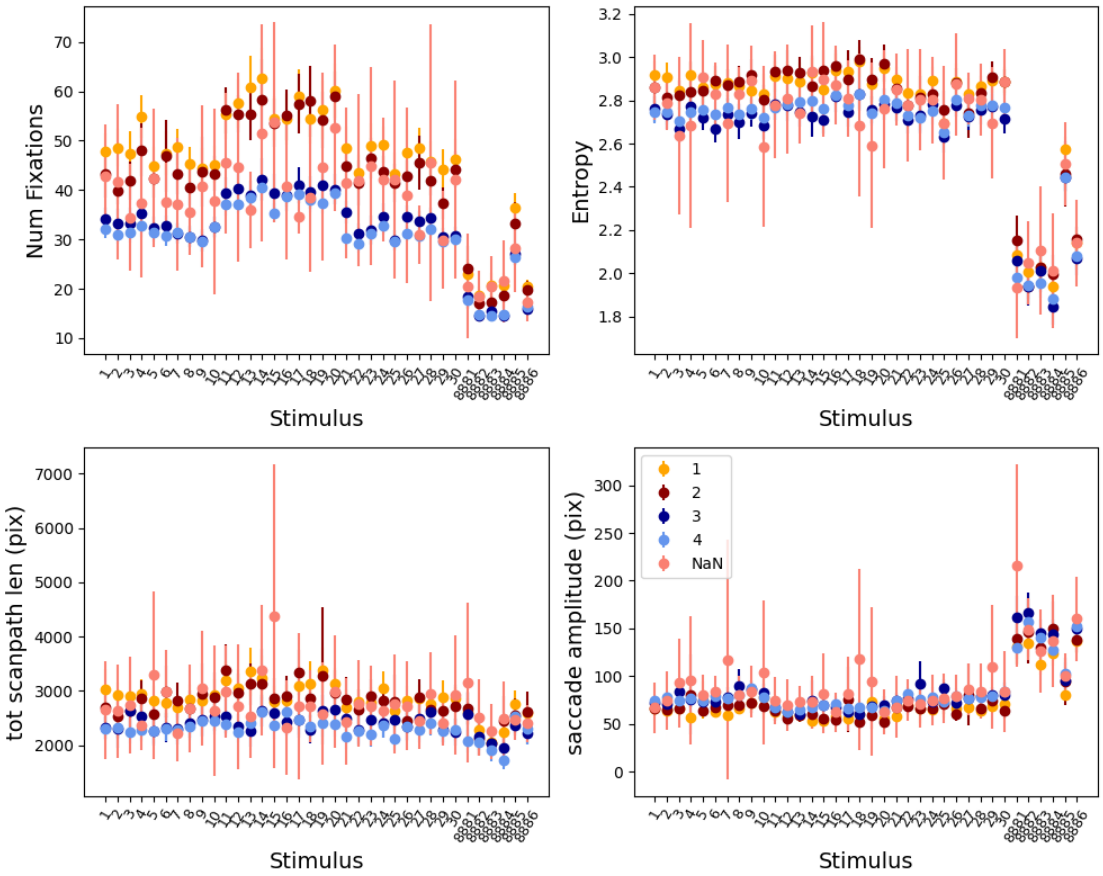


Figure 2.6.: Dysfluent Readers of German: Descriptive Analysis by Groups

In Table 2.2 we can see the statistical t-test on different eye-movement metrics performed between each class. We can easily notice that there are strong differences comparing classes 1 and 2 with classes 3 and 4, whereas class 1 and 2 do not have any metric with a p-value below 0.05, and the same is comparing the classes 3 and 4. Regarding the subjects with no label, we can see how we have differences in the saccade amplitudes with the class 1 and for the number of fixation with the class 4. We can inspect in more detail the number of fixations and we can see that even if they are actually larger in the NaN class, they are still lower than the classes 1 and 2. For this reason we consider more

important the difference in saccadic amplitude. This leads us to divide the entire dataset into two general classes: Class 0 (composed of Groups 3, 4 and the subjects without a group) and Class 1 (composed of Groups 1 and 2). The final objective will therefore be to attempt to classify these two classes correctly.

We will also test the best model without considering these subjects with no class, which are still a minority and should not have a big impact on the final results. **This will be necessary only for the Graz subjects**, since the Munich subjects with no group were in the 29 subject with problem in some tasks (as explained above in subsection 2.2.4). Therefore, if the result obtained is worse it means that, on average, subjects were correctly classified to the class 0. While if we would obtain better results, we can conclude that these subjects are sometimes misclassified and that therefore not including them allows us to better separate the two classes.

The Figure 2.7 shows the analysis of the differences between the classes, as determined by this new subdivision. We can see that, for each stimulus and feature, the mean value of the two groups differs significantly. In particular, the number of fixations is the feature with the clearest separation.

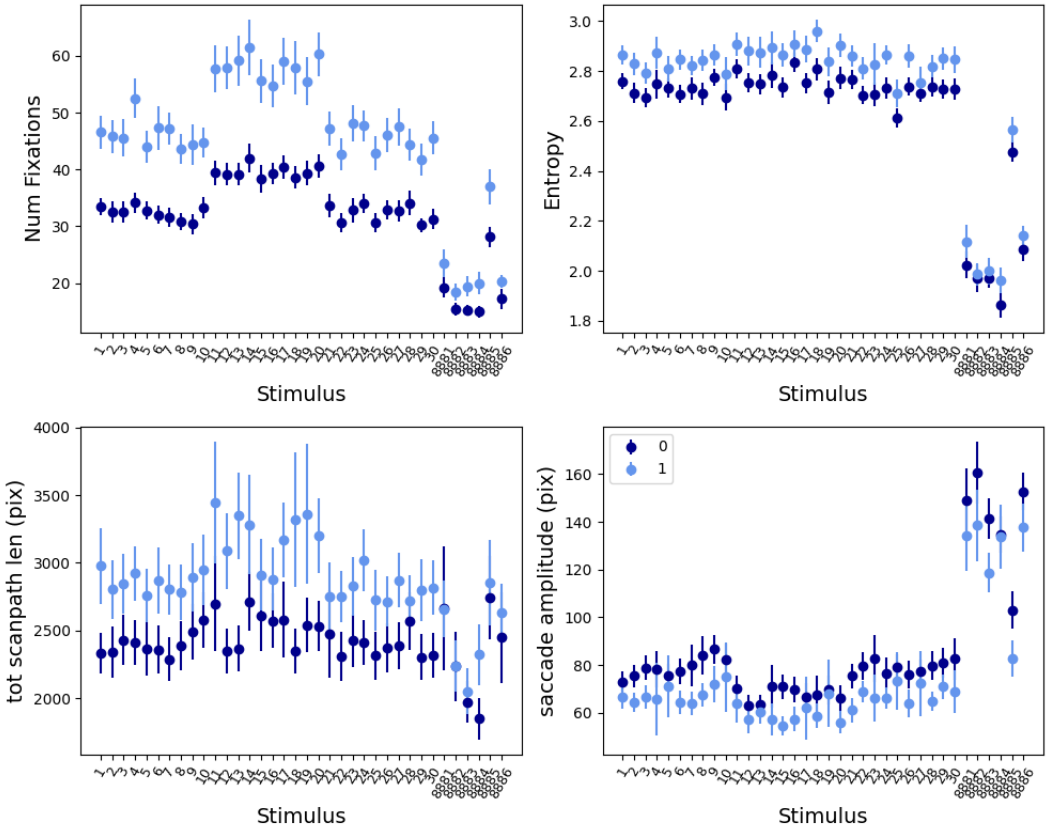


Figure 2.7.: Dysfluent Readers of German: Descriptive Analysis by final group: Control (Group 0) and Deficit (Group 1) subjects

Group A	Group B	Metric	t-value	p-value
1	2	Entropy	0.1324	0.895
		Num Fix	1.3474	0.1817
		Scanpath Lengths	0.4605	0.6464
		Saccade Amplitudes	-0.9003	0.3707
1	3	Entropy	3.6801	0.0004
		Num Fix	8.0472	0.0000
		Scanpath Lengths	3.8636	0.0002
		Saccade Amplitudes	-3.4763	0.0008
1	4	Entropy	3.4717	0.0007
		Num Fix	10.7323	0.0000
		Scanpath Lengths	4.9120	0.0000
		Saccade Amplitudes	-3.2382	0.0016
1	NaN	Entropy	1.1389	0.2595
		Num Fix	1.8211	0.0738
		Scanpath Lengths	0.3884	0.6992
		Saccade Amplitudes	**-2.6738**	**0.0098**
2	3	Entropy	3.1652	0.0023
		Num Fix	7.4740	0.0000
		Scanpath Lengths	2.7249	0.0081
		Saccade Amplitudes	-2.1288	0.0367
2	4	Entropy	2.8253	0.0058
		Num Fix	10.0504	0.0000
		Scanpath Lengths	3.4704	0.0008
		Saccade Amplitudes	-1.8268	0.0709
2	NaN	Entropy	1.1129	0.2736
		Num Fix	1.3531	0.1850
		Scanpath Lengths	0.1065	0.9158
		Saccade Amplitudes	-1.9556	0.0588
3	4	Entropy	-0.4253	0.6715
		Num Fix	1.4783	0.1422
		Scanpath Lengths	0.8367	0.4046
		Saccade Amplitudes	0.4789	0.6329
3	NaN	Entropy	-0.6833	0.4975
		Num Fix	-1.9803	0.0532
		Scanpath Lengths	-1.4748	0.1465
		Saccade Amplitudes	-0.7981	0.4286
4	NaN	Entropy	-0.4667	0.6421
		Num Fix	**-2.9385**	**0.0044**
		Scanpath Lengths	-1.7966	0.0766
		Saccade Amplitudes	-1.0722	0.2872

Table 2.2.: Dysfluent Readers of German - Group Comparisons by t-test on Eye Movement Metrics

3. Methodology

In this chapter, we present the statistical tests and the four classification methods used in the project. We explain in detail how each method works and how it was applied in the categorization process. Before proceeding, it is important to clarify that, since we have datasets with different characteristics, each method required minor adaptations to suit the specific context. Any modifications to the tests or model architectures will be described in the following sections or directly alongside the presentation of results.

3.1. Preliminary Data Analysis

In the dataset presentation we observed some clear differences in the visual paths of the two groups. Therefore, we decided to proceed inspecting in more detail the differences of some specific features like: Saccade Angles, Fixation (total number, duration, and entropy), Progression, Regression, and Total Reading Time.

Firstly, we will analyse the scanpath angles of the two groups by means of two statistical tests: Watsons U^2 test[53] and a MANOVA approach. These tests have been evaluated in previous works and are recommended as optimal for handling circular data especially on large datasets[54], as in our case. All the other features will then be analysed using another statistical test, Hedges' g effect size, as proposed in the paper[13].

3.1.1. Watsons U^2 test

Watsons U^2 test is a non-parametric rank-based test for equality of two circular distributions. Large U^2 values indicate greater discrepancy between the two empirical cumulative distribution functions (ECDFs) on the circle.

Let n_1 and n_2 be the sizes of the two samples, with $N = n_1 + n_2$. Let $\theta_1, \theta_2, \ldots, \theta_N$ be the sorted combined sample (modulo 2π), and let

$$F_1(\theta_i) = \frac{1}{n_1} \sum_{j=1}^{i} \mathbf{1}_{\text{label}_j=0}, \quad F_2(\theta_i) = \frac{1}{n_2} \sum_{j=1}^{i} \mathbf{1}_{\text{label}_j=1} \tag{3.1}$$

be the empirical CDFs of sample 1 and sample 2 evaluated at each sorted data point θ_i. Define the difference at each point:

$$D_i = F_1(\theta_i) - F_2(\theta_i) \tag{3.2}$$

Then, **Watsons U^2 statistic is given by:**

$$U^2 = \frac{n_1 n_2}{N^2}\left(\sum_{i=1}^{N} D_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} D_i\right)^2\right) \tag{3.3}$$

To assess the statistical significance of the observed Watsons U^2 statistic, we employed a test with 1000 random permutations.

3.1.2. MANOVA

Multivariate Analysis of Variance (MANOVA)[1] is a statistical test used to compare the means of two or more groups across multiple dependent variables. To bring circular angles into a linear multivariate framework, we converted each angle θ into its two Cartesian components: $x = \cos(\theta)$, $y = \sin(\theta)$. Then we treated (x,y) as the two dependent variables and the group label (group 0 - Control vs. 1 - Dyslexic) as the single factor.

To better understand the results that will be presented later, let us list and define all the statistic:

Let H be the hypothesis sumsofsquaresandcrossproducts (SSCP) matrix and E the error SSCP matrix obtained in the MANOVA. Let the $s = \min(p, \nu_H)$ nonzero eigenvalues of $E^{-1}H$ be

$$\lambda_1, \lambda_2, \ldots, \lambda_s,$$

and denote by $m = \text{rank}(H)$ and $\nu_E = \text{rank}(E)$ the corresponding degrees of freedom.

Wilks' Lambda Λ

$$\Lambda = \frac{\det(E)}{\det(E+H)} = \prod_{i=1}^{s} \frac{1}{1+\lambda_i}. \tag{3.4}$$

A small value of Λ indicates that the betweengroup effects (captured by H) are large relative to withingroup variation (captured by E).

Pillais Trace V

$$V = \text{tr}\big(H\,(H+E)^{-1}\big) = \sum_{i=1}^{s} \frac{\lambda_i}{1+\lambda_i}. \tag{3.5}$$

Because it accumulates the proportions of explained variance, Pillais criterion is often the most robust to violations of assumptions. With V that ranges from 0 to 1 and larger value indicate stronger effect.

HotellingLawley Trace T

$$T = \text{tr}\big(E^{-1}H\big) = \sum_{i=1}^{s} \lambda_i. \tag{3.6}$$

[1] MANOVA Algorithms: `https://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/24.0/en/client/Manuals/IBM_SPSS_Statistics_Algorithms.pdf`

This is the sum of the eigenvalues, and thus measures total signaltonoise across all discriminant dimensions. For this reason, it is more sensitive to differences in group means. Larger value of T indicate stronger effect.

Roys Greatest Root Θ

$$\Theta = \frac{\lambda_1}{1 + \lambda_1} \tag{3.7}$$

It focuses solely on the largest eigenvalue (λ_1), i.e. the single strongest linear combination differentiating the groups. Also in this case, larger value of Θ means stronger effect.

Each of these four statistics is ultimately transformed into an Fstatistic with appropriate numerator and denominator degrees of freedom (here, ν_H and ν_E) to test:

H_0 : no multivariate effect (all group means equal) versus H_1 : at least one linear combination differs.

A small $p$value (e.g. $p < 0.05$) leads to the rejection of H_0 and conclude there is a significant multivariate effect.

In addition to the test regarding the two groups, we also have the results of the Intercept, which tests whether the mean vector of the dependent variables is significantly different from zero. This step is necessary in order to proceed with the actual test.

3.1.3. Hedges' g Statistic

Hedges' g is a measure of effect size that quantifies the difference between two groups means, similar to Cohen's d. It is especially effective for small sample numbers because it is computed as the difference between two groups means divided by a pooled standard deviation, with a correction factor to reduce bias. This bias correction is typically recommended when sample sizes are smaller than 50. For larger size both statistics produce very similar results[2].

Let $\bar{X}_1$ and $\bar{X}_2$ be the sample means, s_1 and s_2 the sample standard deviations, and n_1, n_2 the sample sizes for the two groups.

Pooled standard deviation:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{3.8}$$

[2]Hedges'g Statistic:`https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/hedgeg.htm`

Cohen's d

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s_p} \tag{3.9}$$

Correction factor

$$J = 1 - \frac{3}{4(n_1 + n_2 - 2) - 1} \tag{3.10}$$

Hedges' g:

$$g = J \cdot d = \left(1 - \frac{3}{4(n_1 + n_2 - 2) - 1}\right) \cdot \frac{\bar{X}_1 - \bar{X}_2}{s_p} \tag{3.11}$$

standard error of Hedges' g:

$$SE_g = \sqrt{\frac{n_1 + n_2}{n_1 n_2} + \frac{g^2}{2(n_1 + n_2 - 2)}} \tag{3.12}$$

Confidence Interval (95%):

$$\mathrm{CI}_{95\%} = g \pm 1.96 \cdot SE_g \tag{3.13}$$

3.2. Similarity Measure

The angle similarity method in `PyEyeSim` is designed to compare scanpaths by analysing the angular relationships between saccades. It provides multiple approaches for quantifying similarity, ensuring robustness in various scenarios.

Since there is no single definition of similarity, the library offers several methods. Some are more general and compare differences in angle distributions overall, while others are better suited to specific cases. In total, there are five main methods:

- **Threshold:** The threshold-based approach determines similarity by checking whether the absolute angular difference between two saccades is below a predefined threshold. Each saccade is transformed to an equivalent representation within the range $[0, 180]$ degrees to ensure consistency. The similarity score is computed as the total number of saccade pairs whose difference falls below the given threshold. This method is particularly suitable when a clear and fixed notion of angular closeness is desired. This allows for a straightforward and interpretable criterion to decide whether two saccades are similar, especially in scenarios where minor directional deviations are acceptable and do not significantly affect the overall pattern recognition.
- **Angle Difference (with peak at $90°$):** The angle difference method quantifies the similarity between two sets of saccades by computing their angular differences. Two saccades are considered maximally different when they form an angle of $90°$. This method is useful for detecting perpendicular movements in scanpaths, which may indicate structured exploration patterns. However, it does not consider differences in the direction of the saccade; for example, $30°$ and $210°$ will be considered equivalent.

- **Angle Difference (with peak at 180°):** This approach is similar to the previous one but considers two saccades maximally different when they form an angle of 180°, indicating movement in completely opposite directions. By emphasizing dissimilarities at 180°, this method is effective for detecting reversal patterns in scanpaths, often associated with systematic searching behaviours.

- **Angle Difference (with peak at 180° and sequence match):** In this case, we optimize the correspondence of one set of saccades to the other maintaining the temporal order.

  The similarity is not computed between all saccade pairs but instead compares the two sequences in order chaining the starting saccade to maximize the match. In the end, the score is penalized based on the difference in sequence length. This method emphasizes the temporal structure of scanpaths, making it particularly suitable for tasks where the sequence of fixations carries meaningful information.

- **Cosine Similarity:** This method represents the distribution of saccade directions as normalized histograms (i.e., probability distributions) over a fixed number of angular bins. This method captures global differences in the overall directionality distribution of the scanpaths. It is especially effective for comparing patterns in tasks where the general spread or bias in movement direction is of interest.

In this project we considered only three of the methods cited above, since they capture differences that are important when comparing scanpaths of reading tasks: Angle Difference with peak at 180°, the version with sequence match, and the Cosine Similarity.

For instance, the Angle Difference with peak at 90° would consider two opposite saccades (i.e. one going from right to left and one going from left to right) to be equal, which obviously would not be true in our case.

Following, we will explain in more detail how these scores are calculated.

Angle Difference (with peak at 180^o) This metric compares all pairs of saccades between two sets by calculating the minimum angular differences, considering circularity (i.e., directions wrap around at 360°). The result is normalized and raised to a power p, which can control the sensitivity to differences:

$$\text{Sim}_{180} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \left(\frac{\min\left(|s_i - t_j|,\, 180 - |s_i - t_j|\right)}{180} \right)^p \tag{3.14}$$

where s_i and t_j are angles from the two sets of saccades, and n, m are the number of saccades in each set.

Angle Difference (180^o with match) This method compares two sequences of saccades by aligning the shorter sequence with a subsequence of the longer one to minimize angular differences. The shorter sequence is slid across the longer one to find the alignment with the lowest average angular difference. Angular differences are

computed circularly (i.e., modulo 360°) and normalized to fall within $[0, 1]$, where 0 represents a perfect match and 1 an opposite direction (180°).

For each alignment, a similarity score is computed as the mean of the powered normalized angular differences. A length mismatch penalty proportional to the difference in sequence lengths is added to discourage large discrepancies.

The final similarity score is the lowest score across all alignments, with an exponent parameter p controlling sensitivity to angular differences:

$$\text{Sim}_{\text{match}} = \min_{\text{shift}} \left(\frac{1}{N} \sum_{i=1}^{N} \left(\frac{\min\left(|s_{1i} - s_{2(i+\text{shift})}|,\ 360 - |s_{1i} - s_{2(i+\text{shift})}|\right)}{180} \right)^p + \frac{|L_2 - L_1|}{L_2} \right) \tag{3.15}$$

where L_1 and L_2 are the lengths of the shorter and longer sequences respectively, and $N = \min(L_1, L_2)$.

Cosine Similarity This method treats each set of saccades as a distribution over angle bins. Each set is first converted into a histogram over the full circular space (e.g., selecting threshold of 10° we obtain 36 bins of 10° each), normalized to form probability distributions. The cosine similarity is computed between the two resulting vectors:

$$\text{Sim}_{\cos} = \frac{A \cdot B}{\|A\|\|B\|} \tag{3.16}$$

where A and B are the normalized histograms of the two saccade angle distributions. This method captures global differences in directional tendencies considering all the angles form 0° to 360°.

For simplicity and to have a more clear comparison of the methods, we will use p=1 for both methods and test two thresholds for the Cosine similarity: 5° and 10°.

Method Comparison

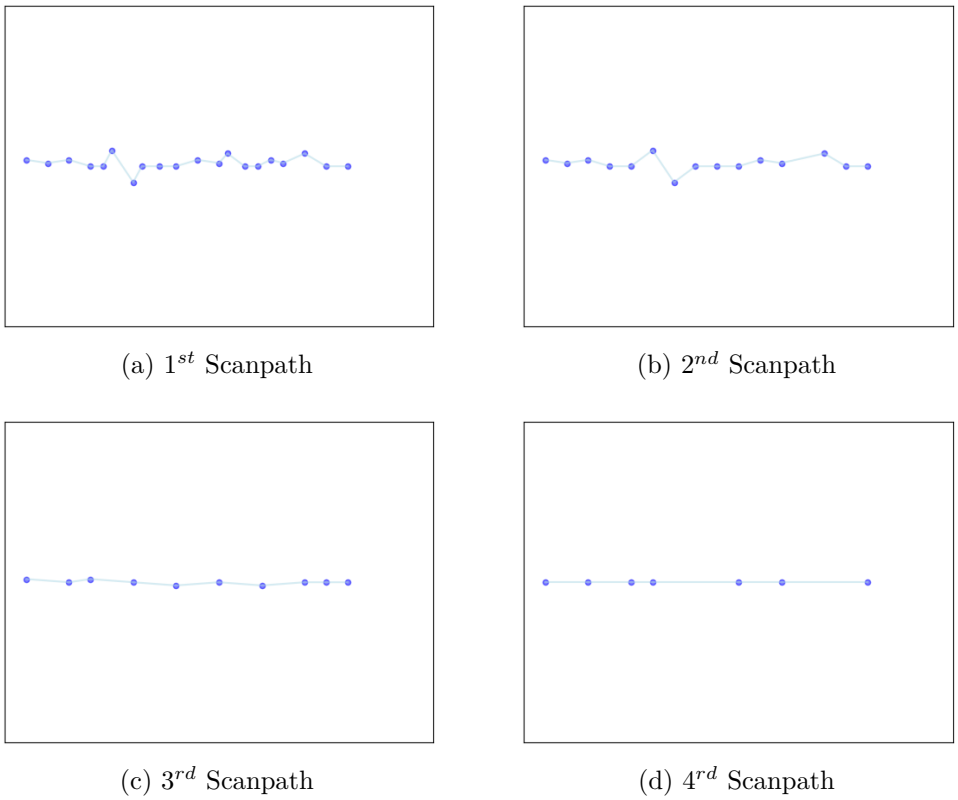
To have a better comparison and understanding of these three methods, we propose a simple example that shows advantages and disadvantages of each score.

We consider four simple scanpaths (Figure 3.1) created manually with the aim of mimicking a reading scanpath, without regression and with all scanpaths starting at x coordinate 5 and ending at the coordinates (80,50). We start with a slow reader that has high fixation number and high variation between the saccade angles (Figure 3.1a). Then we have a completely opposite example, a perfect reader with no variation between saccade angles and few fixation (Figure 3.1d). To complete the comparison we created other two scanpaths, one with characteristics more similar to the first subject (Figure 3.1b) and the other one more similar to the perfect reader (Figure 3.1c). For this reason, we expect that the last two subjects are considered closely similar by our similarity measures

and instead the first two subjects present high dissimilarity compared to the perfect reader.

Below, we provide a more detailed description of the main characteristics of each scanpath:

1. The first one has the highest number of fixations (20) with also the highest mean saccadic angle difference (41.35°)
2. The second presents 15 fixation points with mean saccadic angle difference of 33.31°.
3. The third one presents 10 fixations and 7.27° of mean saccadic angle difference.
4. The last scanpath is a straight line (0° of mean saccadic angle difference) with only 7 fixations.



(a) 1^{st} Scanpath (b) 2^{nd} Scanpath

(c) 3^{rd} Scanpath (d) 4^{rd} Scanpath

Figure 3.1.: Examples of simple scanpaths for testing and comparing similarity scores

We have the first two scanpaths with more fixations and a 30/40 degree high angle variation. The third scanpath have a reduction in fixations and a small difference in angles degrees. In the last one we have only 7 fixations (less than half compared to the first case) and aligned in a perfect straight line.

To give a more precise definition of the desired outcome, we expect the similarity measures to give a much lower score (higher similarity) on scanpaths 3 and 4 than on scanpaths 2 and 4, which should also be lower than the score of scanpaths 1 and 4 (i.e. $Similarity(3,4) \ll Similarity(2,4) < Similarity(1,4)$).

Therefore, in Table 3.1, we can see the comparison of the results obtained from the different similarity measure:

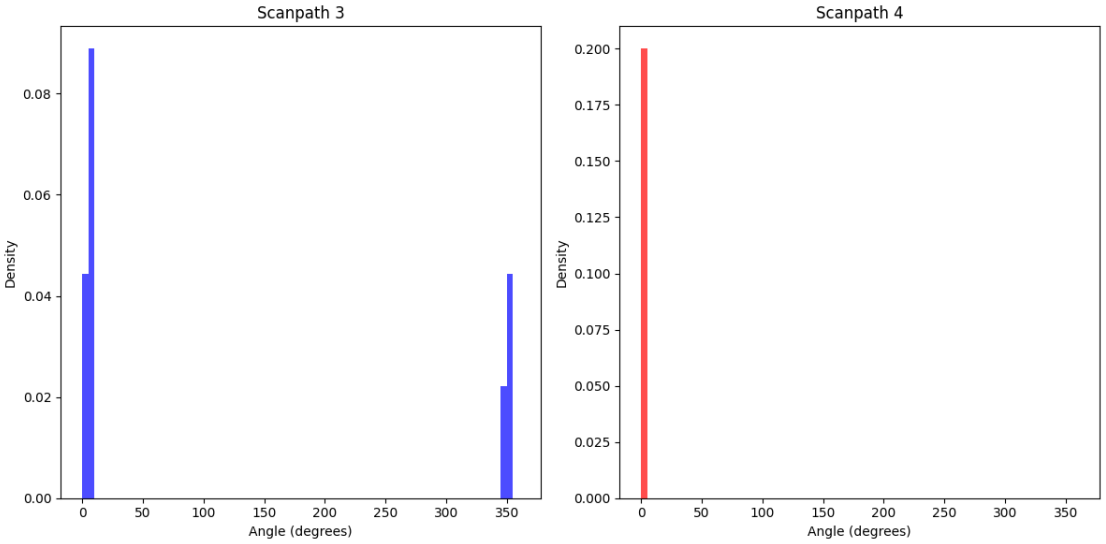
Method	**Similarity(3,4)**	**Similarity(2,4)**	**Similarity(1,4)**
Peak 180	0.0281	0.1136	0.1463
Peak 180 (with match)	0.3545	0.6533	0.7969
Cosine (Threshold 5°)	0.6	0.2546	0.2697
Cosine (Threshold 10°)	0.063	0.2929	0.2857

Table 3.1.: Comparing similarity scores using example scanpaths

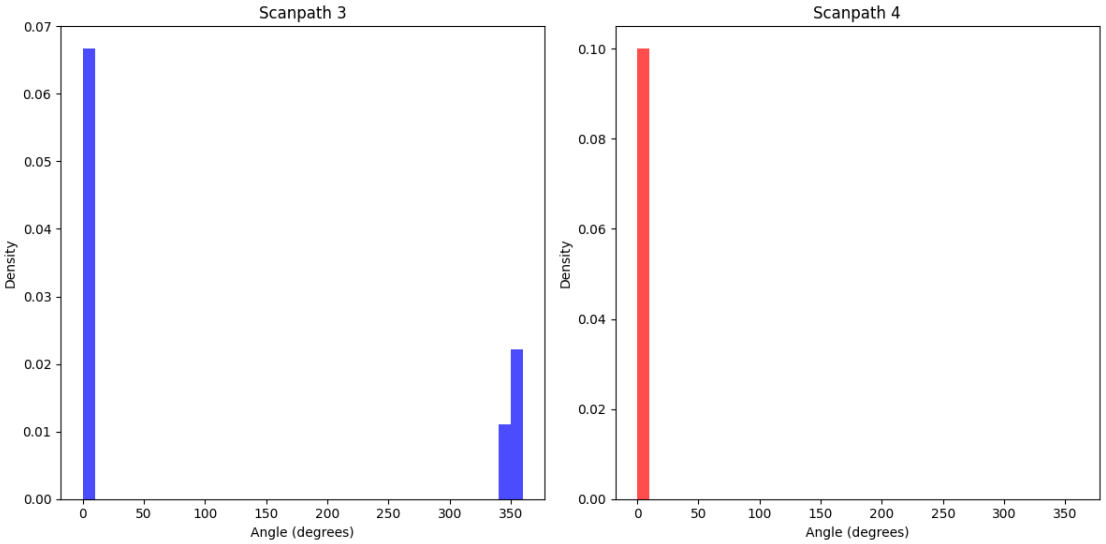
- **Peak 180**: In this case, we can see that we achieve the desired result, even though all the similarity score are very low considering that the range it is form 0 to 1.
- **Peak 180 (with match):** With this method we can observe that other than obtaining the desired result, we have a better subdivision of the score obtained.
- **Cosine (Threshold 5°):** In this case we did not obtain the expected result. The scanpath 3 and 4 results to be the most dissimilar since with bins of 5° we have only few angles that are in the same bins for the two histogram (as we can see better in Figure 3.2). Moreover, the difference between scanpath 1 and 4 and 2 and 4 is very small (as with peak 180).
- **Cosine (Threshold 10°):** By changing the threshold we can completely change the results obtained previously and now we correctly have that scanpath 3 and 4 have the highest similarity. However, now the similarity between 2 and 4 is higher than 1 and 4. In Figure 3.2 we can visualize the histogram of the scanpath 3 and 4 with the two different angle of threshold. We can clearly see that using bins of 5° the density of angle between 0 and 5 degrees is very low. Instead using bins of 10° we have higher correspondence, having grouped together the angles between 0 and 10 degrees.

In summary, the Peak 180 method correctly compares the proposed scanpaths. However, because the scoring scale ranges from 0 to 1, the differences between multiple scanpaths can be very subtle. A similar issue can be observed with the Cosine method, where the choice of threshold plays a critical role: increasing or decreasing it can completely alter the resulting scores. Moreover, the Cosine method is more general, as it evaluates differences across all angles from 0° to 360°. This can overly penalize scanpaths that are in fact very similar, as shown in the example with a 5 ° difference. By contrast, the Peak 180 method with the matching process, handles angles differently and introduces two

additional factors that are crucial for analysing reading scanpaths: the order and the length of the sequence. As we will see in the classification results, incorporating these aspects in the calculation of the similarity score can bring to notable improvements.



(a) Histogram of the Scanpath 3 and 4 with threshold of 5°



(b) Histogram of the Scanpath 3 and 4 with threshold of 10°

Figure 3.2.: Cosine Similarity - Histogram of the two scanpath (with different threshold)

3.2.1. Classification Process

PyEyeSim allows the computation of a similarity measure for each stimulus and for every pair of subjects. In this way it is possible to select a subset of training/testing subjects and compute the average and standard deviation of the similarity scores for each group.

Once the scores have been computed, we evaluate two different classification approaches:

- **Threshold-based classification:** For each test subject, the similarity score for each group is computed. The subject is then assigned to the group with the higher average similarity score, or it is also possible to define a specific threshold using the mean and standard deviation.
- **Neural network-based classification:** The average similarity score and its standard deviation are used as input features to a multi-layer perceptron (MLP), which is trained to classify the subject into one of the predefined groups.

Problem with the Threshold-based classification

During testing on the ETDD70 Dataset, we noticed that dyslexic subjects show in average a low degree of similarity between their scanpaths. Moreover, this result is even lower to the average similarity between Dyslexic and Control subjects. In Table 3.2, it is possible to observe an example of what we have described. We calculated the similarity score for the entire stimulus using Peak 180 and the result within the Control Group is the lowest one, meanwhile the one within Dyslexic subjects is the highest score (more dissimilar).

Group	**Control**	**Dyslexic**
Control	0.363	0.391
Dyslexic	0.391	0.413

Table 3.2.: ETDD70: Task 1 - Average Similarity Score within and between groups

Therefore, it is not possible to obtain good results in the classification process by just comparing the score of the two groups. The only way to proceed is to apply a threshold in the similarity score for only the Control group and all the subjects with a result above this threshold will be classified as Dyslexic.

From our tests, we have noticed that, even though there have been improvements, this method still contains too many misclassified subjects. To further improve the results, we also introduced the standard deviation of the score, noting that dyslexic subjects generally have a higher variance. At this point, manually setting a threshold becomes very complex, because other than one boundary for the similarity score of the Control group we need to check also the standard deviation of the score for the two groups. For this reason we decided to use these features in a machine learning algorithm and trying to optimize the classification.

Since in our study we will already define the structure for a neural networkbased classification model, we proceed to evaluate the inclusion of similarity features to additional key eye-tracking measures, in order to assess whether they lead to improvements in the classification performance. In section 3.4 we will present the details of this method.

3.3. Gaussian Hidden Markov Model

Hidden Markov Model (HMM) is a stochastic model used to represent systems where an underlying process evolves over time but it is not directly observable. Instead, what we observe are related outputs (or emissions) which provide indirect information about the hidden states [10].

HMMs assumes that the system follows a Markov process, meaning that the future state of the system depends only on its present state, not on the sequence of past states.

In particular, a specific type of HMM with Gaussian emission has been chosen from the library hmmlearn [3], in which each hidden state generates an observable value that follows a Gaussian distribution.

In figure 3.3 it is possible to see the key components of a Hidden Markov Model with Gaussian emissions:

1. **Hidden States** (S_t): Discrete states representing different hidden states of the system. The system exists in one of a finite number of discrete states at any time step t. In eye movement analysis, states represent different fixation points.

2. **Transition Probabilities** (A): The probability of transitioning from one hidden state to another is represented as a matrix $A = [a_{ij}]$, where:

$$a_{ij} = P(S_{t+1} = j \mid S_t = i) \tag{3.17}$$

   and the sum of probabilities for each row of A is 1.

3. **Emission Probabilities (Gaussian Distribution)**: Each hidden state emits an observed value that follows a Gaussian (Normal) distribution:

$$X_t \mid S_t = i \sim \mathcal{N}(\mu_i, \Sigma_i) \tag{3.18}$$

   where μ_i is the mean and Σ_i is the covariance matrix for multidimensional observations.

4. **Initial State Probabilities** (π): This defines the probability of starting in each hidden state.

[3] hmmlearn - Unsupervised learning and inference of Hidden Markov Models:`https://pypi.org/project/hmmlearn/`

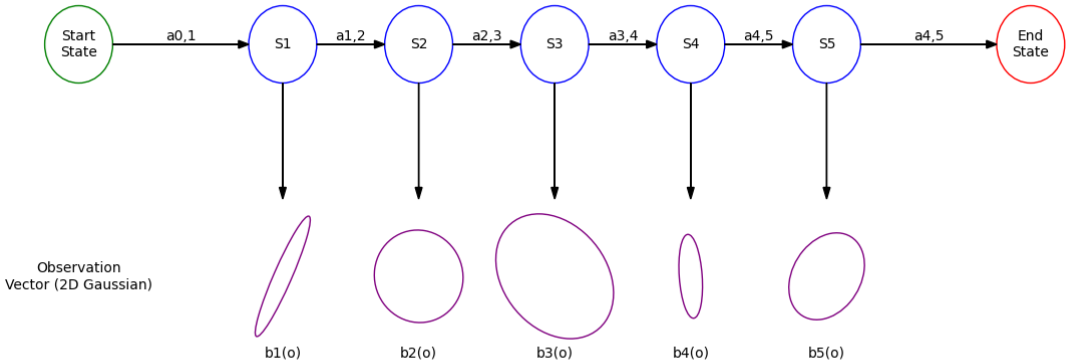


Figure 3.3.: Gaussian Hidden Markov Model (GHMM) Schema

Key advantages of HMMs, particularly for modeling eye movement data, include:

- Efficient handling of sequential data.
- Capability to model temporal dependencies in time series data.
- Capture of probabilistic state transitions and observation uncertainties.

While Hidden Markov Models (HMMs) offer interesting insights in this domain, there are several disadvantages that make them unsuitable for certain scenarios. After conducting various tests on the ETDD70 Dataset, we identified problems regarding the possibility of fitting models with such a high number of components, which can lead to convergence issues.

Convergence problems may occur for two main reasons:

- **Transition Matrix**: The rows of the transition matrix do not sum to 1.0. This issue is often due to numerical approximation errors in the library. However, it also indicates that some hidden states are rarely visited, which can occur when there are very few fixations for one state, preventing the model from properly fitting the components.
- **Covariance Matrix Issues**: When using a full covariance matrix, it must be symmetric and positive-definite. The training process involves Cholesky decomposition of covariance matrices, which requires all eigenvalues to be non-negative. However, small numerical errors can accumulate during computation, causing some eigenvalues to become slightly negative, when they should be close to zero. This results in the model failing to fit properly.

These constraints force us to select fewer components but this leads to larger observations that include multiple words. To better understand this concept, we present two examples in which we attempt to fit a Hidden Markov Model (HMM).

In Figure 3.4a we fitted the model to a high number of components (one for each word) using spherical covariance type (that requires lower amount of data to fit). It is important to note that in this case the fitting process was forced by repeating the test until it encountered no error.

Instead, in the second example, we tried to fit a model with lower number of components using the tied covariance type.

We can observe that in Figure 3.4a there are very large and overlapping observations. Increasing the number of components it does not resolve the problem and also it may cause the model failing to fit.

In Figure 3.4b instead, we obtained better results, but we are constrained to use a lower number of components and this will make the model less precise in identifying data from the same distribution. For this reason, it is difficult to use such models to classify our subjects.

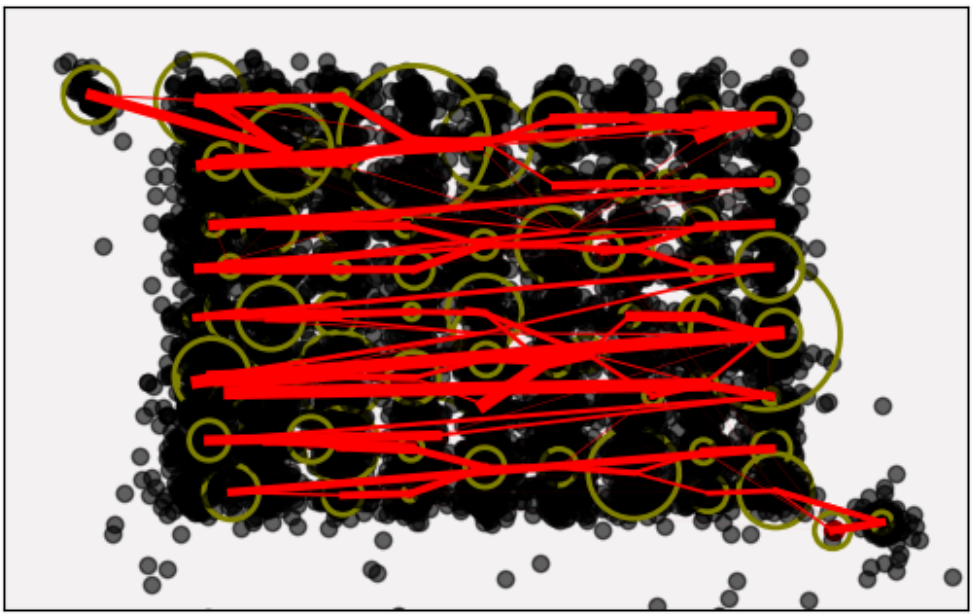
In this example we used the first Task of the ETDD70 Dataset, with low number and well separated syllables. By repeating the same test on the other two tasks, which contain more fixations, we would need to reduce more the number of components. This would make it difficult to obtain results even similar to those shown in Figure 3.4b.

Overall, HMMs work well when applied to stimuli with a limited number of key elements. In such cases, they can provide valuable insights.

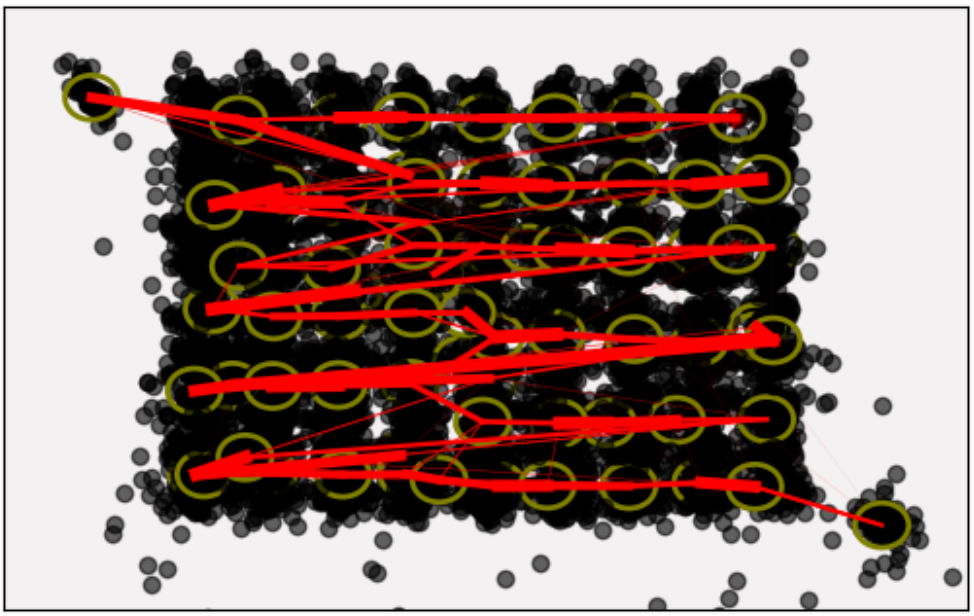
Well-fitted HMMs offer several advantages:

- Allow visualization of observation sequences to identify the most important features/patterns in a stimulus.
- Can generate new scanpaths that mimic real ones.
- Can score scanpaths to determine whether a given path exhibits similar characteristics.

In the context of learning disorder recognition, the previously discussed limitations make HMMs less effective for long textual stimuli, such as those in the ETDD70 Dataset. Instead, they can be used in short sentences like those present in the Dysfluent Readers of German Dataset.



(a) Fitted model with 102 Components and Spherical Covariance



(b) Fitted model with 70 Components and Tied Covariance

Figure 3.4.: ETDD70: Task 1 - GHMM Fitting Tests

3.3.1. Model Fitting

Preliminary tests show that HMMs are subject to a high randomized process and setting a fixed random seed can lead to suboptimal results.

For this reason, in PyEyeSim we created a pipeline that allows to repeat the fitting process multiple times in order to collect more models and keep the best one.

After obtaining the results, the models are ranked based on their Bayesian Information Criterion (BIC) scores, where the lowest value is the best. More details on why we use this criterion will be provided in subsubsection 3.3.4.

This way, we can also evaluate correctly the effect of different number of components and even easily compare differences between models fitted on data of different classes.

In Figure 3.7a we can see an example of results of this pipeline, where for each group and number of components we have minimum, maximum, and average score obtained by the model.

3.3.2. Data Preparation

Lastly, we want to discuss about the changes made to the original data. After some testing, we found it necessary to reduce extreme fixations too far from the text, thus creating a model with more concentrated and more generalizable observations.

The advantage of this model is that it requires only the fixation sequences. However, we noticed that by using the tied covariance type we could find some observations fitted to data very far from the stimulus. This could be a problem for the classification of unseen data and in general for the explainability of the model.

As reported in the work of Hayward J. Godwin, Charlotte E. Lee & Denis Drieghe, "researchers in this field should consider, test, and check the consequences of their cleaning and analytic decisions. Moreover, the reasoning behind such decisions should be made as transparent as possible, highlighting any supporting evidence of prior experimentation or empirical validation of cleaning procedures, where such evidence exists"[55].

Therefore, we will show with some clear practical examples of the problem that we encountered by fitting the model on the original Dysfluent Readers of German Dataset. Then we will explain the solution proposed: how we transformed the data and some results examples.

In Figure 3.5 we can see, for different stimuli, hidden Markov model fitted on the original data. The main problem visible in this figure is that in some cases the model state are positioned outside the image and this is not helpful for the classification process and neither for the understandability of the model.

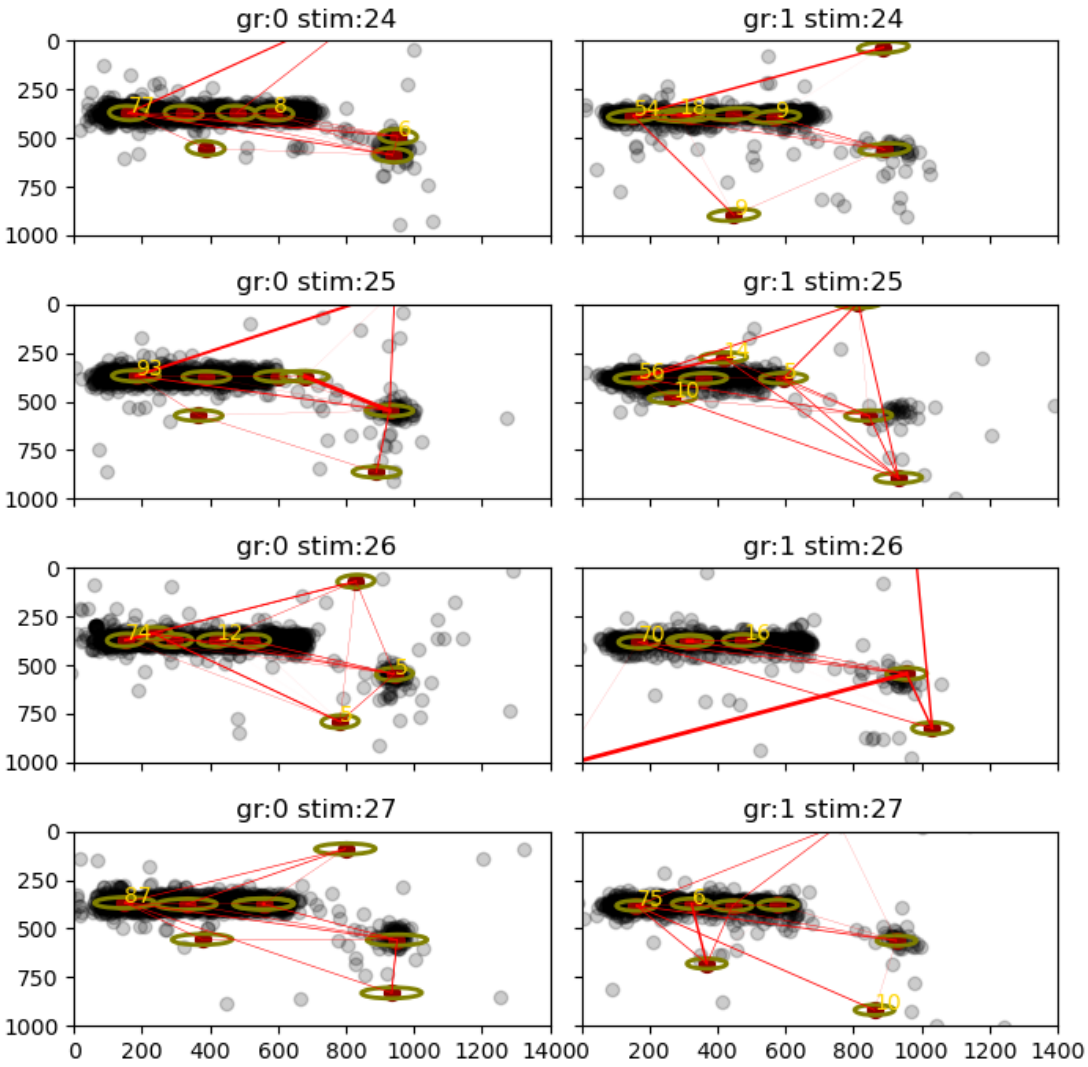


Figure 3.5.: Dysfluent Readers of German - Fitted GHMM using the original dataset

For this reason we used the following algorithm to remove outliers from the dataset using the Z-score method, applied separately for each stimulus.

Algorithm 1 Remove Outliers using Z-score

Require: DataFrame D, x coordinates, y, *stimulus* coordinates, threshold T
Ensure: DataFrame D' with outliers removed

$D' \leftarrow$ empty list
for all unique values s in *stimulus* column **do**
    $G \leftarrow$ all rows in D where $stimulus = s$
    Compute μ_x, σ_x, μ_y, σ_y from G
    for all row r in G **do**
        $z_x \leftarrow \left|\frac{r[x]-\mu_x}{\sigma_x}\right|$
        $z_y \leftarrow \left|\frac{r[y]-\mu_y}{\sigma_y}\right|$
        if $z_x < T$ **and** $z_y < T$ **then**
            Append r to D'
        end if
    end for
end for
return D'

The threshold has been selected in base of the visualization of the results obtained in all the stimuli. The desired result was to remove the furthest fixations while still retaining the last fixation (the one used to terminate the task). In the end, we used a threshold of 4.2 for the Graz subjects and 4.1 for the Munich ones.

Comparing Figure 3.5 with Figure 3.6 we can observe that by cleaning the dataset from extreme fixations, we obtain models with observations that are more in correspondence of the text and no observations outside the screen.

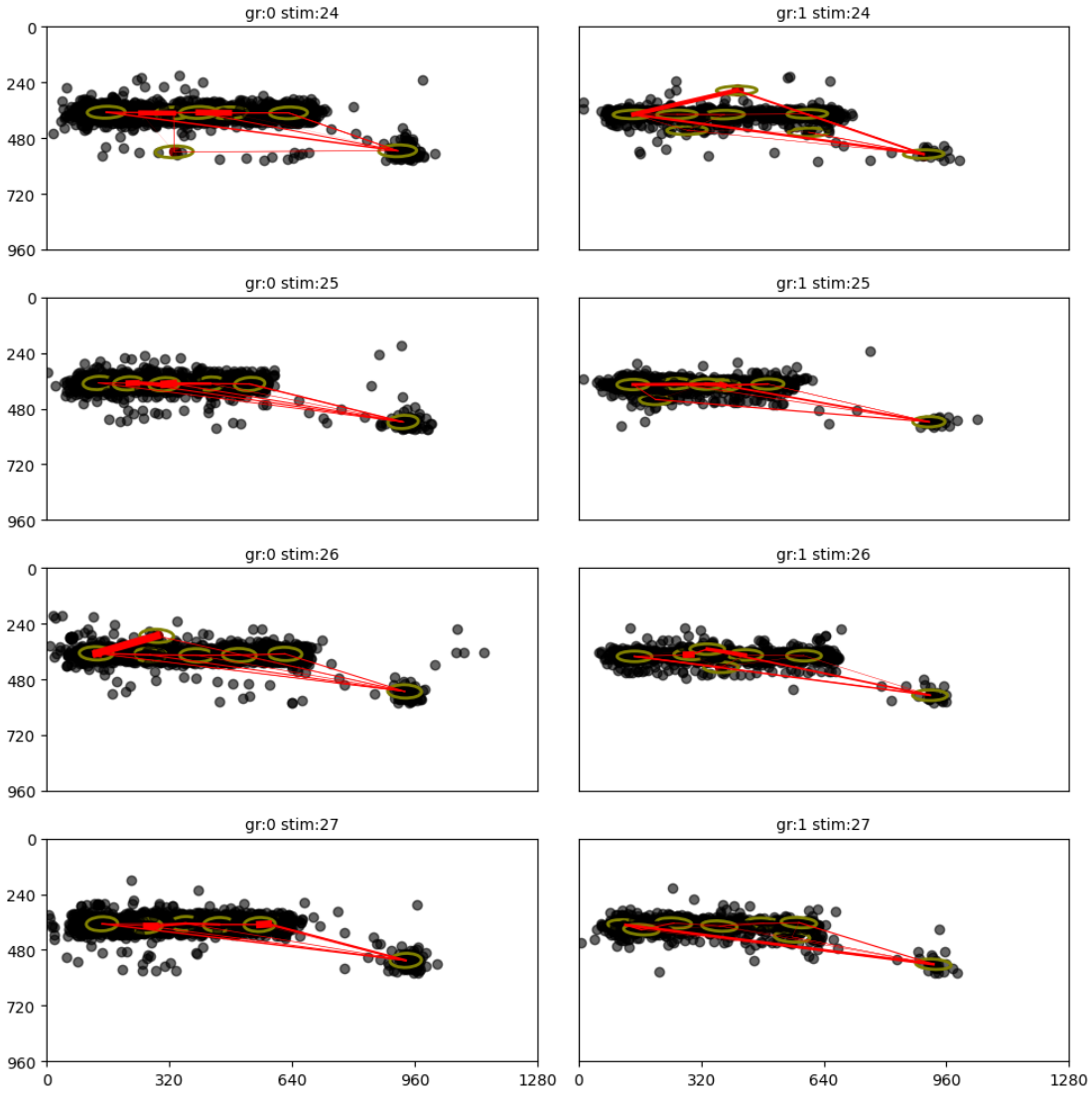


Figure 3.6.: Dysfluent Readers of German - Fitted GHMM using the modified dataset

3.3.3. Parameter Selection

Selecting an appropriate **number of components** and **covariance type** for a Gaussian Hidden Markov Model (GHMM) is a non-trivial task that requires careful evaluation. The number of components directly affects the models capacity to capture the underlying dynamics of the observed eye-movement patterns. Meanwhile, the covariance type determines how flexible the model is in representing the shape and orientation of the state emission distributions, influencing its ability to capture correlations between features and the overall expressiveness of each hidden state.

Number of Components

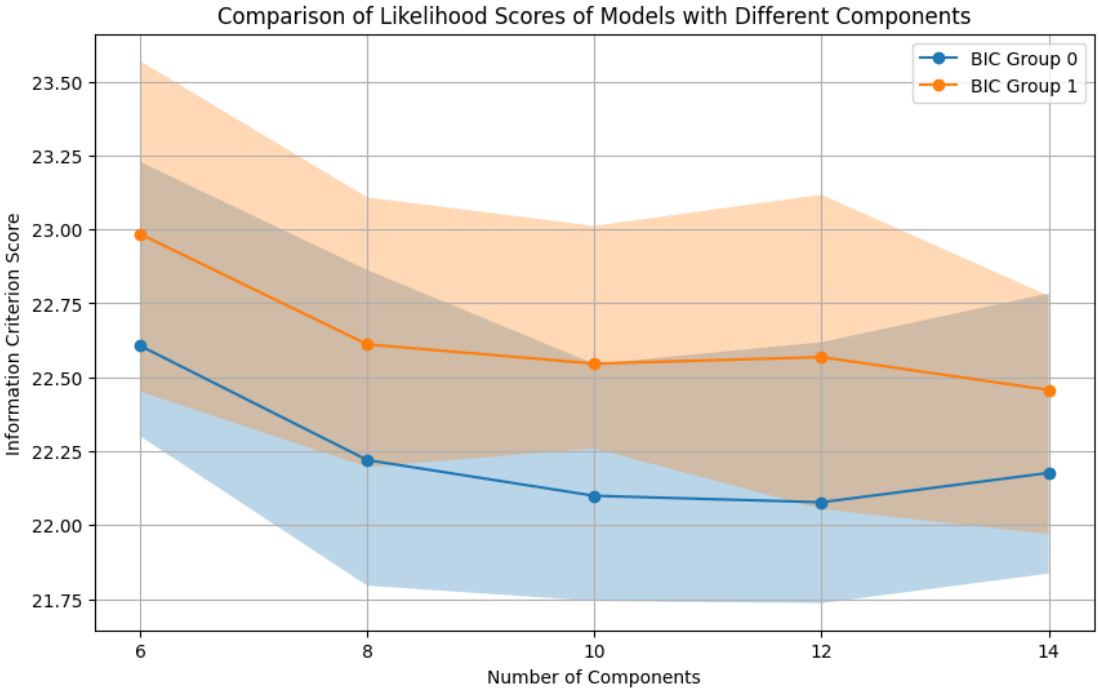
Knowing the dataset characteristics allows us to estimate a reasonable range of component values to test. For example, based on prior knowledge about the stimuli characteristics, we can hypothesize a valid range of valid number of states. However, empirical evaluation across different number of components remains essential to identify the most appropriate configuration.

In PyEyeSim, this process is facilitated through a dedicated pipeline designed for evaluating multiple model configurations. The pipeline systematically fits GHMM with varying numbers of components using the fixation data of all subjects and then compares their performance based on a model selection criterion.

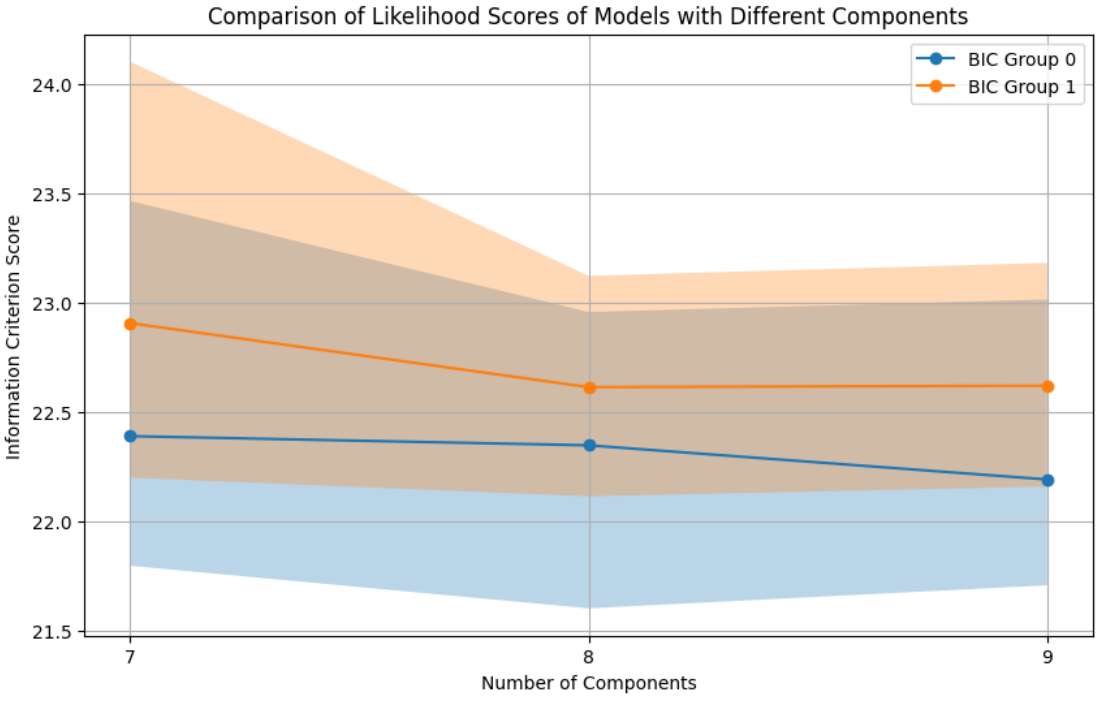
As illustrated in Figure 3.7, the GHMM evaluation pipeline outputs a comparison plot of Bayesian Information Criterion (BIC) scores across different component numbers for the two groups, in this case Group 0 (Control subjects) and Group 1 (Dyslexic subjects). Each line represents the average BIC score obtained over multiple runs, while the shaded regions correspond to the minimum and maximum values observed. This approach helps mitigate the stochastic variability inherent in HMM training, ensuring that model selection is based on stable and representative results rather than on potentially suboptimal single-run outcomes.

These plots allow to compare different models and identify the optimal number of components, corresponding to the minimum BIC value. In some cases, a clear minimum can be observed, indicating an optimal point (or range of values). Otherwise, an "elbow" or "knee of a curve" region may emerge and can be used as a cutoff point. Beyond this point, adding further components results in only small or no improvements in performance.

This comparison also provides insight into how the model complexity interacts with the underlying cognitive profiles of the two groups. For instance, if the Dyslexic group consistently shows higher BIC values or a need for more components to achieve comparable fit quality, this could indicate greater variability or less consistent eye-movement patterns compared to the Control group.



(a) Stimulus 13 - 5 number of components (6,8,10,12,14)



(b) Stimulus 13 - 3 number of component (7,8,9)

Figure 3.7.: GHMM Pipeline results for number of components evaluation

For the tests, we considered 5 different components from 6 to 14, increasing the number by 2 for each component. In Figure 3.7a we can observe a clear jump in the score between 6 and 8 for both groups. Instead, from 8 to 14 there are not significant changes in the BIC score. We analysed in more detail the interval around 8 components and in Figure 3.7b it is possible to observe that 8 and 9 have similar performance while 7 components have a worse BIC score, especially for the group 1.

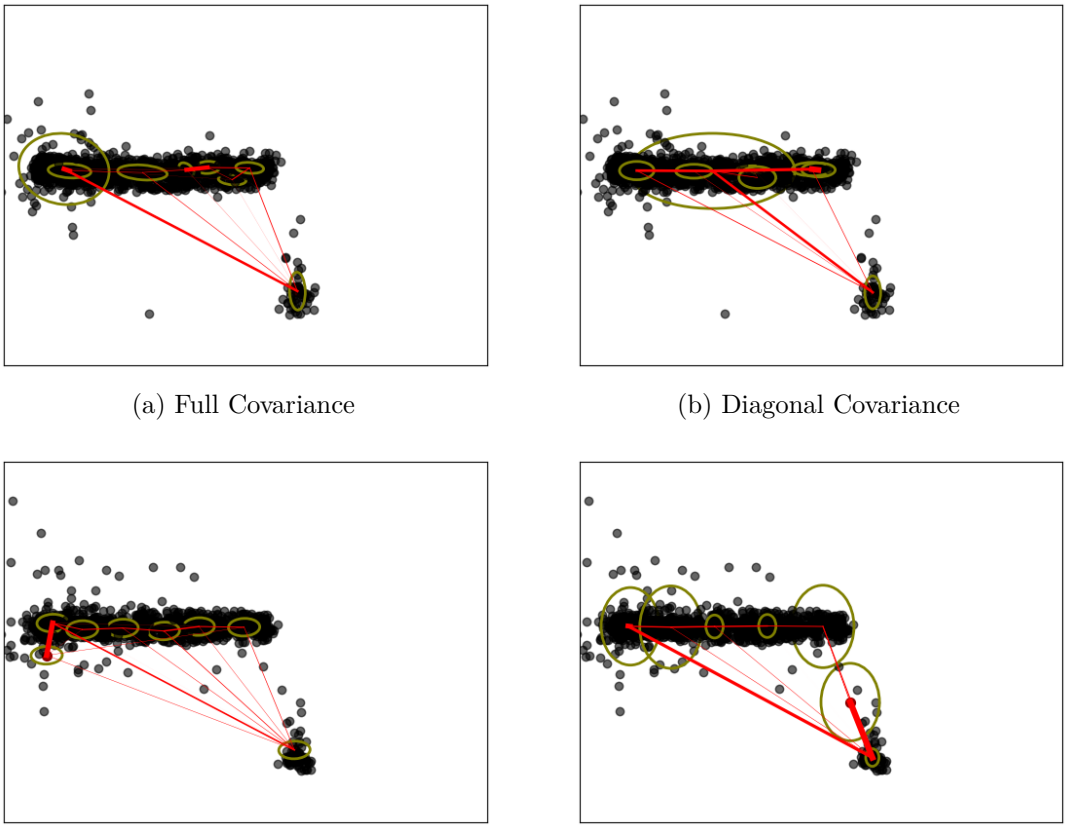
In multiple stimuli of the Dysfluent Readers of German Dataset, we observed that 8 components generally perform well. Using the same fixed number for every test allows us to create a more generalizable model that should work well for each subject and stimulus.

Covariance Type

Since we are working with a low number of components, we can consider all types of covariance: **Full**, **Diagonal**, **Tied**, and **Spherical**.

- **Full**: Each Gaussian component has its own general covariance matrix. This allows for modelling correlations between features, providing the most flexibility. However, it is computationally expensive and may lead to overfitting if data are limited.
- **Diagonal**: The covariance matrix is constrained to be diagonal, assuming that the features are uncorrelated. This significantly reduces the number of parameters to estimate while still allowing for individual variance per feature. It is a good balance between complexity and expressiveness.
- **Tied**: All components share the same general covariance matrix. This assumes that while different components may have different means, the shape and orientation of their distributions are the same. It offers a compromise between the flexibility of the full covariance and the simplicity of the diagonal one, reducing the risk of overfitting and lowering computational cost.
- **Spherical**: The covariance matrix is assumed to be a multiple of the identity matrix, i.e., all features have the same variance and are uncorrelated. This is the most constrained model, requiring the least amount of data to estimate parameters, but it may oversimplify the data structure.

Following, we propose an example showing the results of models fitted on the same data but using different covariance type.



(a) Full Covariance

(b) Diagonal Covariance

(c) Tied Covariance

(d) Spherical Covariance

Figure 3.8.: GHMM Covariance Type Comparison

From Figure 3.8, it is noticeable that the full and diagonal covariance types produce observations that cover a wide range of words and in some cases overlap with other observation states. Spherical covariance on the other hand, is a valid option. In multiple tests, we noticed similar observations to the tied covariance but, as we can see in Figure 3.8d, sometimes the observation can be very large and far from the text. With the tied covariance, we obtain observations that are tightly aligned with individual words. Having all distributions with the same shape and orientation permits obtaining more robust results, without problems of large covariance like in the spherical example.

3.3.4. Score Methods and Classification Approach

HMMs, unlike other machine learning methods, do not have a direct prediction system. The model simply gives a Likelihood score based on how similar the input data are to the patterns learned from the sequence of hidden states during training.

For this reason, we need to build an additional classification system that interprets these scores to determine the most likely class for a given subject.

Scoring Methods

The most commonly used scoring methods, implemented also in the `hmmlearn` library are: Likelihood, Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC).

Likelihood The likelihood score measures how well the model explains the observed data. It is defined as the probability of the observed data given the model parameters.

$$\mathcal{L} = P(\mathbf{O}|\lambda) \tag{3.19}$$

where $\mathbf{O}$ represents the observed data and λ represents the model parameters.

Advantages:

- Direct measure of how well the model fits the data.
- Easy to compute and interpret.

Disadvantages:

- Tends to favor more complex models with more parameters, which can lead to overfitting.
- Does not penalize for the number of parameters in the model.

Bayesian Information Criterion (BIC) A criterion for model selection among a finite set of models. It is based on the likelihood function but includes a penalty for the number of parameters to discourage overfitting.

$$\text{BIC} = -2\ln(\mathcal{L}) + k\ln(n) \tag{3.20}$$

where $\mathcal{L}$ is the likelihood of the model, k is the number of parameters, and n is the number of observations.

Advantages:

- Penalizes the number of parameters, thus discouraging overfitting.
- Useful for comparing models of different complexities.

Disadvantages:

- Can sometimes over-penalize, leading to underfitting.
- Assumes that the model errors are independent and identically distributed, which might not always be the case.

Akaike Information Criterion (AIC) Another criterion for model selection. It balances the goodness of fit of the model and the complexity of the model by including a penalty term for the number of parameters.

$$\text{AIC} = -2\ln(\mathcal{L}) + 2k \tag{3.21}$$

where $\mathcal{L}$ is the likelihood of the model and k is the number of parameters.

Advantages:

- Penalizes model complexity, reducing the risk of overfitting.
- More flexible than BIC in some scenarios, as it does not assume a specific distribution of model errors.

Disadvantages:

- The penalty may not be sufficient in cases with very large datasets, potentially leading to overfitting.
- Does not directly account for the number of observations as BIC does.

Since the number of components is important in our case, we decided to use BIC for select the number of components (as we saw in the previous section) and evaluating the subject data.

Classification process

After establishing the model scoring mechanism, we proceed to explain the subjects classification process.

As mentioned above, this only applies to Dysfluent Readers of German Dataset, as it is the only one with valid stimuli where models can be trained on a limited number of components and thus avoiding problems such as those mentioned above.

Firstly, since the length of each sentence and each word varies between stimuli, we cannot create a single model and train it with all the data. Instead, we must create at least one model for each stimulus. There are two options: we can create a single model, train it with data from a single class, and then define a threshold based on the training data score to decide whether a subject belongs to the same class. Alternatively, we can create two models, each trained with data from a specific class, and classify a subject according to the model that returns the best score (the lowest BIC score).

We tested both methods and the best one proved to be the second. Although it is more computationally demanding (since it requires training two models instead of one), it achieved on average significantly better results than using a single model. The main problem with the first method is that setting an ideal threshold is not an easy task. Even after finding the optimal value for the training data, the results with the test data are not very good.

In addition, we observed that using a model trained solely on data from dyslexic subjects makes it even more difficult to distinguish between the two groups. As noted in the previous section, dyslexic individuals exhibit considerable variability at the level of the visual pathway, which causes the model to assign high scores to a wide range of subjects. On the other hand, using a model trained on data from control subjects, results in many false positives (i.e., subjects who exceed the threshold despite not being dyslexic). However, when comparing the two models, we found that although some subjects receive

high scores from the control-subject model, their scores are still comparatively lower than those obtained from the dyslexic-subject model, making the latter more effective overall.

Another important factor we noticed is that increasing the number of stimuli also increases the accuracy of the voting system of the two models. For this reason, we will perform multiple tests on different stimuli subsets to find which group of stimuli obtains better performance and in which case we obtain the highest accuracy.

In conclusion, we are going to create a set of HMMs (number_of_stimuli x number_of_folds x number_of_groups) and evaluate each test subject for each stimulus considered. The grading system consists of assigning a positive grade to the class corresponding to the model that returns the best BIC score. In the end, the class with the most votes will be considered the test subject's class.

In the case of parity between the two classes, we decided to classify the subject as dyslexic, since in this case it is preferable to have more false positives than false negatives.

3.3.5. Summary Guidelines

In this section we presented advantages, disadvantages, and challenges of using Gaussian Hidden Markov Model with fixation data. The model does not require particular calculations or feature extraction and it offers clear visualization of emission and transition probabilities, which can help to understand differences in visual scanpaths. However, it is necessary to carefully study and test the input data and the parameters used for training the model.

Another important issue is dealing with high number of components. With the ETDD70 Dataset, it was not possible to fit sufficient number of components to cover the stimulus text.

Using the Dysfluent Readers of German Dataset, we have tested different model parameters:

- The **tied** covariance type is optimal for obtaining observations in correspondence with the words, reducing overlapping and large emissions.
- Using **8 components** resulted in the lowest number of components with the best BIC score and the most generalizable between the two groups and each stimulus.

For the classification process, we will define a majority voting system where a model will be trained for each stimulus and for each group. Finally, the new subject will be classified based on the class for which it received the most votes for the various stimuli.

3.4. Multi Layer Perceptron (MLP)

We decided to structure the MLP in the same way as presented by Sedmidubsky et al. [11] in their work on the ETDD70 Dataset. The only difference is that we will try to keep the structure as similar as possible between the different stimuli in order to check the actual performance with the same features and input values used. The objective is to create a model with standard features that perform stably on different stimuli.

In this section, we will look at the architecture of the neural network, the features used and other important details.

3.4.1. Features Extraction

We extract the same features for different ROI in order to obtain the input value for the MLP.

For Task 2 and 3, the image has been subdivided in 10x10 ROI and for Task 1 we subdivided in 9x10 ROI (9 vertical and 10 horizontal), since in the stimulus we have exactly 9 columns and 10 rows. Form these ROI, we have extracted 4 features with the similarity measure (comparing only the saccades inside the ROI). In addition, we have 10 global features and 4 features based on the similarity measure, this time calculated using the entire set of saccade. This way we obtain a maximum of 414 values.
Regarding the Dysfluent readers of German Dataset, since the length of the stimuli precludes the use of subdivisions in ROIs, only global features will be used. The input of the MLP corresponds to the global features extracted from each stimulus, so we can obtain an input vector of similar size to those used in the previous dataset. The maximum size reachable is 14 features by 30 stimuli that corresponds to an input of 420. As said in subsection 2.2.4, since we can not remove the Graz subjects, we decided to fill the input corresponding to the problematic stimuli with zeros. This allow us to use all the subjects from Graz and therefore to test the MLP model.

The features are then concatenated in a single vector that will be used as input for the MLP and normalized using a `StandardScaler`[4].

ROI Features

- Number of Fixation
- Mean of Fixation Duration
- Number of Revisits (saccades that start from outside the ROI and ends inside)
- Landing of the first coordinate

[4]Sklearn StandardScaler:`https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html`

Global Features

- Number of fixation
- Average saccadic amplitude
- Standard deviation of the saccadic amplitude
- Average of the fixation duration
- Standard deviation of the fixation duration
- Total scanpath length
- Total reading time
- Entropy of the fixation distribution
- Number of regressions (saccades higher than 90° and lower than 270°)
- Ratio of the number of regressions to the number progressive saccades
- Mean of the group similarity for Dyslexic subjects
- Mean of the group similarity for Non-Dyslexic subjects
- Standard deviation of the group similarity for Dyslexic subjects
- Standard deviation of the group similarity for Non-Dyslexic subjects

In PyEyeSim it is possible to calculate the similarity score also for specific ROI. However, since for the second dataset we can use only global features and we aim to create an architecture that can be generalised over different type of stimuli, we decided to calculate the score using the complete scanpath of the subjects. Therefore, for both datasets, the similarity measure will be used as a global feature.

We define six different setups to study the performance of the selected features and identify the most effective ones. In particular, regarding the novel similarity measure, we aim to test whether, on average across all tests, it provides an improvement in the classification process. All the tests that include the similarity measure are repeated and compared using the four similarity measure presented before.

1. **Original Features:**
   We consider the original features presented in the paper, that consists in the 4 features extracted form the ROI and 6 global feature: **Number of fixations, Average fixation duration, Total reading duration, Average saccadic amplitude, Ratio of progressive/regressive saccades, Number of regressions**; The first subset is necessary in particular for the German Dataset where we do not have any results to compare. Although, since we made changes to the architecture, it will also be useful for the ETDD70 Dataset.

2. **Original and Additional Feature:**
   We consider an additional set of features that include the standard deviation of saccadic amplitude and fixation duration, and also the entropy of the fixation distribution. The fixation entropy is integrated in multiple functions inside PyEyeSim and tested with others feature in the preliminary analysis. We will test if this feature in combination with the original ones can improve the model accuracy.

3. **Original Features with Average and Standard Deviation of the Similarity Score:**
   For each subject, we calculate the average and standard deviation of the similarity measure for the two groups and use the values as global features of the model.

4. **Original Features with Average and Standard Deviation of the Similarity Score, without Average saccadic amplitude:**
   The Similarity Score already includes information regarding the amplitude of the saccade. We will also see in the preliminary analysis that the mean saccade amplitude does not have clear difference between the two groups, so we decided to include this test in the comparison. This way, we can check if the same pattern is reflected in the classification process.

5. **Original Features with Average of the Similarity Score:**
   We also test the same feature without including the standard deviation and checking whether this improves or worsens the accuracy of the model.

6. **Original Features with Average of the Similarity Score, without Average saccadic amplitude:**
   In the same way, we repeat the test by removing the average saccadic amplitude.

7. **All feature:**
   Finally, we test all the features that we have extracted (average and standard deviation). Also in this, case we repeated the test for each similarity measure.

3.4.2. Model Architecture

- **Input layer:** every layer applies an affine linear transformation to the input data

$$y = xA^T + b$$

- **Hidden layer:** for every hidden layer we divide the number of input value by 2 until we reach a number of input of 2 digits (<100).

- **Drop Out:** for every hidden layer we applied a 20% dropout.

- **Output layer:** from the previous hidden layer we have 2 values, in input and by using Softmax of dimension 1, we generate the output of the model.

Although the code used in the paper is available, we preferred to rewrite the structure in order to adapt feature extraction with the functions integrated directly into PyEyeSim. The network was implemented using PyTorch, employing Xavier uniform initialization for all layers to ensure a stable starting point for training. Xavier initialization is designed to keep the scale of the gradients roughly the same across all layers, helping to prevent issues such as vanishing or exploding gradients during backpropagation. For optimization, we used Stochastic Gradient Descent (SGD) with a learning rate of 0.2 (based on the best results in the paper) and Cross Entropy as the loss function, which is well-suited for multi-class classification tasks.

Additionally, we explored the effect of different momentum values within the SGD optimizer. In particular we noticed that for obtaining the same results reported in the paper, we needed to apply a momentum of 0.8. For this reason we have tested also a lower momentum of 0.5 to control if we can get further improvements. Momentum helps accelerate convergence and avoid local minima by incorporating information from previous updates. In some cases, tuning the momentum parameter led to noticeable improvements in accuracy. A detailed analysis of the impact of momentum on model performance will be presented in chapter 4.

The training process involved 5-fold stratified cross-validation to maintain class distribution, 20 epochs per fold and Checkpoint saving for the best model based on loss. In the first dataset, to keep the same approach reported in the paper, we used sklearn library `StratifiedKFold`[5] to get the same dataset splitting (55 subject for training and 15 for validation). Instead, for the German Dataset, we applied a the sklearn `train_test_split`[6] with stratify for balancing the class with 90% for training and 10% for testing.

3.5. Convolutional Neural Network (CNN)

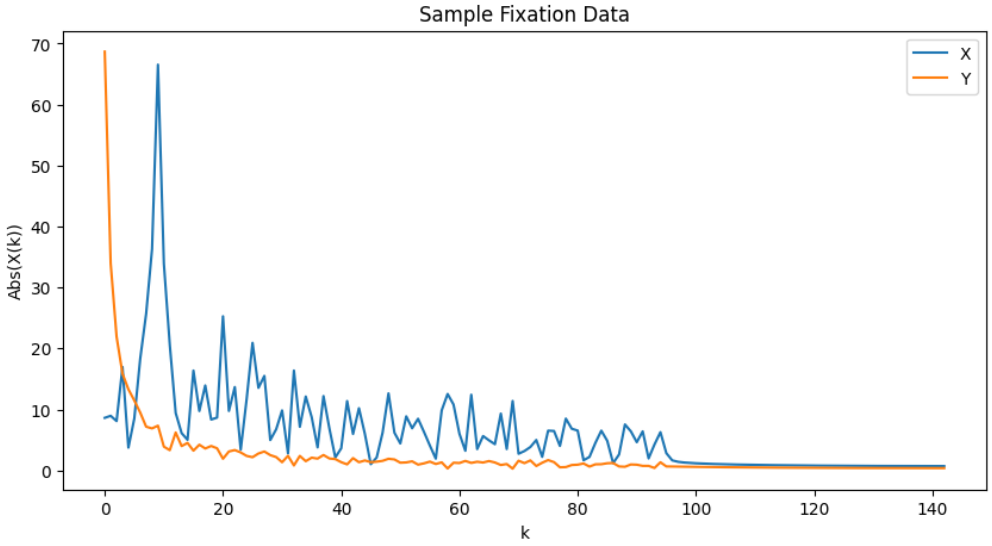
For the CNN approach, we implemented a methodology inspired by the paper of Boris Nerui, Jaroslav Polec, Juraj Skunda, and Juraj Kaur[8]. The process involves transforming fixation data into a continuous signal through several steps:

- **Signal Preparation**:
  - Extract x and y coordinates of fixations for each participant.
  - Calculate pad length by checking the slowest reading subject.
- **Signal Processing Pipeline**:
  1. Apply Discrete Cosine Transform (DCT) Type III for signal interpolation.
  2. Extend spectrum with zeros for the calculated pad-size.

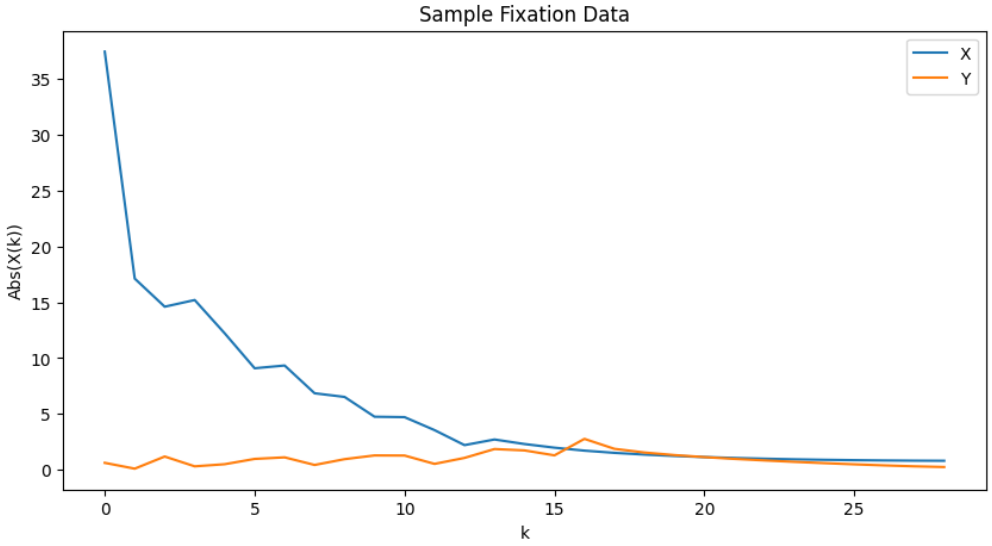
[5]Sklearn StratifiedKFold:`https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html`

[6]Sklearn train_test_split:`https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html`

3. Perform Inverse DCT to reconstruct continuous signals.
4. Apply energy correction based on original sequence length.
5. Compute Discrete Fourier Transform (DFT).
6. Calculate magnitude spectrum.



(a) ETDD70 CNN Input - random subject from Task 1: Magnitude spectrum of time domain interpolated signal (x and y coordinates)



(b) Dysfluent Readers of German - random subject from random stimulus: Magnitude spectrum of time domain interpolated signal (x and y coordinates)

Figure 3.9.: Magnitude spectrum of time domain interpolated signal (x and y coordinates)

3.5.1. Model Architecture

As described in the paper, we explore three distinct convolutional neural network (CNN) architectures.

CNN - 2 Layer

- Input layer: 2 channels (x and y coordinates)
- Two convolutional blocks in sequence, each block consists of:
  - 1D convolution with output channels 8 and 16 respectively
  - Kernel size 3, stride 1, and padding 1
  - Batch Normalization
  - ReLU activation
  - MaxPooling with kernel size 2 and stride 2
- Flattening of features after last block
- Two fully connected layers:
  - First fully connected layer with ReLU activation
  - Second fully connected output layer producing raw logits
- Output layer corresponds to number of classes; use with CrossEntropyLoss

CNN - 3 Layer

- Input layer: 2 channels (x and y coordinates)
- Three convolutional blocks in sequence, each block consists of:
  - 1D convolution with output channels 8, 16, and 32 respectively
  - Kernel size 3, stride 1, and padding 1
  - Batch Normalization
  - ReLU activation
  - MaxPooling with kernel size 2 and stride 2
- Flattening of features after last block
- Two fully connected layers:
  - First fully connected layer with ReLU activation
  - Second fully connected output layer producing raw logits
- Output layer corresponds to number of classes; use with CrossEntropyLoss

CNN - 4 Layer

- Input layer: 2 channels (x and y coordinates)
- Four convolutional blocks in sequence, each block consists of:
  - 1D convolution with output channels 8, 16, 32, and 64 respectively
  - Kernel size 3, stride 1, and padding 1
  - Batch Normalization
  - ReLU activation
  - MaxPooling with kernel size 2 and stride 2
- Flattening of features after last block
- Two fully connected layers:
  - First fully connected layer with ReLU activation
  - Second fully connected output layer producing raw logits
- Output layer corresponds to number of classes; use with CrossEntropyLoss

The models were trained using the following configuration of hyperparameters and evaluation strategies:

- **Loss function:** Binary Cross-Entropy Loss was used to optimize the models for the binary classification task.
- **Optimizer:** SGD optimizer tested with different learning rate (0.01 and 0.1) and different momentum (1 and 0.5).
- **Cross-validation:** To ensure robustness and generalizability, we applied also in this case 5-fold cross-validation.
- **Training duration:** Each fold was trained for 50 epochs to allow sufficient learning while minimizing overfitting.
- **Batch size:** Experiments were conducted with different batch sizes (8, 16, 32 and 64 depending to the dataset) to analyse the impact on convergence and generalization.

4. Analysis and Results

In this section, we will look at the results of the analysis performed on the two datasets and the results obtained through the classification models.

Firstly, as mentioned above, we will carry out various statistical tests to check for differences between the two groups under examination. In particular, we will try to understand whether, as proven in many previous studies[34, 12, 32, 52, 13], differences between the scanpaths of dyslexic subjects and controls can also be observed in our two datasets.

Subsequently, we will present the results obtained by the classification process, using the methods described in the previous chapter. The aim here is not only to evaluate their performance on each dataset individually, but also to compare the results in order to identify which approach proves to be the most accurate and generalizable across the different experimental settings. In doing so, we will provide a clearer picture of the strengths and limitations of the various models and assess their potential for reliably distinguishing between the two subject groups.

4.1. Results of the Statistical Tests

We first provide a visual representation of the saccade angle distributions for each dataset, comparing Dyslexic and Control subjects. These plots are then complemented by statistical tests performed on the distributions to assess whether significant differences exist between the two groups under investigation.

Then, we proceed by presenting and discussing the results obtained by the Hedges g size effect, tested on the metrics that then we will also use as input of our neural network.

4.1.1. ETDD70 Dataset

Comparing saccadic angle distribution between Dyslexia and Control

All the saccade angle plots are represented with a histogram chart in polar coordinates (with a bin size of 5°) showing three charts: two for the separated classes and one with the overlay of both. Then, using bin size of 1°, we focus on the most important angle ranges, where the majority of the saccades are located:

- Progressive Saccade: from 0° to 45°and 315° to 360°.
- Regressive Saccade: from 135° to 225°.

For the MANOVA test we will report and discuss the most important information. A table with all the results can be found in section A.2 of the appendix.

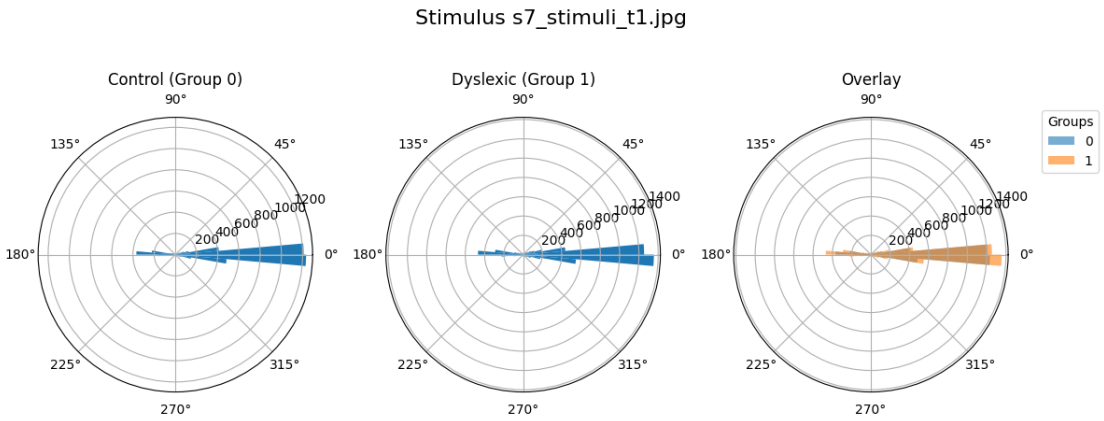
Starting from **Task 1**, the Control group exhibited a total of 5375 saccades, with a mean amplitude of 1.72° and a standard deviation of 63.84°. In contrast, the Dyslexic group showed a higher number of saccades (6776), with a mean of 1.50° and a standard deviation of 71.74°. This difference in both quantity and dispersion already suggests the presence of distinction between the two groups, as also seen in the dataset presentation. To further explore these differences, we proceed with a more detailed analysis of their angular distributions.

In Figure 4.1, we can observe that the overall distributions of the two groups show subtle but notable differences. Specifically, the Control group presents a more balanced distribution of saccades between 355° and 5°, whereas the Dyslexic group displays a higher concentration around 360°. This effect is more evident in Figure 4.1b. Furthermore, as shown in Figure 4.1c, Dyslexic subjects exhibit a greater number of regressive saccades within the range of 170°180°, while the Control subjects show fewer saccades in that region, with their distribution appearing more compact and concentrated.

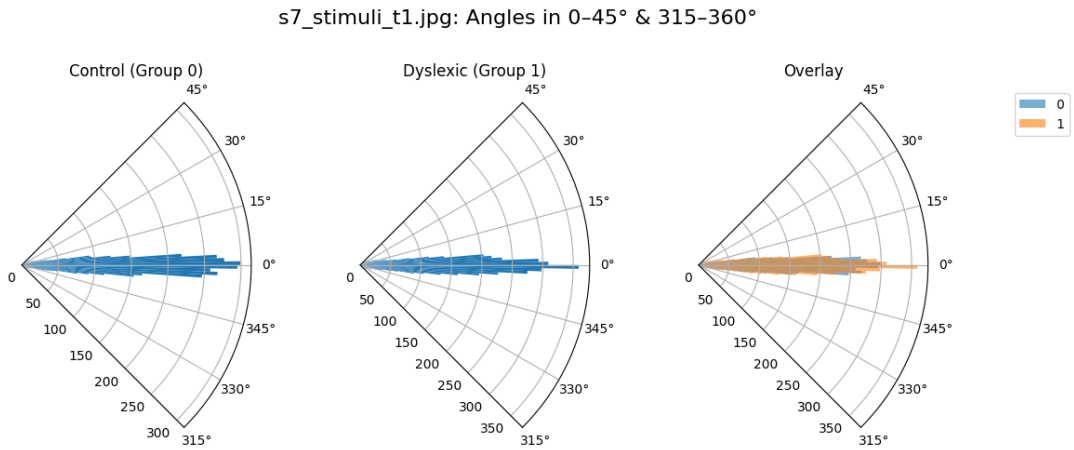
By analysing the statistical differences in the angular distribution we found out that:

- The **Watsons U^2 test** yielded a statistic of $U^2 = 2.1996$ with a p-value < 0.001, indicating that we can reject the null hypothesis of no difference between the two circular distributions.
- Consistently, the **MANOVA test** produced analogous results. After accounting for the intercept, the variable Label has a statistically significant effect on the joint distribution of the dependent variables. While statistically significant, the effect size is relatively small. All four multivariate tests agree on the significance
  - Wilks Lambda showed a small deviation ($\Lambda = 0.9975$), while Pillais Trace, HotellingLawley Trace, and Roys Greatest Root all exhibited same values (0.0022), reflecting small but statistically significant differences.
  - The F-value was 15.3789 with a p-value < 0.0001, confirming a statistically significant overall difference in the angular distributions between the Control and Dyslexic groups.

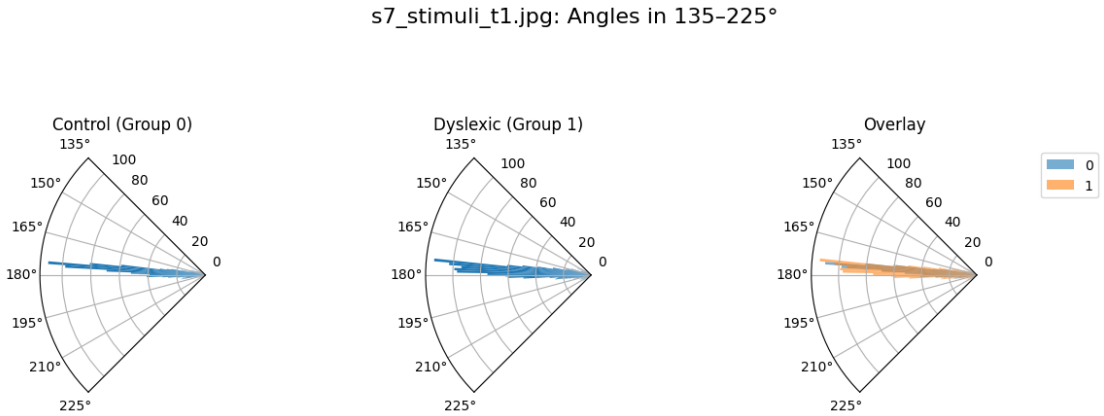
In summary, both tests demonstrate agreement in revealing statistically significant but subtle differences between the two distributions.



(a) Complete Angles Distribution (bins of 5°)



(b) Progressive Angles Distribution (bins of 1°)



(c) Regressive Angles Distribution (bins of 1°)

Figure 4.1.: ETDD70: Task 1 - Comparison of the Saccade Angles Distribution for Control and Dyslexic subjects, considering the first stimulus with ninety syllables

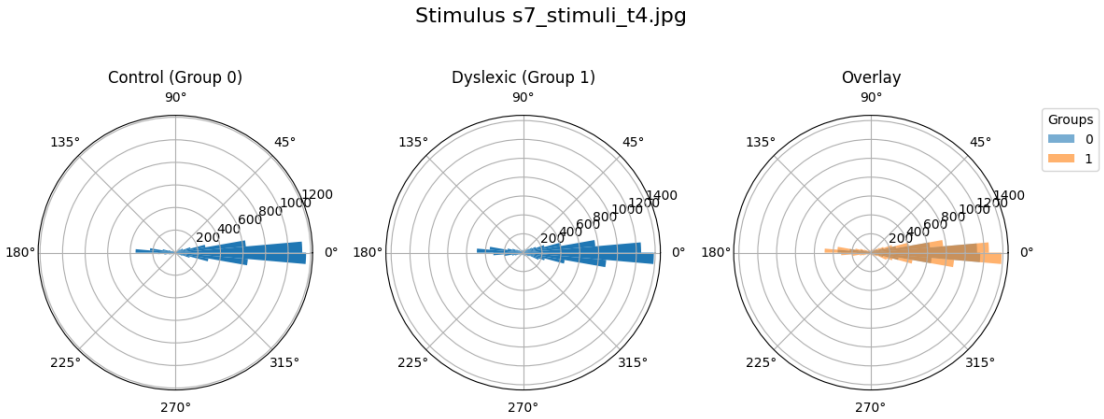
For Task 2, the differences between the two groups become even more pronounced. We observe a further increase in the number of saccades, with 6259 recorded for the Control group and 9632 for the Dyslexic group. Regarding the mean and standard deviation, we obtained results similar to those seen previously: 1.42° and 62.84° for Controls, and 1.03° and 74.22° for Dyslexic subjects. We recall that the two tasks have very different characteristics. In the previous case, we had single words and pseudo-words spaced apart, while in this task we have complete sentences. Therefore, it is normal to see significant differences, including an increase in saccades.

By analysing the images, we can see patterns similar to those seen in the first task. In Figure 4.2b and Figure 4.2c we can notice in the Overlay plot how the angle distribution of the Dyslexic group is more wider and with multiple bins that reach high number of frequency. In particular, in the Regressive distribution it is possible to notice a high number of saccade angles present only in the Dyslexic subjects.

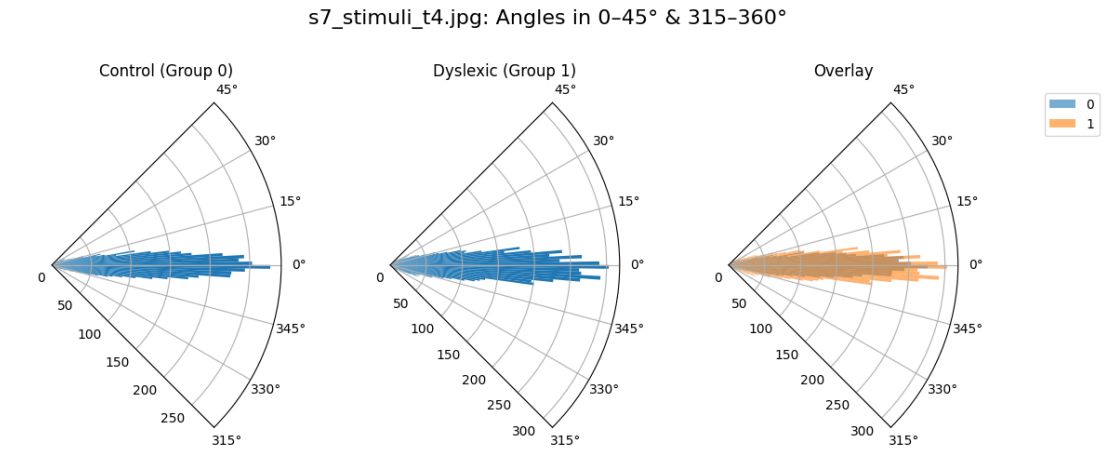
Next, we proceed to present and discuss the results of the two statistical tests:

- **Watson's Uš test** reported a value of $U^2 = 6.5281$ with p-value < 0.001. Compared to the previous task, we have an increase in the statistical value, which means that the difference between the two distributions is more significant.
- The **MANOVA test** also shows an increase in the significance of the deviation between the two distributions, with a p-value still below 0.0001.
  - The statistic value of the tests are: 0.9951 for the Wilks Lambda, 0.0049 for Pillai's Trace, and the other two test equal to 0.0050.
  - The F-Value reflect the increasing of statistic significance, increasing the resulting value to 39.3709.

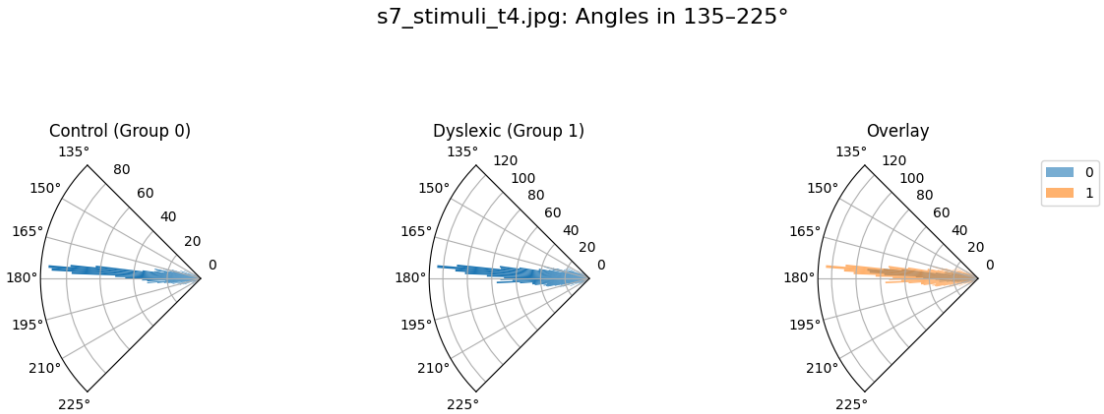
These findings show that Task 2 reveals even greater statistically significant differences between Dyslexic and Control subjects compared to Task 1.



(a) Complete Angles Distribution (bins of 5°)



(b) Progressive Angles Distribution (bins of 1°)



(c) Regressive Angles Distribution (bins of 1°)

Figure 4.2.: ETDD70: Task 2 - Comparison of the Saccade Angles Distribution for Control and Dyslexic subjects, considering the second stimulus with Meaningful Text

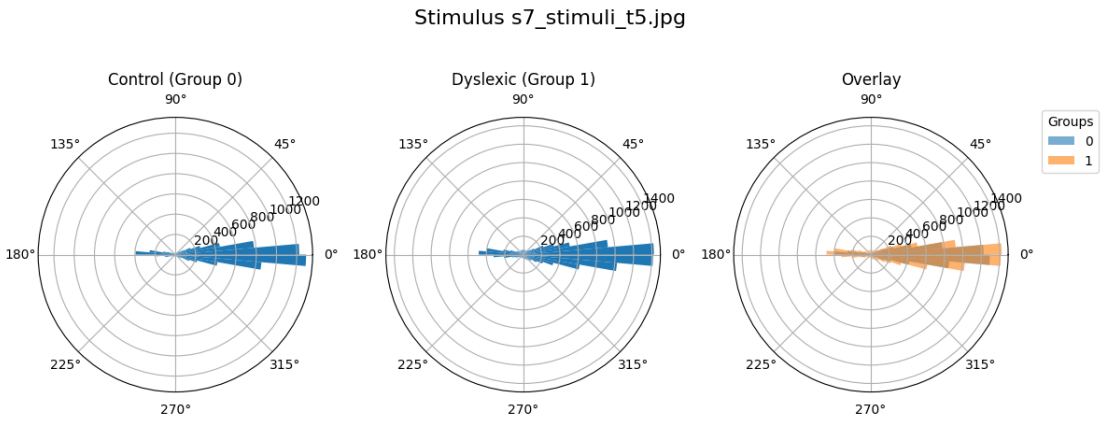
For **Task 3**, we obtained values that are consistent with those found in Task 2, as expected from the similar characteristics of the two stimuli. In both groups, we observed an increase in the number of saccades: 8,308 for the Control group and 12,082 for the Dyslexic group. The mean saccade amplitude was 1.24° and standard deviation of62.12° for the Control subjects, and 1.00° with 73.74° for the Dyslexic subjects. These values are very close to those reported above.

In Figure 4.3 we can see that the same patterns described for the previous stimulus are still present in this one. The Dyslexic subjects have a high and wide concentration of saccades between 355° and 5°, better visible in Figure 4.3b. In Figure 4.3c, we can notice that Dyslexic subjects present again a high variation of Regressive saccades, with even a higher frequency compared to the Control group.

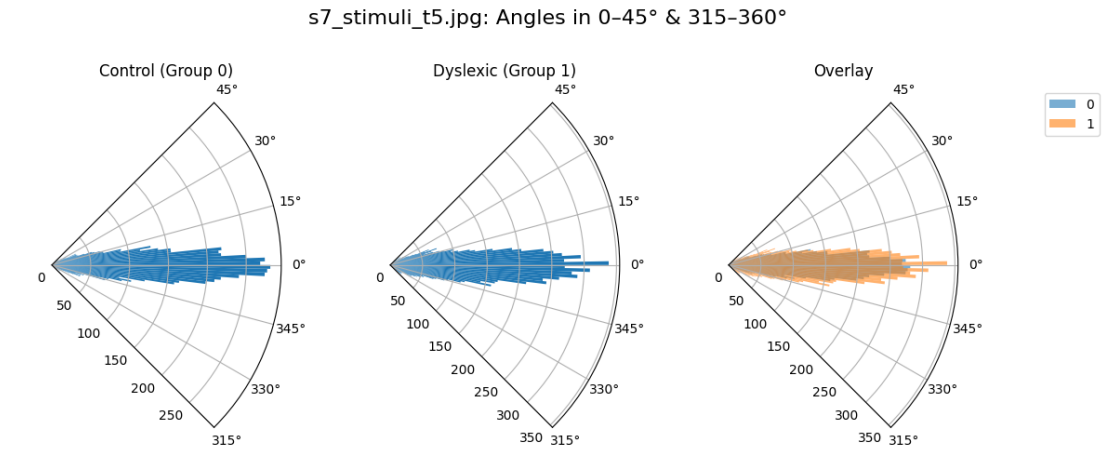
The two statistical tests show again a slight increase in the significance and validity of the difference between the two distributions:

- Compared to the previous task, **Watson's Uš test** has a similar U^2 of 6.8471, with p-value < 0.001.
- The **MANOVA test** shows an increase in the significance, always with a p-value < 0.0001.
  - Wilks' lambda = 0.9945, Pillai's trace = 0.0055 and the other two test equal to 0.0056.
  - The F-Value slightly increased to 56.6821.

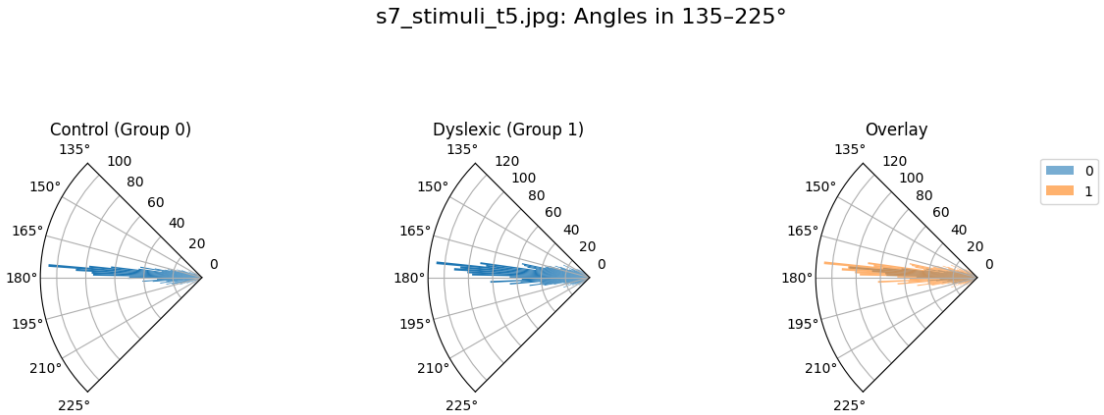
In conclusion, across all three tasks, statistical tests revealed small but significant differences in saccade distributions between Control and Dyslexic subjects. These findings are consistent with previous analyses conducted on this database. This further support the importance of investigating in more detail potential scanpath differences between these two groups.



(a) Complete Angles Distribution (bins of 5°)



(b) Progressive Angles Distribution (bins of 1°)



(c) Regressive Angles Distribution (bins of 1°)

Figure 4.3.: ETDD70: Task 3 - Comparison of the Saccade Angles Distribution for Control and Dyslexic subjects, considering the second stimulus with Non-Words

Comparing metrics statistic between Dyslexia and Control

To analyse the features, which we will take into consideration for the classification process, we show for each stimulus the results of Hedges' g effect size (where effect sizes were considered significant if the 95% CI did not include zero), as presented in the paper[13].

From the images (Figure 4.4, Figure 4.5, and Figure 4.6) we can observe, in concordance with what was reported in the paper cited above, that almost all the features have significant effect size and in general with the Dyslexic subjects showing lower values than the Control group.

Fixationbased metrics

- **Number of fixations:** As seen in chapter 2, Dyslexic readers make substantially more fixations than controls in all three tasks, with all confidence intervals (CI) excluding the zero, indicating highly reliable group differences.
- **Total reading duration and Fixation duration:** Dyslexic participants read more slowly, with longer mean duration for each fixation
- **Fixation entropy:** Entropy of fixation locations is higher in dyslexia, suggesting more dispersed gaze patterns.

Regression metrics

- **Number of regressions:** Compared to the Control subjects, Dyslexic subjects have a higher number of saccades above 90° and below 270° .
- **Progression/regression ratio:** Control readers have higher forwardtobackward saccade ratios, indicating more efficient forward reading saccades.

Saccadeamplitude metrics

- **Mean saccade amplitude:** For Task 1 and Task 3 (Figure 4.4 and Figure 4.6) the CI for mean amplitude crosses zero, indicating no reliable difference in that condition.
- **Saccade amplitude (std):** Variability in saccade sizes is greater in control readers, with positive effect sizes, reflecting more diverse saccade angles between subjects.

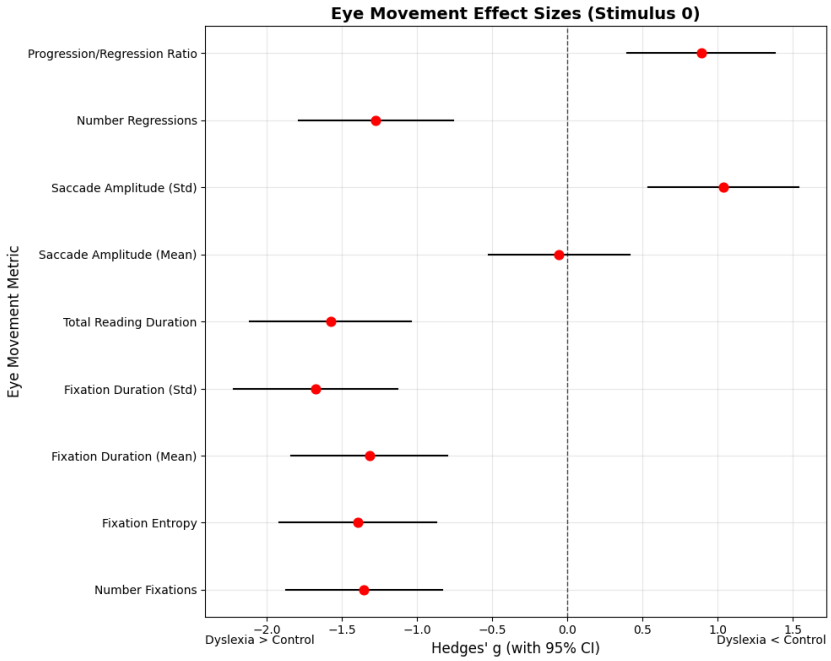


Figure 4.4.: ETDD70 - Task 1 Hedges' g effect size

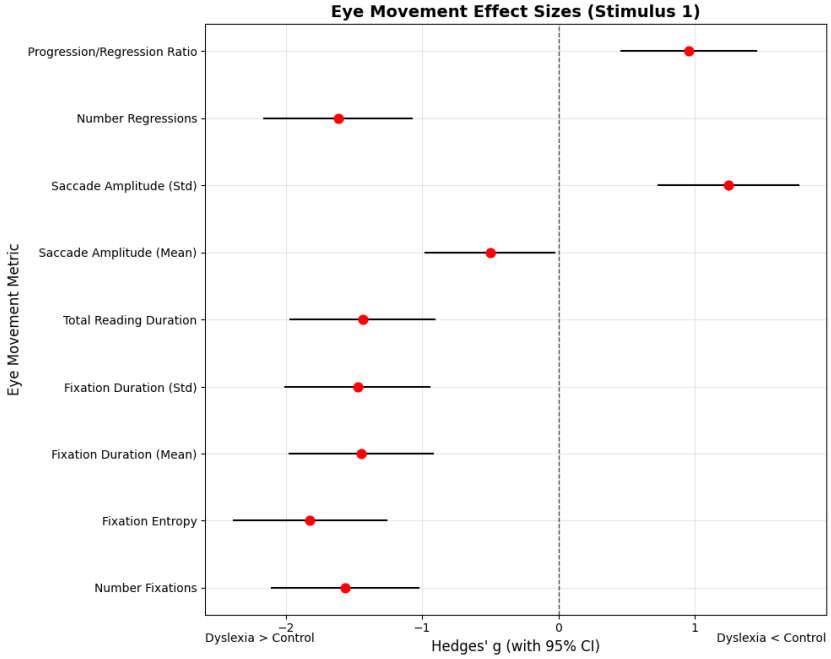


Figure 4.5.: ETDD70 - Task 2 Hedges' g effect size

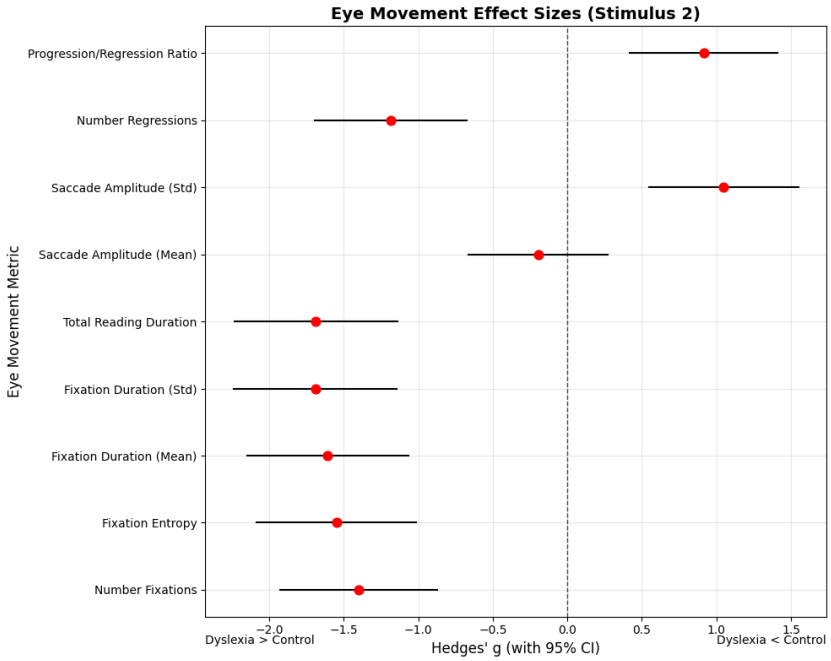


Figure 4.6.: ETDD70 - Task 3 Hedges' g effect size

4.1.2. Dysfluent Readers of German Dataset

Comparing saccadic angle distribution between Dyslexia and Control

Compared to ETDD70, this dataset has a higher number of stimuli. For representing the results in a more concise way, we will show the distributions by considering 5 subgroups:

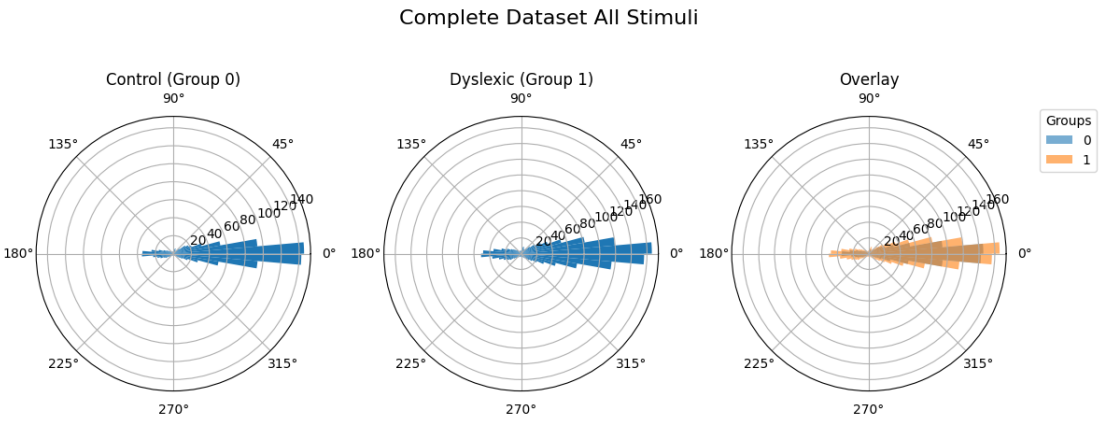
- All 36 stimuli, also considering the 6 practising tasks with only real words.
- The 30 stimuli containing pseudowords and nonwords (without the 6 tasks with real words).
- The 6 practising tasks containing only real words.
- The stimuli containing pseudowords and nonwords, but not considering the stimuli containing only nonwords (from stimulus 1 to 10 and from 21 to 30).
- The stimuli containing only nonwords (from stimulus 11 to 20)

Since we also have an unbalanced group size, to obtain a better visualization, the number of bins is normalized by the number of subjects in each class. This way, we can have a more direct comparison to what we observed in the previous dataset, in particular regarding the frequency of specific bins range. Then, similar to what we did for the previous dataset, we will report and discuss the results of the two statistical tests. For the MANOVA tests, the complete result tables can be found in the section B.2 of the appendix.

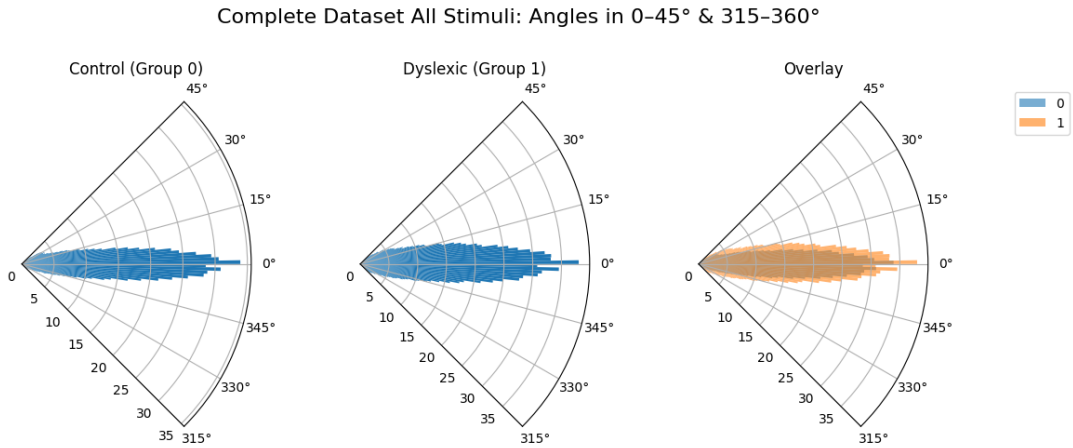
We start by considering all the stimuli. In this case, the Control group present a total of 130856 saccades, that in average are 1200 for each subject, Mean equal to 2.66° and 70.92° of standard deviation. For the Dyslexic group we have 127007, in average 1568 for each subject, Mean of 0.98° and standard deviation of 77.50°. We can notice that, also for this dataset, the Control subjects present a lower standard deviation and in average a lower number of saccades.

In Figure 4.7we can visualize the distribution of the two groups. The saccade frequency of Dyslexic subjects is notably higher in the ranges from 0° to 10° and from 350° to 360° compared to Control subjects. In each plot, it is evident that the distribution of Control subjects is contained within the broader distribution of Dyslexic subjects. These differences are reinforced by the results obtained by the statistical test. In fact, both Watsons U^2 and MANOVA reject the hypothesis that there is no difference between the two distributions. Following, we report the obtained results:

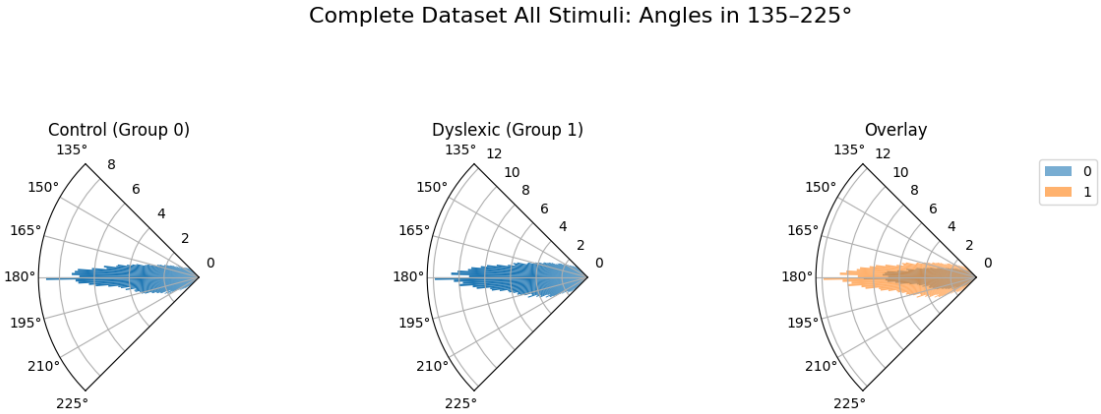
- The **Watsons** U^2 shows a value of 38.7578, with a p-value of 0.0010.
- Every component of the **MANOVA test** have strong evidence against the null hypothesis, even though with a very small effect size.
  - Wilks' Lambda value is equal to 0.9982 and all the other three tests have the same value equal to 0.0018.
  - The F-value is 238.3810 with a p-value < 0.0001.



(a) Complete Angles Distribution (bins of 5°)



(b) Progressive Angles Distribution (bins of 1°)



(c) Regressive Angles Distribution (bins of 1°)

Figure 4.7.: Dysfluent Readers of German - Comparison of the Saccade Angles Distribution for Control and Dyslexic subjects, for all the stimuli

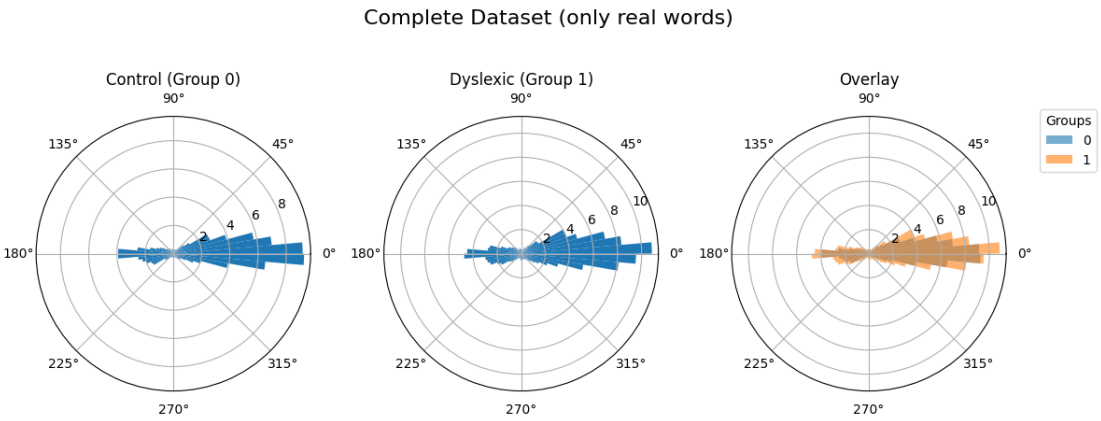
After seeing that there are statistical differences between the two distributions on all the stimuli, we proceed to analyse in detail the subgroups defined above. We start by keeping only the six stimuli containing real words, which were probably used as a preliminary test to familiarise participants with the tasks required.

For these stimuli, Control group has 11841 saccades, in average 100 for each subject, with angular mean of 3.54° and standard deviation of 87.82°. The Dyslexic group has 10521 saccades, 130 in average, with mean of 1.52° and standard deviation of 88.87°. Among these data, what stands out most is the standard deviation of the two groups, which, compared to what we have seen in all previous cases, on these stimuli results to be very close between the two groups.

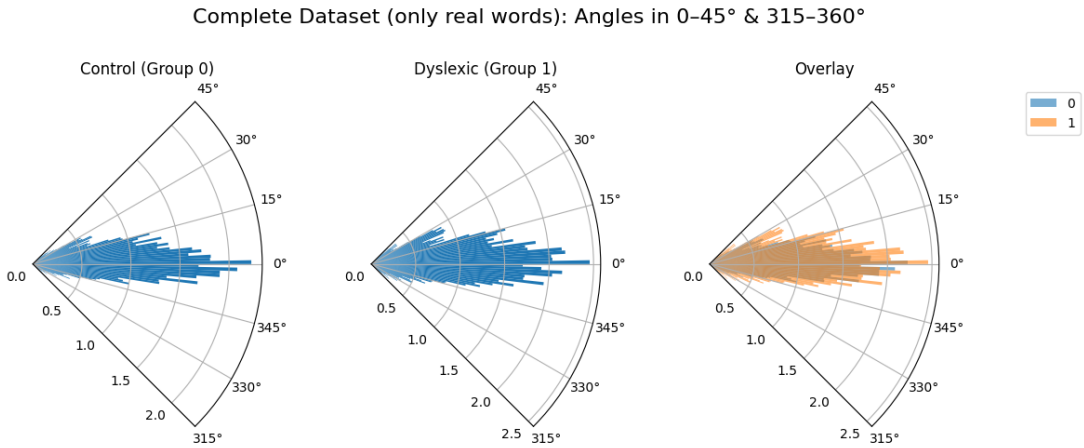
In Figure 4.8, we can notice that beside the difference in the quantity of saccades, we can not identify elements that suggest a difference between the two distributions. In particular, in the overlay plot, we can notice that almost all the bins that appear in one group have correspondence also in the other. Usually the Control group is more compact compared to the Dyslexic one. Only in Figure 4.8c we can see a little difference between the two groups in the angles from 165° to 180°.

This analysis, carried out using graphs, is confirmed by two statistical tests that differ greatly from what has been obtained so far:

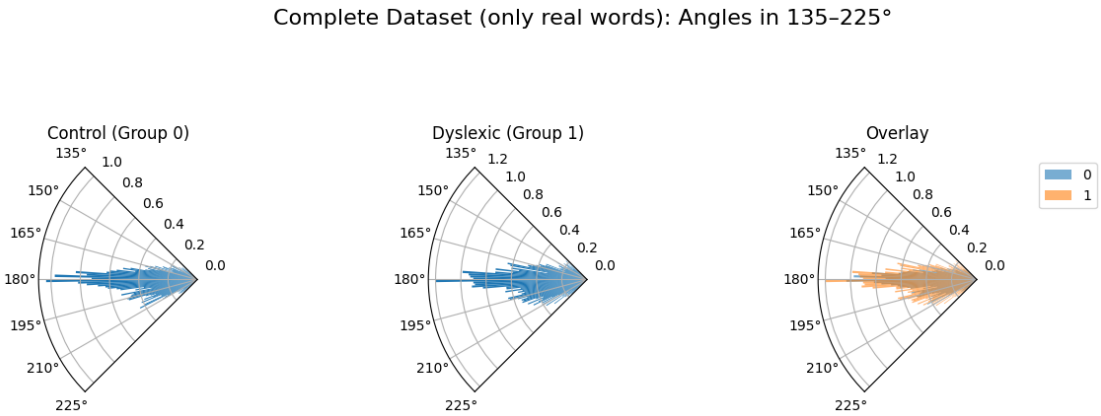
- The **Watsons** U^2 result is very low 0.9281 with also an important increasing of p-value to 0.0040. Although the p-value remains low, the increase obtained suggests that the two distributions do not differ significantly.
- Similarly, the **MANOVA test** shows a very high increase in the p-value and for this reason we can no longer reject the null hypothesis.
  - Very high Wilks' Lambda value equal to 0.9998 and important decreasing of the other three tests to 0.0002.
  - The F-value is 2.1854 with a p-value equal to 0.1125.



(a) Complete Angles Distribution (bins of 5°)



(b) Progressive Angles Distribution (bins of 1°)



(c) Regressive Angles Distribution (bins of 1°)

Figure 4.8.: Dysfluent Readers of German Dataset Dataset - Comparison of the Saccade Angles Distribution for Control and Dyslexic subjects, considering practice stimuli with only real words

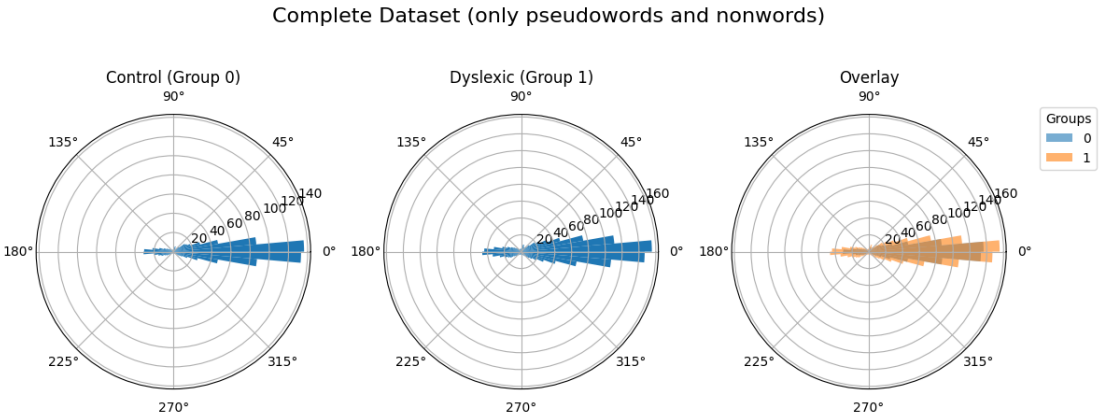
In the opposite way to what we did previously, we will now proceed to analyse the 30 stimuli containing pseudowords and nonwords, without considering the real words stimuli.

In this subset we have for the Control group 119015 total saccades with an average of 1000 for each subject, mean angle of 2.60°, and standard deviation of 69.38°. Instead for the Dyslexic group instead we have 116486 saccades, in average 1438, mean angle of 0.94° and standard deviation of 76.55°. Compared to the previous case we can notice a reduction of number of saccades, in particular for the Control subjects, and angle standard deviation.

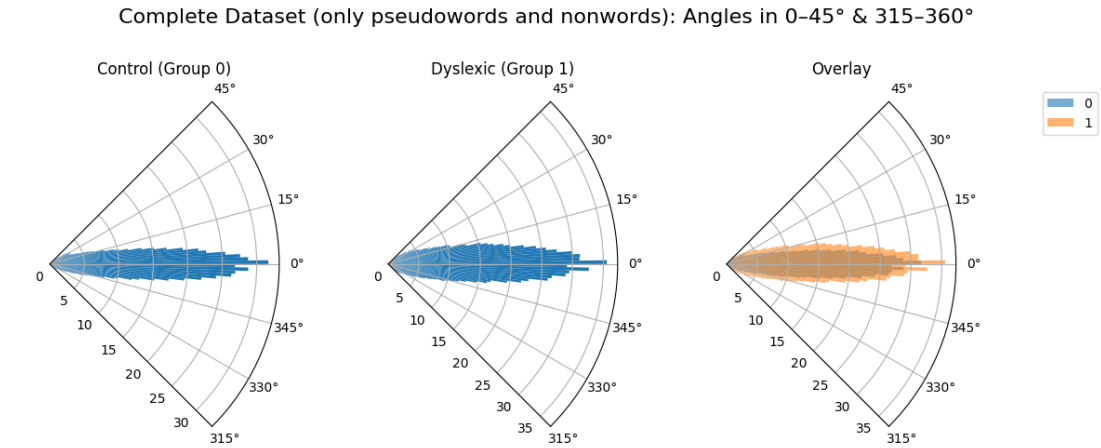
In Figure 4.9 we have a similar observation to the previous one, with the Dyslexic group that presents a much wider distribution and with more saccades compared to the one of the Control subjects.

The statistical tests confirm the presence of differences in the angular distribution, with a slightly increasing of significance:

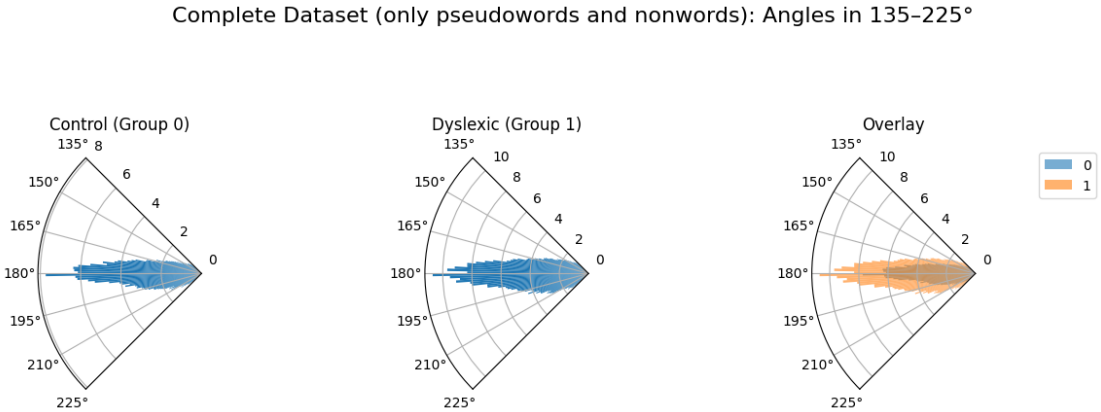
- The **Watsons** U^2 returns the exact same value obtained in the first subset: 38.7578, with a p-value of 0.0010.
- The **MANOVA test** instead presents small differences in the statistic value:
  - Wilks' Lambda equal to 0.9978 and the other three tests equal to 0.0022.
  - The F-value is 261.9562 with a p-value < 0.0001.



(a) Complete Angles Distribution (bins of 5°)



(b) Progressive Angles Distribution (bins of 1°)



(c) Regressive Angles Distribution (bins of 1°)

Figure 4.9.: Dysfluent Readers of German - Comparison of the Saccade Angles Distribution for Control and Dyslexic subjects, considering only pseudowords and nonwords (without real words stimuli)

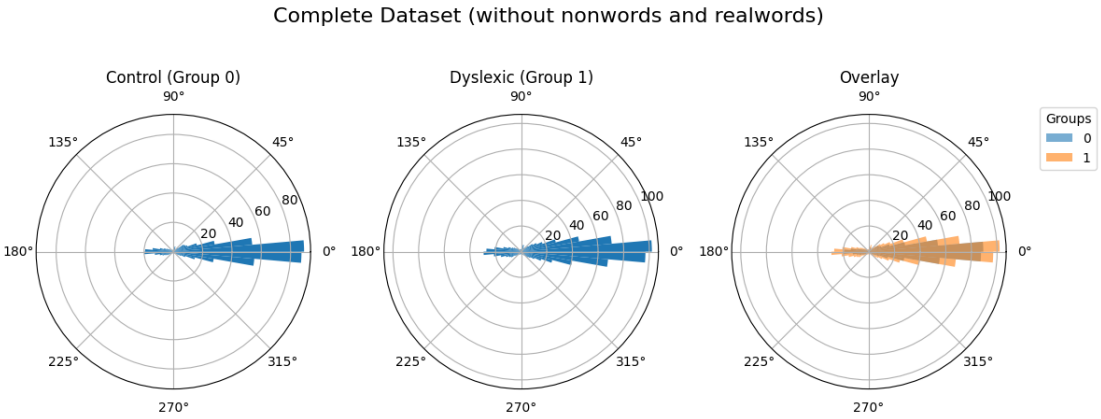
For the last two subsets, we begin by removing stimuli that contain only nonwords (and of course removing again the six stimuli with real words). This way we obtain 20 stimuli with a mix of pseudowords and nonwords.

In the Control group we have 73814 saccades, 620 for each subject, with mean of 2.77° and standard deviation of 69.26°. The Dyslexic group has 71679 saccades, 884 in average, with mean of 1.05° and standard deviation of 75.967°. We can observe that these data are very similar to those obtained when nonwords stimuli are also considered.

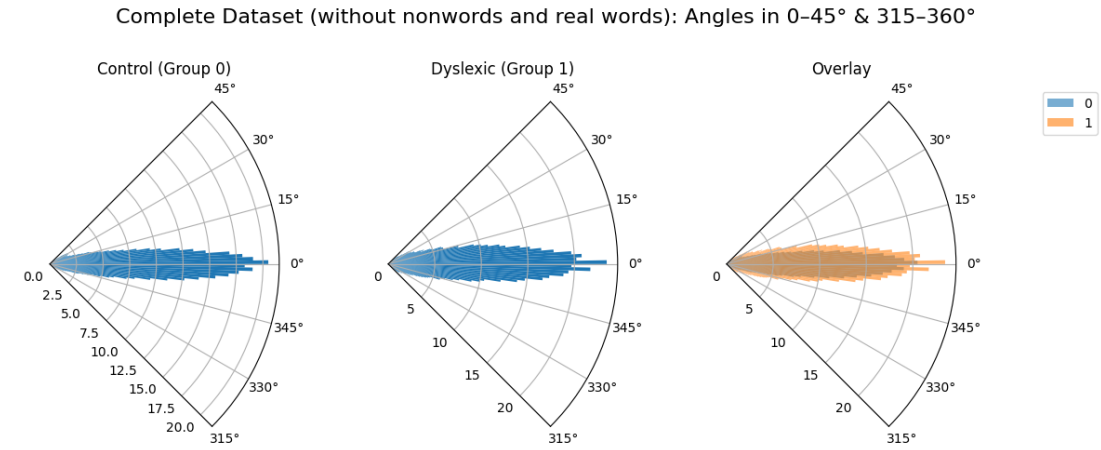
The same pattern is present in Figure 4.10 were all the plots are quite similar to what we have seen in the previous subsets. Only in the Overlay of Figure 4.10c we can note that the difference between the two groups is slightly reduced. Meanwhile, by comparing Figure 4.10b to Figure 4.9b it is possible to see an overall reduction in the distribution of both groups.

Therefore, we continue testing the two distributions. Even in this case we can notice that the results are quite similar to what obtained above:

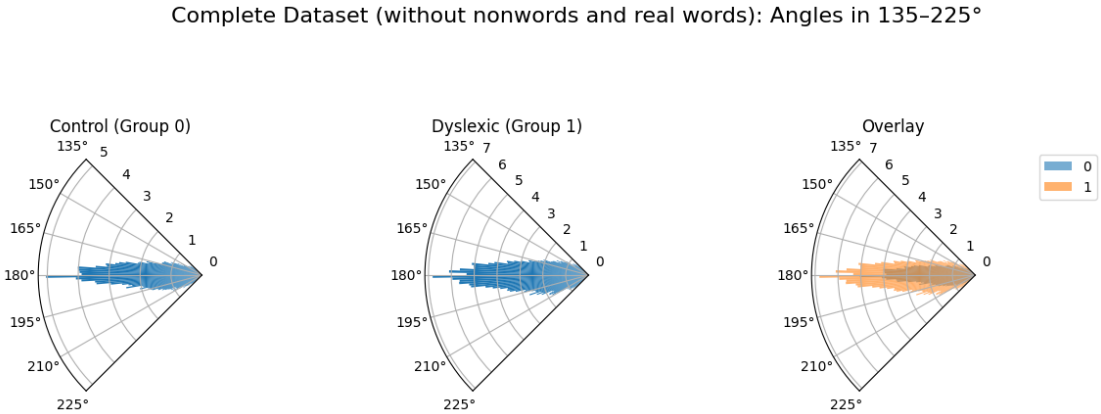
- The **Watsons** U^2 reports a lowering of the statistic value to 22.9921, with a p-value of 0.0010.
- The **MANOVA test** also shows a decreasing of the value of the four tests, but always reflecting a statistical importance of the distribution difference.
  - Wilks' Lambda equal to 0.9980 and the other three tests equal to 0.0020.
  - The F-value is 148.3100 with a p-value < 0.0001.



(a) Complete Angles Distribution (bins of 5°)



(b) Progressive Angles Distribution (bins of 1°)



(c) Regressive Angles Distribution (bins of 1°)

Figure 4.10.: Dysfluent Readers of German - Comparison of the Saccade Angles Distribution for Control and Dyslexic subjects, considering stimuli with pseudo words and nonwords (without stimuli containing only nonwords and real words)

In the last subset we consider only the 10 stimuli containing nonwords, excluding the one with mix of pseudowords and nonwords, and the one that has only real words.

In this case the Control subjects have 45201 saccades, 380 in average, with mean of 2.33° and standard deviation of 69.57°. The Dyslexic group has 44807 saccades, 553 for each subject, with mean of 0.76° and standard deviation of 77.49°. Here the difference of standard deviation between the two groups is a little bit more evident.

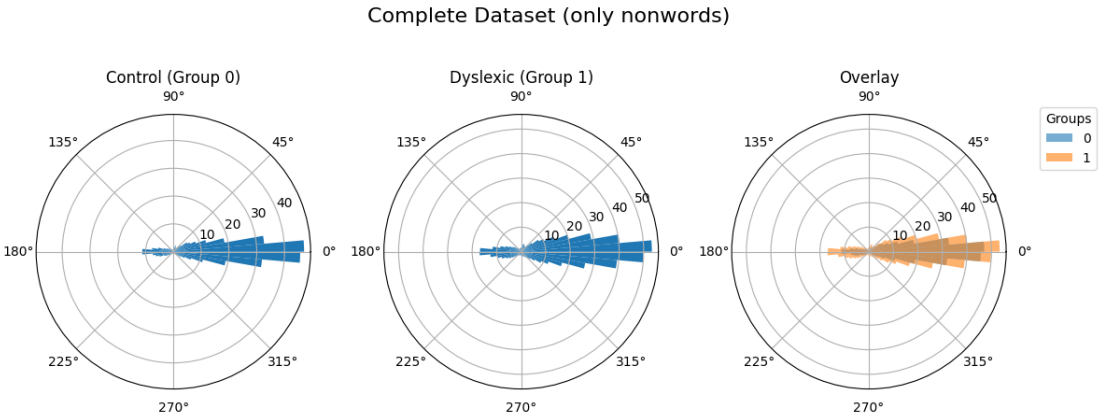
Comparing Figure 4.11 with Figure 4.10 it is very difficult to view significant difference. However, in the progressive saccade, it is possible to see that for both the groups we have an increasing of the angle distribution, reflecting also the increment of standard deviation.

In this case, we obtain two slightly different results for the two statistical tests: the value reported by Watson's U^2 test shows a reduction in the statistical value, while MANOVA reports an increase in the significance of the test:

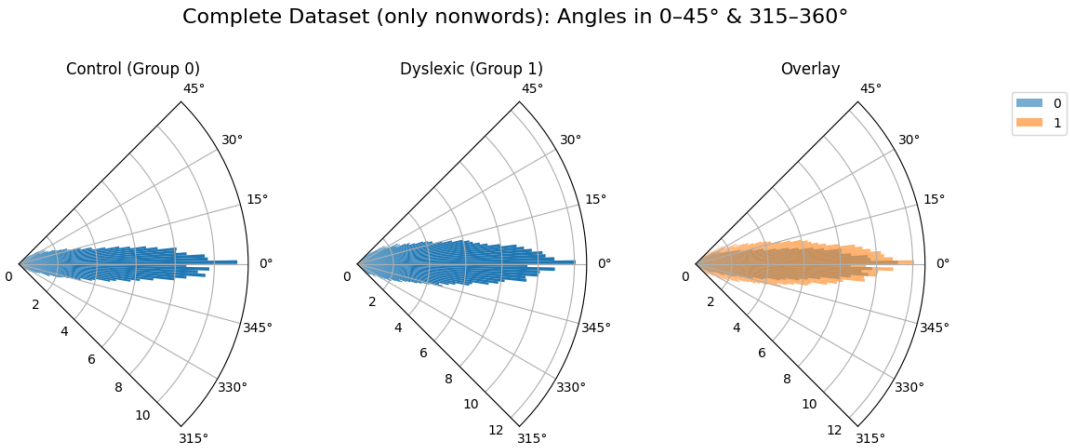
- The **Watsons** U^2 equal to 17.2100, with a p-value of 0.0010.
- The **MANOVA test** reported an increment compared to the other subsets
  - Wilks' Lambda equal to 0.9974 and the other three tests equal to 0.0026.
  - The F-value is 115.6169 with a p-value < 0.0001.

In conclusion, clear differences in the Saccade Angle Distributions between the two groups were observed across all stimulus subgroups, with the exception of the first six stimuli containing only real words. For these stimuli, the MANOVA results did not provide sufficient evidence to conclude that the distributions differ between groups.

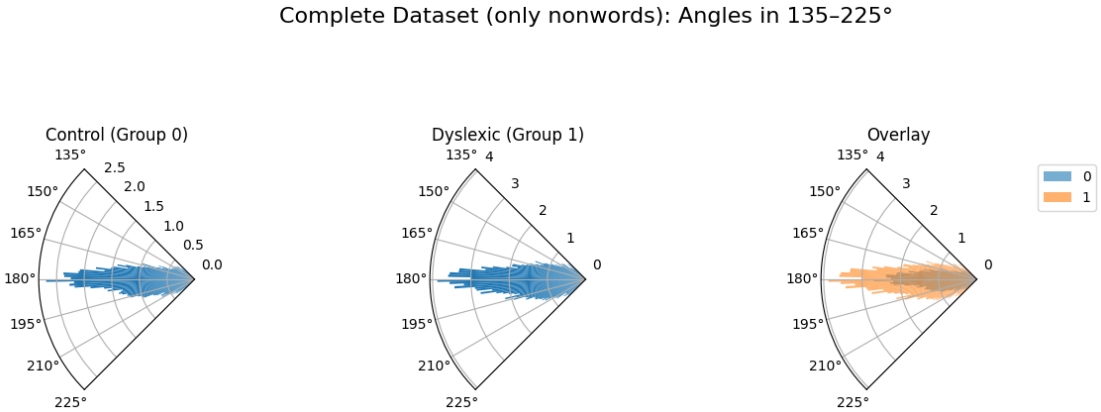
We also conducted MANOVA tests on each individual stimulus. All stimuli associated with the practice task, which also consist only of real words, showed p-values greater than 0.05. Among the remaining 30 stimuli, only stimulus 28 (which contains both pseudowords and nonwords) yielded a p-value above the significance threshold ($p = 0.0644$).



(a) Complete Angles Distribution (bins of 5°)



(b) Progressive Angles Distribution (bins of 1°)



(c) Regressive Angles Distribution (bins of 1°)

Figure 4.11.: Dysfluent Readers of German - Comparison of the Saccade Angles Distribution for Control and Dyslexic subjects, considering stimuli with only nonwords

Comparing metrics statistic between Dyslexia and Control

For the Hedges g analysis, we present the overall effect sizes computed across all stimuli, performing separate tests for the Graz and Munich subject groups. As explained in subsection 2.2.4, these two groups exhibit significant differences in their fixation coordinate distributions. Therefore, they are analysed and classified separately. Additionally, since these eye-movement metrics will later serve as input features for the MLP model, it is important to evaluate their discriminative power through effect size analysis for each group independently. Moreover, the spatial differences in fixation locations could influence the comparison of effect sizes, particularly for the Fixation Entropy metric, which is inherently dependent on spatial distribution.

In Figure 4.12, the results for the Graz subjects show that only the Mean Saccade Amplitude does not cross zero. Meanwhile, all other features display clear differences between the two groups. The pattern is consistent with the results obtained for the ETDD70 Dataset, except that the Fixation Duration (mean and standard deviation) metrics are closer to zero, indicating slightly weaker effects.

For the Munich subjects, shown in Figure 4.13, a generally similar pattern is observed, with minor variations in effect magnitude. When comparing the exact values, most features show smaller effect sizes compared to the Graz group, except for Fixation Duration (mean and standard deviation), which remain comparable across both datasets.

These results, in combination with what was observed in the other statistical tests, provide strong evidence of systematic differences between the two groups. The sampling of visual information by dyslexic readers appears to be slower and more effortful, reflecting a fundamentally different reading process compared to typical readers.

Building upon these statistical evaluations, we now proceed to test the practical effectiveness of these eye-movement features in the classification process, with the aim of automatically identifying the group to which each subject belongs.

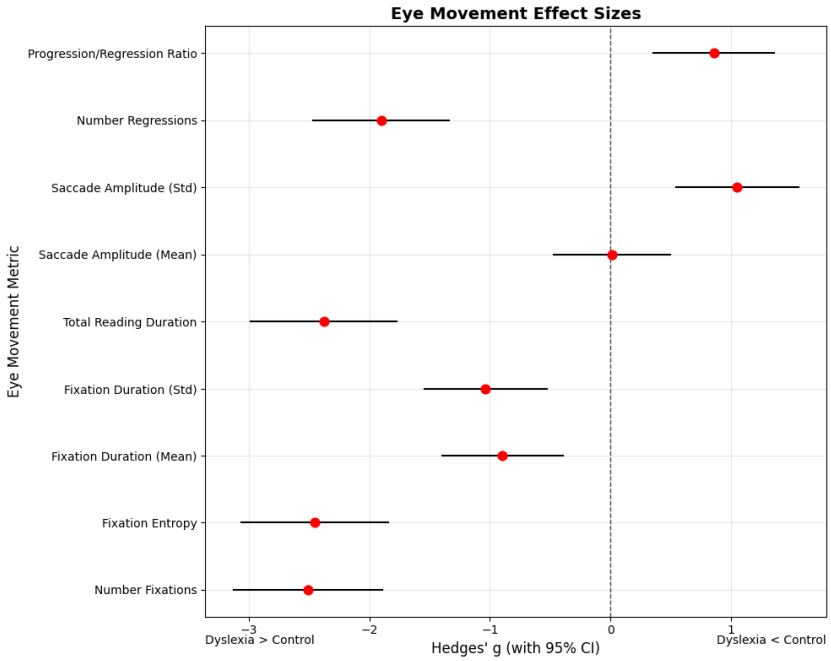


Figure 4.12.: Dysfluent Readers of German - Graz Subjects Hedges' g effect size

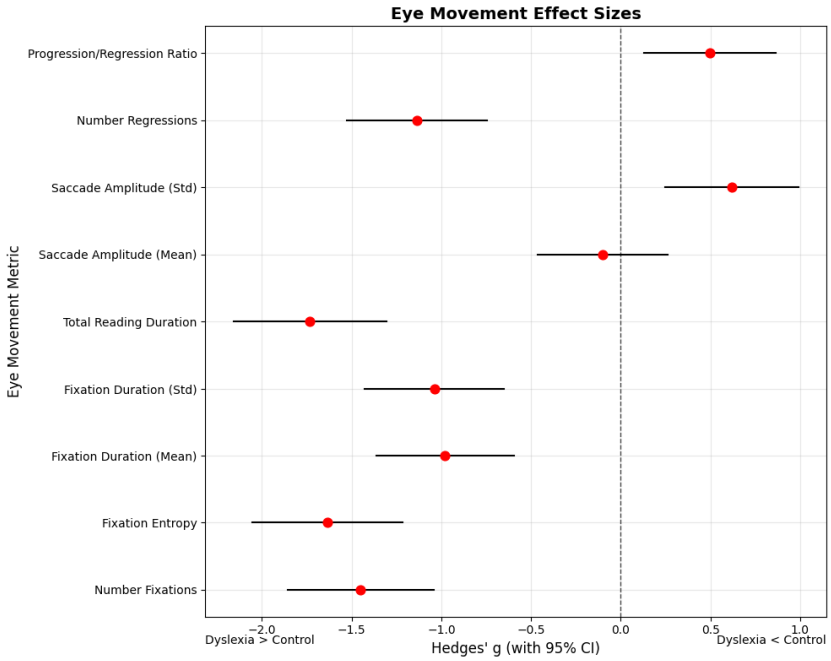


Figure 4.13.: Dysfluent Readers of German - Munich Subjects Hedges' g effect size

4.2. Classification Results

We will present the classification results for the two datasets: ETDD70 and Dysfluent Reader of German. As mentioned above, for the second dataset we will consider the subjects from Graz separately from those from Munich.

For each dataset we will present the results subdivided by the stimuli used and based on the changes of input features or parameters of specific models.

Regarding the MLP test, we will use different groups of eye-tracking measures as input. In particular, we will compare the results using only standard features and introducing the novel similarity measure.

We consider three similarity measures, one of which with two different thresholds:

- Cosine (with Threshold = 10°)
- Cosine (with Threshold = 5°)
- Peak 180
- Peak 180 (with match)

We will have four feature subsets, each of them tested two times with momentum 0.8 and 0.5:

- **Subset 1**: Original and Similarity (Average & Standard Deviation)
- **Subset 2**: Original and Similarity (Average & Standard Deviation, no mean saccade amplitude)
- **Subset 3**: Original and Average Similarity
- **Subset 4**: Original and Average Similarity (no mean saccade amplitude)

In addition, there will also be a final test where we will consider all the features at our disposal.

Then, we will also present and compare the results obtained with the other two models.

4.2.1. ETDD70 Dataset

In this dataset, we have three different stimuli and we repeated the experiments for each stimulus, as presented in the original paper. We will also compare our results with those reported in the paper to identify which methods and configurations led to improved or reduced performance. Finally, we will compare the results obtained using the MLP method with those obtained using the CNN.

Since the original paper does not specify the momentum used, we will report the value as double question mark (**??**).

The results obtained for each stimulus and model will be presented and briefly discussed in the following sections.

Task 1

For the first task, we will present the results obtained with the MLP model over the different input subsets.

In Table 4.1 we report the results of the paper compared to the original setup that we recreated in PyEyeSim and the additional features we extracted. In this case, the variation between the original input and architecture presented in the paper and our setup was only about one feature. We did not consider the "Number of fixations and saccades, w/o the incoming/outgoing saccade" in order to standardize the features extracted from the ROI to those of Task 2 and 3. As explained in the methodology section this has the aims to generalize the model proposed. Even if this means a lower classification accuracy, we still want to have a better comparison of the same features in different configurations.

Features	**Momentum = ??**	**Momentum = 0.8**	**Momentum = 0.5**
Original Paper	**91.43 ± 5.35**		
Local		84.29 ± 11.43	**90 ± 3.5**
Local and Global		82.86 ± 13.25	81.43 ± 14.71
Local, Global and Additional		88.57 ± 5.71	87.14 ± 7

Table 4.1.: ETDD70: Task 1 - MLP classification average accuracy and standard deviations (5-folds) with original and additional features (best results in **bold**)

From the results we can notice that the Original Paper has the highest accuracy, and also the Local features extracted by us obtained very similar results (lower accuracy but also reducing the standard deviation, i.e., more stable results).

Instead, adding the global features did not seem to have a good impact in the accuracy. This could be the reason why in the paper they have decided to extract only local features for this stimulus. Anyway, in the last test, by adding new global features we can observe a notable improvement approaching the best performance.

From Table 4.2, we can notice that by including the similarity measure, we are able to achieve a major improvement over the best result obtained in the paper. From **91.43% ± 5.35%**, we increased the accuracy and reduced the standard deviation to **92.86% ± 4.52%**, using Peak 180 (with and without match). In all the tests the best subset of features, the one that reported the highest accuracy and lower standard deviation in average, was the **Subset 4** with Average Similarity Measure **without mean saccade amplitude**. Moreover, the average saccadic amplitude showed no significant difference between the two groups. This is in agreement with what we have observed in the preliminary analysis.

The similarity measure in combination with the most important global features can provide an input value that significantly improves the results. To be more precise, we need to mention that we removed one global feature. The second-best results (without considering the similarity measure) were obtained using only the local features. In this

context, we also need to demonstrate that the 'Local and Global' test in Table 4.1 would not yield high accuracy results if the mean saccadic amplitude was excluded. This aims to demonstrate that the accuracy improvement is entirely attributable to the similarity measure features.

Similarity Measure	**Features**	**Momentum = 0.8**	**Momentum = 0.5**
Cosine (Threshold = 10°)	Subset 1	84.29 ± 12.29	87.14 ± 8.33
	Subset 2	87.14 ± 11.42	84.29 ± 14.57
	Subset 3	87.14 ± 13.09	87.14 ± 12.29
	Subset 4	88.57 ± 8.57	91.43 ± 2.86
	All	84.29 ± 9.48	82.86 ± 13.25
Cosine (Threshold = 5°)	Subset 1	87.14 ± 5.35	84.29 ± 5.35
	Subset 2	85.71 ± 10.10	81.42 ± 14.71
	Subset 3	87.14 ± 10.50	88.57 ± 9.69
	Subset 4	90.00 ± 7.28	91.43 ± 2.86
	All	84.29 ± 9.48	82.86 ± 13.25
Peak 180	Subset 1	88.57 ± 9.69	82.86 ± 8.57
	Subset 2	88.57 ± 10.69	85.71 ± 12.78
	Subset 3	87.14 ± 10.50	85.71 ± 11.95
	Subset 4	91.42 ± 5.35	**92.86 ± 4.52**
	All	85.71 ± 10.10	85.71 ± 10.10
Peak 180 (with match)	Subset 1	90.00 ± 9.69	85.71 ± 6.39
	Subset 2	88.57 ± 10.69	85.71 ± 12.78
	Subset 3	88.57 ± 11.61	85.71 ± 11.95
	Subset 4	88.57 ± 8.57	**92.86 ± 4.52**
	All	84.29 ± 11.43	82.86 ± 13.25

Table 4.2.: ETDD70: Task 1 - MLP classification average accuracy and standard deviations (5-folds) with original, additional and similarity measure features (best results in **bold**)

As can be seen from the results in Table 4.3, although improvement is evident, the accuracy obtained by removing the mean saccadic amplitude is far from that obtained using the average similarity score. Therefore, we can confirm that for this stimulus excluding the mean saccadic amplitude generally improves the accuracy and that the highest accuracy is achievable by including similarity-based measures in the input features.

Features	Momentum = 0.8	Momentum = 0.5
Local and Global	82.86 ± 13.25	81.43 ± 14.71
Local and Global (without mean saccadic amplitude)	88.57 ± 9.69	85.71 ± 4.52

Table 4.3.: ETDD70: Task 1 - MLP classification accuracy and standard deviations Local and Global features with and without mean saccadic amplitude

In Table 4.4 we present the results obtained via the three CNN models. As anticipated in the methodology chapter, there will be three tests for each model with different batch sizes: 8,16 and 32. Regarding MLP, we tested two different momentum values: 1 (default value) and 0.5.

Model	Batch Size	Momentum = 1	Momentum = 0.5
CNN - 2 Layer	8	42.86 ± 12.78	65.71 ± 14.57
	16	54.29 ± 16.66	**74.29 ± 10.69**
	32	62.86 ± 11.43	71.43 ± 9.04
CNN - 3 Layer	8	65.71 ± 11.43	51.43 ± 14.57
	16	68.57 ± 14.00	54.29 ± 5.71
	32	57.14 ± 15.65	60.00 ± 5.71
CNN - 4 Layer	8	68.57 ± 14.00	65.71 ± 11.43
	16	71.43 ± 9.04	68.57 ± 18.95
	32	57.14 ± 20.20	57.14 ± 18.07

Table 4.4.: ETDD70: Task 1 - CNN models classification average accuracy and standard deviations (5-folds, best results in **bold**)

The best performance was obtained by the smallest CNN (2-layer) using a batch size of 16 and momentum of 0.5. However, the accuracy is significantly lower to that obtained with MLP and we can see how generally the results show low accuracy with very high standard deviation.

Task 2

In Table 4.5, the input features extracted for the MLP were exactly the same as in the paper. The local features were taken from 100 ROIs (10x10 divisions) and the Global features were the same as in the previous stimulus. Then, in Table 4.6 we compare the performance introducing also the similarity measure.

Features	**Momentum = ??**	**Momentum = 0.8**	**Momentum = 0.5**
Original Paper	**90 ± 7.28**		
Global		**92.86 ± 9.04**	88.57 ± 7.28
Global and Additional		85.71 ± 7.82	87.14 ± 5.35

Table 4.5.: ETDD70: Task 2 - MLP classification average accuracy and standard deviations (5-folds) with original and additional features (best results in **bold**)

Similarity Measure	**Features**	**Momentum = 0.8**	**Momentum = 0.5**
Cosine (Threshold = 10°)	Subset 1	88.57 ± 7.28	88.57 ± 7.28
	Subset 2	81.43 ± 11.61	81.43 ± 11.61
	Subset 3	85.71 ± 6.39	85.71 ± 9.04
	Subset 4	87.14 ± 5.35	88.57 ± 3.50
	All	84.29 ± 7.00	84.29 ± 7.00
Cosine (Threshold = 5°)	**Subset 1**	**90 ± 5.71**	88.57 ± 7.28
	Subset 2	81.43 ± 9.69	81.43 ± 11.61
	Subset 3	87.14 ± 7.00	85.71 ± 9.04
	Subset 4	87.14 ± 5.35	88.57 ± 3.50
	All	82.85 ± 9.69	82.85 ± 9.69
Peak 180	Subset 1	88.57 ± 7.28	88.57 ± 7.28
	Subset 2	87.14 ± 13.85	85.71 ± 12.78
	Subset 3	90.00 ± 7.28	**91.43 ± 8.33**
	Subset 4	88.57 ± 3.50	88.57 ± 3.50
	All	90.00 ± 9.69	85.71 ± 7.82
Peak 180 (with match)	**Subset 1**	**90 ± 5.71**	88.57 ± 7.28
	Subset 2	80.00 ± 10.50	81.43 ± 11.61
	Subset 3	88.57 ± 8.57	87.14 ± 10.50
	Subset 4	87.14 ± 5.35	87.14 ± 5.35
	All	85.71 ± 7.82	84.29 ± 7.00

Table 4.6.: ETDD70: Task 2 - MLP classification average accuracy and standard deviations (5-folds) with original, additional and similarity measure features (best results in **bold**)

We observe that replicating the original method yields slightly different results compared to those reported in the paper. This discrepancy may be attributed to minor variations in the architecture, random initialization, or the specific momentum setting we selected. Despite this, as discussed for the previous stimulus, the two outcomes remain fully comparable: although our model shows a slight increase in accuracy, it also exhibits a higher standard deviation. However, unlike the previous case, incorporating the additional features did not lead to any improvement in classification performance.

Firstly, by comparing the accuracy, we can note that this time we did not manage to improve the performance obtained in the paper. However, multiple results show similar accuracy to those in Table 4.5. In particular, the best result was obtained in the second subset using Peak 180 (91.43 ± 8.33, which is a result falling between the best two reported above). Finally, if we consider the best result taking into account both accuracy and standard deviation across the 5 folds, we can see that subset 1 with the Cosine method (threshold 5°) achieved an accuracy of 90 ± 5.71, thus reaching the same performance obtained in the paper while reducing the standard deviation.

We will now present the results obtained through the three CNN models. In Table 4.7 we can observe that the first version of CNN is the one that performed best, in this case even with a smaller batch size, compared to the previous task. For this stimulus we can notice that every model managed to obtain an accuracy above 70%. However, performance is still far from those obtained through MLP.

Model	**Batch Size**	**Momentum = 1**	**Momentum = 0.5**
CNN - 2 Layer	**8**	51.43 ± 17.14	**82.86 ± 16.66**
	16	71.43 ± 15.65	80.00 ± 11.43
	32	77.14 ± 11.43	80.00 ± 11.43
CNN - 3 Layer	8	48.57 ± 17.14	74.29 ± 5.71
	16	57.14 ± 23.90	74.29 ± 5.71
	32	74.29 ± 10.69	77.14 ± 14.57
CNN - 4 Layer	8	65.71 ± 7.00	77.14 ± 14.57
	16	68.57 ± 16.66	71.43 ± 12.78
	32	71.43 ± 15.65	74.29 ± 14.00

Table 4.7.: ETDD70: Task 2 - CNN models classification accuracy and standard deviations (5-folds, best results in **bold**)

Task 3

Now we will see the results obtained for the last stimulus in the dataset. The only variation made for the MLP is in the number of ROIs selected: in the paper 102 ROIs are used while in our case (again to keep the model as standard as possible) we considered 100 ROIs, exactly as in the previous task.

In the original paper it is not explicitly specified, but we can presume that they used 6 horizontal and 17 vertical divisions to obtain the 102 ROIs. This strategy probably allowed an improvement in accuracy as can be seen from the results obtained in Table 4.8. The additional features did not seem to bring any improvements, and also increased the standard deviation.

Features	**Momentum = ??**	**Momentum = 0.8**	**Momentum = 0.5**
Original Paper	**88.57 ± 5.71**		
Global		82.86 ± 11.61	82.86 ± 9.69
Global and Additional		81.43 ± 13.25	80 ± 13.85

Table 4.8.: ETDD70: Task 3 - MLP classification average accuracy and standard deviations (5-folds) with original and additional features (best results in **bold**)

In Table 4.9 we report the results obtained by including the similarity measure feature. We can notice that with the Cosine similarity measure (10° threshold) we have the same accuracy of the paper but lower standard deviations, from 5.71% to 3.50%. Meanwhile, with Peak 180 we also have an improvement in the accuracy, with the original features and average similarity score (without mean saccadic amplitude). By using the similarity measure, we improved the performance obtained with only the original features and the previously best results of the paper. This demonstrates once again how similarity measures can significantly increase the accuracy of the model.

Similarity Measure	Features	Momentum = 0.8	Momentum = 0.5
Cosine (Threshold = 10°)	Subset 1	82.86 ± 10.69	81.43 ± 11.61
	Subset 2	80.00 ± 11.43	82.86 ± 9.69
	Subset 3	81.43 ± 11.61	82.86 ± 13.25
	Subset 4	85.71 ± 4.52	**88.57 ± 3.50**
	All	82.86 ± 9.69	78.57 ± 16.29
Cosine (Threshold = 5°)	Subset 1	82.86 ± 10.69	80.00 ± 10.50
	Subset 2	80.00 ± 11.43	82.86 ± 9.69
	Subset 3	81.43 ± 11.61	81.43 ± 11.61
	Subset 4	85.71 ± 4.52	87.14 ± 5.35
	All	82.86 ± 9.69	78.57 ± 16.29
Peak 180	Subset 1	81.42 ± 13.25	78.57 ± 10.10
	Subset 2	78.57 ± 12.78	82.86 ± 9.69
	Subset 3	82.86 ± 13.25	81.43 ± 14.71
	Subset 4	87.14 ± 5.35	**90 ± 3.50**
	All	80.00 ± 13.10	78.57 ± 16.29
Peak 180 (with match)	Subset 1	82.86 ± 10.69	81.43 ± 8.57
	Subset 2	80.00 ± 11.43	81.43 ± 9.69
	Subset 3	80.00 ± 13.09	80 ± 13.09
	Subset 4	85.71 ± 6.39	84.29 ± 8.33
	All	82.86 ± 10.69	78.57 ± 16.29

Table 4.9.: ETDD70: Task 3 - MLP classification average accuracy and standard deviations (5-folds) with original, additional and similarity measure features (best results in **bold**)

Now we compare the results with the one obtained by the CNN. In Table 4.10 we can see that the best average result over the 5 fold is 80.00% ± 7.00%, which is 10%± 3.50% lower than MLP best result. Also we can notice that in average, the performance is very low with results even lower than 50%.

Model	Batch Size	Momentum = 1	Momentum = 0.5
CNN - 2 Layer	8	45.71 ± 16.66	68.57 ± 14.00
	16	65.71 ± 11.43	65.71 ± 17.14
	32	71.43 ± 15.65	65.71 ± 21.38
CNN - 3 Layer	8	68.57 ± 14.00	65.71 ± 11.43
	16	71.43 ± 9.04	68.57 ± 18.95
	32	57.14 ± 20.20	57.14 ± 18.07
CNN - 4 Layer	8	68.57 ± 5.71	60.00 ± 16.66
	16	**80.00 ± 7.00**	68.57 ± 16.66
	32	51.43 ± 14.57	71.43 ± 12.78

Table 4.10.: ETDD70: - CNN models classification accuracy and standard deviations (5-folds, best results in **bold**)

In chapter 5, we will inspect and discuss in more detail what problems CNN could have with this dataset, compared instead to the promising performance reported in the original paper.

4.2.2. Dysfluent Readers of German - Graz Subjects

In this section, we present the results obtained from the Graz subject. We begin by explaining in details the performance of the MLP model using the different feature combinations tested. Next, we follow the same structure used previously to report the results of the three CNN architectures. Finally, we conclude with the results achieved using the GHMM approach.

MLP

As mentioned in the methodology chapter, as there are some changes in the architecture of the MLP, we no longer used the division into ROIs. Instead, the input vector was based entirely on global features and concatenated similarity measures for each stimulus presented to the participant. These changes, coupled with the fact that we no longer have baseline values for comparing accuracy, led us to slightly different tests that nevertheless retain validity when comparing feature efficiency. Since we have no other data in this dataset on which to compare the classification results, we used the three standard tests with global and additional features, but without the similarity measures.

Features	Momentum = 0.8	Momentum = 0.5
Global	91.11 ± 8.31	**95.56 ± 5.44**
Global and Additional	88.89 ± 9.93	91.11 ± 8.31
Global (without mean saccadic amplitude)	86.67 ± 8.31	86.67 ± 8.31

Table 4.11.: Dysfluent Readers of German: Graz Subjects - MLP classification average accuracy and standard deviations (5-folds) with original and additional features (best results in **bold**)

Only by using the original features, we already obtained a very high accuracy. We can also notice that we have no improvements when adding additional features or when removing the mean saccadic amplitude.

Similarity Measure	Features	Momentum = 0.8	Momentum = 0.5
Cosine (Threshold = 10°)	Subset 1	93.33 ± 5.44	93.33 ± 5.44
	Subset 2	93.33 ± 8.89	91.11 ± 10.89
	Subset 3	93.33 ± 5.44	93.33 ± 5.44
	Subset 4	93.33 ± 5.44	91.11 ± 4.44
	All	93.33 ± 5.44	**100 ± 0**
Cosine (Threshold = 5°)	Subset 1	91.11 ± 8.31	93.33 ± 5.44
	Subset 2	95.56 ± 5.44	95.56 ± 5.44
	Subset 3	93.33 ± 5.44	95.56 ± 5.44
	Subset 4	88.89 ± 7.03	88.89 ± 7.03
	All	95.56 ± 5.44	97.78 ± 4.44
Peak 180	Subset 1	88.89 ± 9.94	91.11 ± 8.31
	Subset 2	88.89 ± 9.94	97.78 ± 4.44
	Subset 3	93.33 ± 5.44	93.33 ± 5.44
	Subset 4	88.89 ± 7.03	91.11 ± 4.44
	All	91.11 ± 8.31	93.33 ± 8.89
Peak 180 (with match)	Subset 1	97.78 ± 4.44	95.56 ± 8.89
	Subset 2	93.33 ± 8.89	**100 ± 0**
	Subset 3	93.33 ± 5.44	93.33 ± 5.44
	Subset 4	95.55 ± 5.44	93.33 ± 5.44
	All	95.56 ± 5.44	95.55 ± 8.89

Table 4.12.: Dysfluent Readers of German: Graz Subjects - MLP classification average accuracy and standard deviations (5-folds) with original, additional and similarity measure features (best results in **bold**)

This time the **Peak 180 with matching sequence** and **Cosine with 10°** have the highest accuracy by correctly classifying all subjects across all 5 folds.

To gain a deeper understanding of the results, we further evaluated the best-performing configuration Peak 180 with matching sequence using a 100-fold stratified cross-validation procedure. This extended evaluation achieved the best average accuracy of 91.89% ± 8.44. The result was obtained with subset 2 (the same as the previous test) and momentum of 0.5. Integrating the similarity measure yields again in better accuracy results. In Table 4.13 it is possible to find the complete comparison between subset and the original features.

Features	**Momentum = 0.8**	**Momentum = 0.5**
Original	87.67 ± 9.54	89.00 ± 9.49
Original (without mean saccade amplitude)	87.67 ± 9.41	88.56 ± 9.36
Original and Additional	86.56 ± 10.23	88.33 ± 10.93
Subset 1	90.89 ± 8.94	91.89 ± 8.72
Subset 2	91.11 ± 8.31	**91.89 ± 8.44**
Subset 3	88.67 ± 8.46	89.44 ± 9.35
Subset 4	89.56 ± 10.16	90.22 ± 9.06
All	90.22 ± 9.33	91.22 ± 8.79

Table 4.13.: Dysfluent Readers of German: Graz Subjects - MLP classification average accuracy and standard deviations - Tested on 100 folds using Peak 180 with match similarity. (best result in **bold**)

Another interesting observation is that by using only the similarity measurement features (average and standard deviation), we obtain a very high accuracy, **91.00**% ± **9.12**%. This result once again reinforces our hypothesis that similarity measures effectively allow to differentiate the scanpaths of Control and Dyslexic subjects.

CNN

Model	Batch Size	Momentum = 1	Momentum = 0.5
CNN - 2 Layer	16	51.67 ± 3.33	81.67 ± 6.24
	32	76.67 ± 6.24	**80.00 ± 6.67**
	64	78.33 ± 6.67	78.33 ± 6.67
CNN - 3 Layer	16	60.00 ± 20.00	76.67 ± 9.72
	32	71.67 ± 10.00	80.00 ± 12.47
	64	73.33 ± 11.06	78.33 ± 6.67
CNN - 4 Layer	16	80.00 ± 13.54	76.67 ± 11.06
	32	71.67 ± 8.50	76.67 ± 11.06
	64	70.00 ± 18.71	76.67 ± 11.06

Table 4.14.: Dysfluent Readers of German: Graz Subjects - CNN models classification average accuracy and standard deviations (5-folds, best results in **bold**)

The smallest model once again demonstrated superior performance within this dataset. Nevertheless, its results were still considerably inferior to those obtained using the MLP.

GHMM

In this section we will look at the results obtained using Gaussian Hidden Markov Models (GHMM). As also explained in the methodology chapter, we used two models (one per group) for each stimulus. A subject is classified according to a majority vote based on the score given by the models.

We conducted five tests, dividing the stimuli into four subsets, plus one test using all available stimuli:

- Stimuli with Pseudohomophones and Nonwords (from stimulus 1 to 10)
- Stimuli with Pseudohomophones and Nonwords (from stimulus 21 to 30)
- Stimuli with Pseudohomophones and Nonwords (from stimulus 1 to 10 and from 21 to 30)
- Stimuli with only Nonwords (from stimulus 11 to 20)
- All Stimuli (from stimulus 1 to 30)

Stimuli	Result
Pseudohomophones and Nonwords (1 to 10)	57.78 ± 10.89
Pseudohomophones and Nonwords (21 to 30)	77.78 ± 14.05
Pseudohomophones and Nonwords (1 to 10 & 21 to 30)	80 ± 10.89
Only Nonwords (11 to 20)	86.67 ± 12.96
Mixed (11 to 30)	88.89 ± 12.17
All Stimuli (1 to 30)	**91.12 ± 8.31**

Table 4.15.: Dysfluent Readers of German: Graz Subjects - GHMM classification average accuracy and standard deviations (5-folds, best results in **bold**)

First, it is evident that when using only the first set of Pseudohomophones and Nonwords, the system struggled to accurately differentiate between the two classes. Instead, by using only the nonwords stimuli, we have already a very high accuracy, even better than the best result of the CNN. However, we can see that when using all the stimuli, we obtained the highest accuracy. Compared with the MLP, we managed to get results in line with multiple tests done with the neural network, using much less information about the scanpath.

4.2.3. Dysfluent Readers of German - Munich Subjects

We now continue the analysis by focusing on the Munich subjects. As with the Graz subjects, we applied a similar methodology to ensure consistency in the evaluation process. This parallel approach enables a comparison between the two subsets.

MLP

Features	**Momentum = 0.8**	**Momentum = 0.5**
Global	88.89 ± 9.93	91.11 ± 10.89
Global and Additional	93.33 ± 8.89	**95.56 ± 5.44**
Global (without mean saccadic amplitude)	88.89 ± 9.93	88.89 ± 9.93

Table 4.16.: Dysfluent Readers of German: Munich Subjects - MLP classification average accuracy and standard deviations (5-folds) with original and additional features (best results in **bold**)

Using the Global and Additional features, it is possible to observe immediately how the model can reach a very high accuracy. Moreover, we can notice that removing the mean saccadic amplitude, seems to not bring an improvement in the performance.

Similarity Measure	Features	Momentum = 0.8	Momentum = 0.5
Cosine (Threshold = 10°)	Subset 1	93.33 ± 8.89	91.11 ± 8.31
	Subset 2	93.33 ± 8.89	91.11 ± 8.31
	Subset 3	93.33 ± 5.44	88.89 ± 9.94
	Subset 4	93.33 ± 5.44	88.89 ± 7.03
	All	93.33 ± 8.89	93.33 ± 8.89
Cosine (Threshold = 5°)	**Subset 1**	**97.78 ± 4.44**	95.56 ± 5.44
	Subset 2	91.11 ± 10.89	91.11 ± 12.96
	Subset 3	91.11 ± 8.31	91.11 ± 8.31
	Subset 4	95.56 ± 5.44	91.11 ± 8.31
	All	95.56 ± 5.44	88.89 ± 9.94
Peak 180	Subset 1	91.11 ± 8.31	93.33 ± 8.89
	Subset 2	93.33 ± 5.44	93.33 ± 5.44
	Subset 3	91.11 ± 8.31	88.89 ± 7.03
	Subset 4	93.33 ± 5.44	93.33 ± 5.44
	All	91.11 ± 8.31	91.11 ± 8.31
Peak 180 (with match)	Subset 1	**97.78 ± 4.44**	95.56 ± 8.89
	Subset 2	91.11 ± 8.31	93.33 ± 8.89
	Subset 3	91.11 ± 4.44	93.33 ± 5.44
	Subset 4	93.33 ± 5.44	88.89 ± 7.03
	All	91.11 ± 8.31	91.11 ± 8.31

Table 4.17.: Dysfluent Readers of German: Munich Subjects - MLP classification average accuracy and standard deviations (5-folds) with original, additional and similarity measure features (best results in **bold**)

From the results obtained, it can be seen that both Cosine and Peak 180 (with match) yielded improvements in accuracy. We can also see that the first subset was the one that got better results. In this case, removing the mean saccadic amplitude did not lead to any improvement.

As we have seen in Table 4.13 for the Graz subjects, we replicated the best test by increasing the number of folds. The model using the cosine method with threshold 5° on 100 folds the best average accuracy of $91.44\% \pm 9.54\%$, but this time with the subset 4 and momentum 0.8. In Table 4.18 we have can see all the results obtained with the different set of features.

Features	Momentum = 0.8	Momentum = 0.5
Original	91.22 ± 9.60	90.22 ± 10.34
Original (without mean saccade amplitude)	91.22 ± 9.60	90.89 ± 9.99
Original and Additional	91.22 ± 9.20	90.33 ± 9.12
Subset 1	89.33 ± 10.41	89.11 ± 10.18
Subset 2	90.11 ± 10.05	89.22 ± 10.24
Subset 3	90.44 ± 10.06	90.11 ± 9.67
Subset 4	**91.44 ± 9.54**	91.11 ± 10.06
All	90.00 ± 10.60	89.56 ± 10.28

Table 4.18.: Dysfluent Readers of German: Munich Subjects - MLP classification average accuracy and standard deviations - Tested on 100 folds using Cosine with threshold 5° similarity. (best result in **bold**)

Overall, the results between the two groups of subjects are very similar and in both the cases we can observe how including the similarity measure features improves the classification model.

CNN

Model	Batch Size	Momentum = 1	Momentum = 0.5
CNN - 2 Layer	16	48.89 ± 5.44	84.44 ± 5.44
	32	75.56 ± 8.31	86.67 ± 4.44
	64	73.33 ± 8.89	84.44 ± 5.44
CNN - 3 Layer	16	75.56 ± 14.74	88.89 ± 0.00
	32	84.44 ± 5.44	**88.89 ± 0.00**
	64	64.44 ± 20.37	82.22 ± 8.89
CNN - 4 Layer	16	80.00 ± 12.96	80.00 ± 4.44
	32	64.44 ± 12.96	80.00 ± 4.44
	64	60.00 ± 19.37	80.00 ± 4.44

Table 4.19.: Dysfluent Readers of German: Munich Subjects - CNN models classification average accuracy and standard deviations (5-folds, best results in **bold**)

This time CNNs achieved good results, in particular the 3 layers model showed a stable average accuracy of 88.89%, over five different folds. In this specific dataset, unlike the Graz subjects, the two classes are balanced and we also have fewer cases where subjects had no data for a specific stimulus. Although these are good results, they are all below 90%, whereas with MLP in multiple tests got significantly better results and even maintained a low standard deviation.

GHMM

As seen in the previous set of subjects, we will now show and comment the results obtained for each stimuli subset.

Stimuli	Result
Pseudohomophones and Nonwords (1 to 10)	86.67 ± 8.31
Pseudohomophones and Nonwords (21 to 30)	84.44 ± 8.89
Pseudohomophones and Nonwords (1 to 10 & 21 to 30)	93.33 ± 8.89
Only Nonwords (11 to 20)	80 ± 12.96
All Stimuli (1 to 30)	**95.56 ± 5.44**

Table 4.20.: Dysfluent Readers of German: Munich Subjects - GHMM classification average accuracy and standard deviations (5-folds, best results in **bold**)

GHMMs were able again to achieve a similar level of accuracy of the MLP, and also outperformed the CNN results. Notably, this was achieved using only fixation order and position, without any specialized feature extraction, indicating strong potential for this approach. Since for both the Graz and Munich subjects yielded results consistent with those obtained via the MLP, in the conclusion we will elaborate and analyse more in details this method.

4.3. Dysfluent Readers of German - Subjects with no Group

In section 2.2, we highlighted the necessity of more closely examining the influence of subjects without a defined group, as we decide to include them into the Control group during classification. Specifically, we aim to evaluate the models performance using the Graz subjects in order to assess their influence on the classification process. As previously discussed, the Munich subjects who were not assigned to any group, were excluded during the data cleaning phase. Therefore, it is not necessary to include them in the current evaluation.

We took into account the best results obtained by Graz subjects using the MLP and we repeated the same test using the respective best similarity score over 100 folds. In Table 4.21, we see in details the new result for each set of feature.

Features	Momentum = 0.8	Momentum = 0.5
Original	91.25 ± 9.10	91.13 ± 10.04
Original (without mean saccade amplitude)	90.63 ± 8.72	91.13 ± 8.89
Original and Additional	90.25 ± 8.94	91.63 ± 8.67
Subset 1	**93.50 ± 7.39**	93.50 ± 8.38
Subset 2	93.00 ± 7.77	93.25 ± 7.59
Subset 3	92.00 ± 8.20	93.13 ± 7.98
Subset 4	92.00 ± 8.39	92.25 ± 8.98
All	93.25 ± 8.19	93.25 ± 8.37

Table 4.21.: Dysfluent Readers of German: Graz Subjects (without ungrouped subjects) - MLP classification average accuracy and standard deviations - Tested on 100 folds using Peak 180 with match similarity. (best result in **bold**)

In Table 4.22, when comparing the best results for the two groups of subjects with the one previously obtained, we can see how removing the non grouped subjects leads to a slightly increasing of the performance. This means that these subjects, or at least some of them, were indeed difficult to classify, and not considering them makes classification easier.

All Subjects	Without No Group
91.89 ± 8.59	93.50 ± 7.39

Table 4.22.: Dysfluent Readers of German - average accuracy and standard deviation with and without no grouped subjects (100-folds)

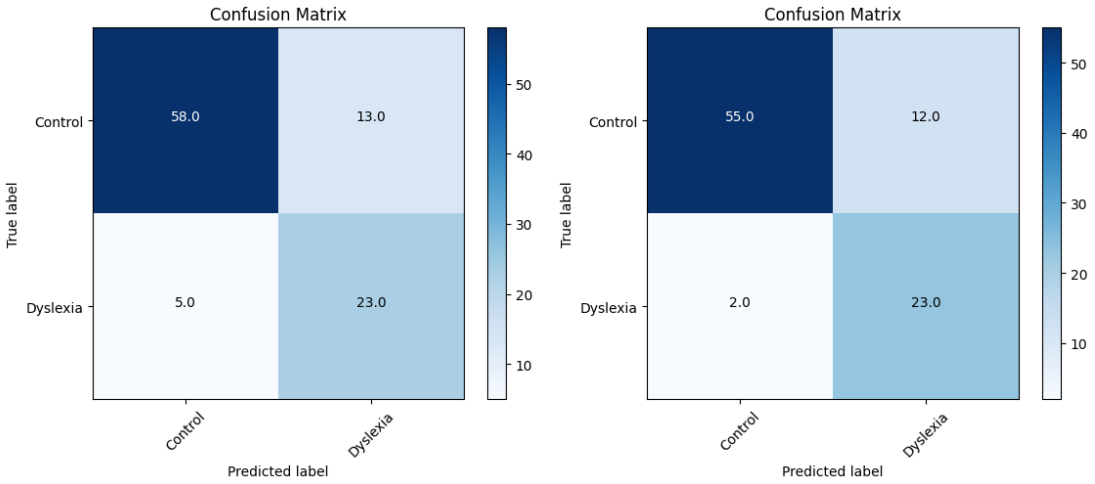
To examine in detail the actual impact of these subjects on the classification outcome, we compare the resulting confusion matrices obtained from the two tests reported above. Since the evaluation involves 100 folds and the same subjects may appear multiple times in the testing sets, each matrix shows the number of subjects that have been predicted as a specific group at least once.

In Figure 4.14, we can observe the two confusion matrices. In the diagonal we have the subjects correctly classified, which Predicted label corresponds to the True label. Instead, in the antidiagonal, we have the misclassified subjects, false positive in the top right (Control subjects predicted as Dyslexic) and false negative in the bottom left (Dyslexic subjects predicted as Control).

In both tests, there are 23 Dyslexic subjects. Meanwhile, there are 60 control subjects in the first test and 56 in the second one. By analyzing the results obtained, we can derive several noteworthy conclusions:

- In both tests we have that all the Dyslexic subjects are correctly classified in at least one fold.

- In the first test over 100 folds, 2 control subjects are never correctly classified. This number decreases to a single Control subject in the second test
  - The same pattern is visible for the false positive, which decreases from 13 in the first test to 12.
  - This result means that between the 4 subjects with no group there is one that it is always misclassified.
- A particularly interesting observation is that removing the ungrouped subject results in a substantial reduction in false negatives. In fact, we can see that from 5 subjects in Figure 4.14a, they decrease to just 2 subjects in Figure 4.14b. This means that by removing these 4 subjects, the neural network has less difficulties to identify Dyslexic subjects.



(a) All Subjects (b) Without No Group

Figure 4.14.: Dysfluent Readers of German: Graz Subjects - Confusion Matrix of unique assigned label (100-folds), with and without "no group" subjects

In conclusion, the result demonstrates that these subjects without a group are sometimes misclassified as dyslexic, and in general they cause a decreasing of the accuracy also in for the prediction of Dyslexic subjects. Therefore the best results obtained with the Graz subjects is 93.50% ± 7.39%.

5. Discussion

5.1. Key Findings

Based on the analysis conducted throughout this thesis and the results obtained, several key findings emerge regarding the use of eye-movement analysis for recognizing learning disorders:

Robust Dyslexia Classification using Eye-Movement Metrics: The study successfully demonstrated that eye-movement metrics, particularly when integrated with machine learning models, provide a robust and generalizable framework for early dyslexia detection. We verified that neural networks trained on eye-tracking metrics achieve excellent results in terms of accuracy and generalizability. Additionally, they demonstrate efficient use of resources, particularly during the training phase, consistent with previous work on the ETDD70 Dataset [11]. By using GHMM, we also introduced another machine learning model that was not considered in previous studies[12, 42, 43, 44, 45]. On short scanpaths, this models has demonstrated results in line with those obtained using neural networks.

Therefore, the study achieved its primary objective of establishing a generalizable framework for distinguishing Control and Dyslexic subjects analysing eye-movement scanpaths.

Statistical Differences and Feature Utility:

- **Distinct Saccade Angle Distributions:** Through both **Watsons Uš and Multivariate Analysis of Variance (MANOVA) tests**, the analysis consistently demonstrated clear and statistically significant differences in the distribution of saccade angles between control and dyslexic subjects. This result is across all stimuli of the ETDD70 Dataset and almost all the ones of the Dysfluent Readers of German Dataset (not for the stimulus 28 and stimuli with real words). This finding highlights the distinct oculomotor patterns associated with learning disorders[34].
- **Feature Relevance and Selection:** The **Hedges' g effect size** analysis was significant in identifying other differences in eye-movement metrics between the two groups, such as fixation count, duration, and regression patterns. All differences are consistent with what has been observed in other previous studies [38, 39, 40], in particular to the one obtained by Franzen et al. [13]. In all datasets, Dyslexic subjects have presented higher patterns in:
  - Number of regressive saccades

- Fixation duration
- Fixation entropy
- Number of fixation

While these features showed strong statistical differences, including all of them did not always improve the classification model's performance. This highlights that statistical significance does not always directly translate to predictive utility in the neural network model.

For example, while **fixation entropy** showed clear statistical differences between the two groups, it never improved classification accuracy and in some tests on the Dysfluent Readers of German Dataset, even led to lower performance.

Meanwhile, the **mean saccadic amplitude** often showed no statistically significant difference between the groups in most Hedges' g tests, with confidence intervals frequently including zero (e.g., ETDD70 Tasks 1 and 3, and generally across the Dysfluent Readers of German Dataset). Although it did show a significant difference in ETDD70 Task 2, its exclusion from feature sets often led to improved or more stable classification accuracy in many MLP tests. For instance, removing this feature was key to achieving the highest accuracy for ETDD70 Task 1 and Task 3. It also contributed to higher accuracy in specific Graz subgroup test.

We will discuss again about this problem in subsection 5.2.2, exploring in more detail the effects of different feature inputs on the neural network.

While Hedges' g can identify features with clear differences between groups, these do not always translate directly into improved predictive utility for machine learning models. Therefore, future work should explore more advanced or adaptive strategies for identifying optimal feature subsets, moving beyond traditional statistical significance to enhance model performance and generalizability.

Multi-Layer Perceptron (MLP) with Novel Similarity Measures:

- The integration of novel saccade-angle similarity measures as input features significantly enhanced the MLP model's performance across both long and short textual stimuli. In all the tests conducted in the previous chapter, we always obtained improvements when integrating this new measure, sometimes without an important impact and other times more significant.

- For the **ETDD70 Dataset** (long textual stimuli), the MLP model, augmented with the saccade-angle similarity measure, yielded notable improvements. For instance, accuracy for Task 1 increased from 91.43%± 5.35% to 92.86%± 4.52% . Similarly, Task 3 improved from 88.57%± 5.71% to 90.00%± 3.5% . Instead for Task 2, by using Cosine (with Threshold of 5°) and Peak 180 (with match), the model achieved a comparable average accuracy of 90.00% while reducing the standard deviation from ± 7.28% to ± 5.71%, indicating more stable performance. When using Peak

180, it improved the accuracy to 91.43% but increasing the standard deviation to ± 8.33%.

- For the **Dysfluent Readers of German Dataset** (containing short sentences), the MLP model consistently achieved the highest performance. Specifically, for both the Graz and Munich participants, the classification accuracy ranged between 97% and 100% under 5-fold cross-validation, and remained stably above 90% even when evaluated with 100 folds. In line with the observations from the previous dataset, incorporating similarity-based features further improved overall accuracy. For the Graz subjects, we also observed strong performance when using only angle similarity measure, achieving an accuracy of 91.00%±9.12, which is comparable to the accuracy obtained when integrating all feature types (91.89% ± 8.44). These results indicate that, for the Graz subjects, the MLP is capable of learning meaningful patterns to differentiate between Control and Dyslexic groups using solely the saccade angle similarity measure.

- The Peak 180 (with match) similarity measure, which is developed within the context of this thesis, proved highly effective. It achieved the highest accuracy with the Graz subjects and performed very good also in the second set of subjects from Munich. In multiple cases the results are very similar to the original version of Peak 180 but from the tests it seems that it can perform better on shorter stimuli.

- The Cosine method, although it is a similarity method not specifically designed for reading tasks, it achieved excellent results in almost all the stimuli proposed. Only with Task 1 of the ETDD70 Dataset it had lower performance compared to the other methods. However, it should be noted that this specific task differs slightly from all the others (as it consists of single words rather than complete sentences). We can therefore conclude that this method is the most consistent, with the only drawback being that it has a manual threshold which can be difficult to determine.

Gaussian Hidden Markov Models (GHMM): For short textual stimuli, the GHMM emerged as a compelling alternative to other neural networks. It demonstrated strong accuracy (over 5 different folds), ranging from 91.11% to 95.55% across the Dysfluent Readers of German Dataset, closely approaching the MLP's efficacy. A significant advantage of GHMM is its simplicity in input data (requiring primarily fixation sequences without extensive feature extraction), its interpretability, and explainability. The model's ability to achieve high accuracy using only the order and position of fixations is very promising.

In Figure 5.1 and Figure 5.2 we can see an example of how this model offers advantages in terms of the analysis and explanation behind its decisional system. We proposed the first four stimuli, containing pseudowords and nonwords. For all Graz and Munich subjects, we used the fitted model to show how we can visualize differences in the starting probability. In fact, we found that by comparing the initial probability states and the

transition matrix of the final model (after the training process), we can see that there are clear differences between the two groups, particularly among the subjects from Graz. Notably, in both the Graz and Munich subgroups, the Control subjects (Group 0) tend to have a higher probability of starting their scanpaths at the beginning of the text.

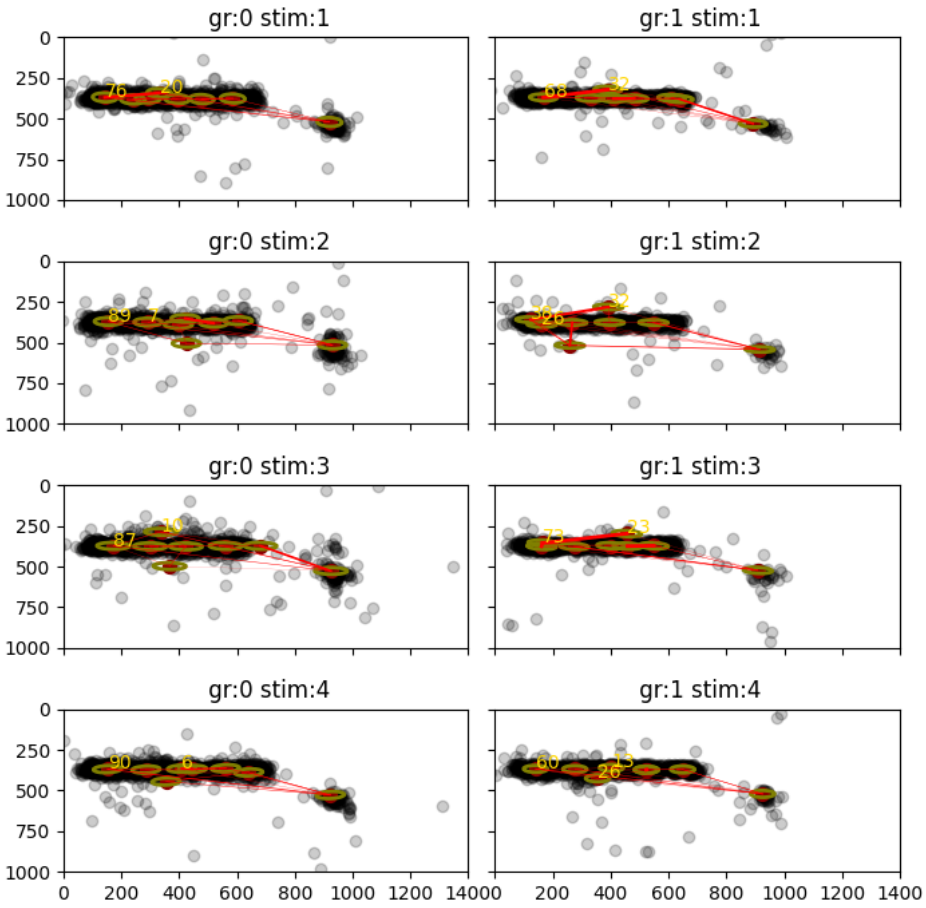


Figure 5.1.: GHMM comparison of Starting Probability between Control and Dyslexic group (Graz subjects)

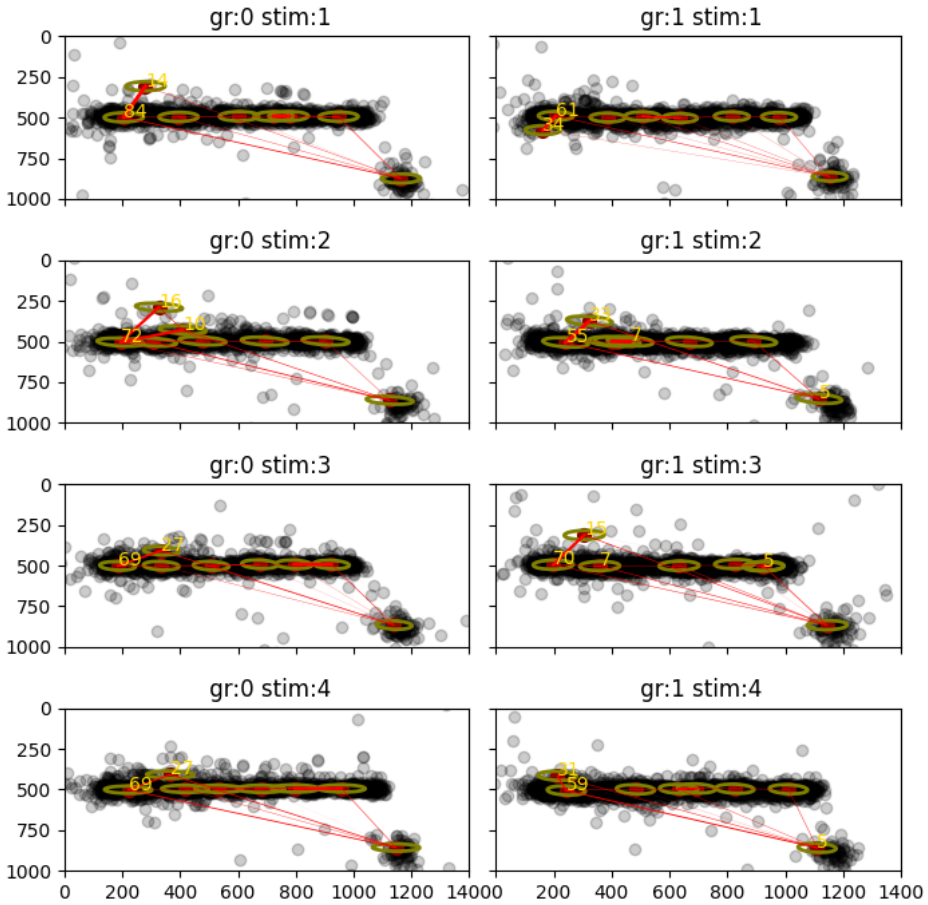


Figure 5.2.: GHMM comparison of Starting Probability between Control and Dyslexic group (Munich subjects)

We can validate and further investigate these differences by analysing the characteristics of the fitted models. In particular, we focus on the start probability and the transition matrix, calculating for the model of each stimulus the entropy of the start probability for the 8 states and the mean of the entropy of the 8x8 transition matrix. Then, with these values we use a related t-test to verify if the two samples have identical average values.

Following, we report the results obtained:

- **Graz subjects:**
  - Start Entropy: statistic=-2.710, p-value=0.0112
  - Transition Entropy: statistic=4.275, p-value=0.0002
- **Munich subjects:**
  - Start Entropy: statistic=-2.651, p-value=0.0129
  - Transition Entropy: statistic=1.122, p-value=0.2708

We can observe that the Graz subjects obtain in both tests a low p-value (the Start Entropy lower than 0.05 and Transition Entropy lower than 0.01). This means that we can reject the null hypothesis that they have identical average. The statistic results show that the starting entropy of the Control subjects is lower than the one of Dyslexic subjects, i.e. the starting entropy of Control subjects is more predictable. Meanwhile, the transition entropy shows an opposite pattern: the Control subjects present in average a higher entropy and this means that the transition matrix is less predictable than that of dyslexic subjects.

The Munich subjects present the same exact results for the Start Entropy as described before, with a p-value smaller than 0.05 and a negative statistic. Instead, the Transition Entropy presents a high p-value and therefore we cannot exclude the null hypothesis. The statistics shows a very small difference between the two means. For this reason, the transition matrices of the two groups have similar entropy.

In Figure 5.3 we can see the compression of the transition matrix between the two groups, calculated from the average of all transition matrices for each stimulus. We can notice that in both group the highest probability is located in the diagonal. This means that if we are in one state with high probability the next state will be again the same one. The explanation for this observation may be that a state contains multiple words, so in a typical scanpath, a subject makes multiple fixations in the same state before moving on to the next one. Then, regarding the statistical difference founded by the t-test, we have difficulties identifying in the image a specific aspect or pattern. However, analysing each transition matrix, we found that in some cases the standard deviation of the probability of each state seems to be much lower in the Control group then compared to the Dyslexic ones. In particular, in the first 10 stimuli (containing pseudowords and nonwords) we have that the standard deviation of the transition matrix of the Control subjects is in average 0.2466 ± 0.0056, lower compared to the Dyslexic average of 0.2678 ± 0.0118. This lower variation can explain the difference and the reason of why the statistic show higher entropy in the Control subjects.

These results allow for other interesting reflections and further evaluation on how stimuli are processed by the two groups. Future work could focus in inspecting and visualizing more in detail the results of each fitted model to investigate deeper differences between the group scanpaths.

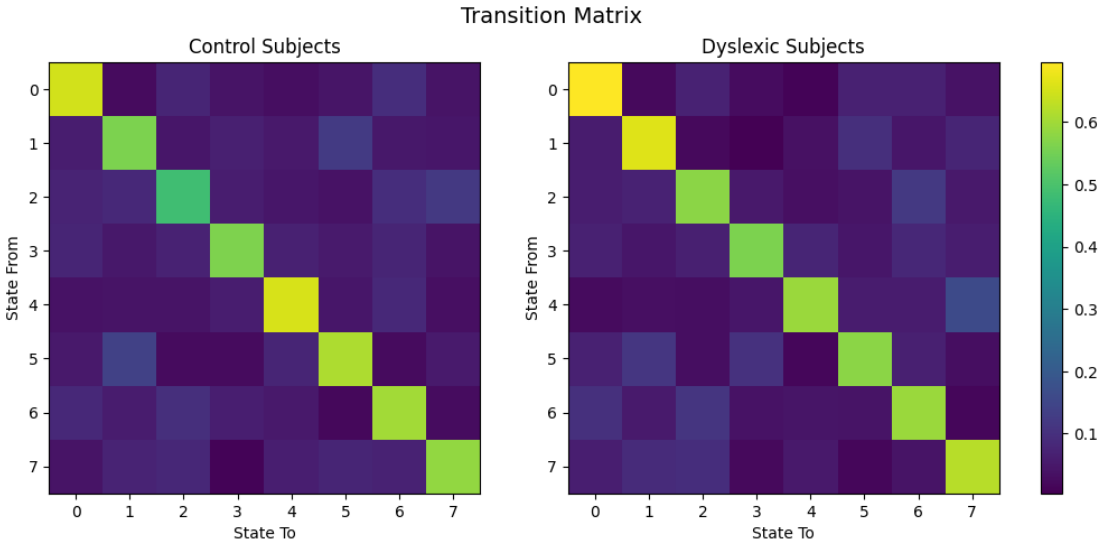


Figure 5.3.: GHMM Average Transition Matrix across the stimuli for Control and Dyslexic group (Graz subjects).

Limited Performance of Convolutional Neural Networks (CNNs): In contrast to MLP and GHMM, the CNN models consistently performed below 90% accuracy on the ETDD70 Dataset. While CNNs achieved better results on the Dysfluent Readers of German Dataset (up to 88.89% for Munich subjects), their performance generally remained below those obtained with MLP and GHMM. In subsection 5.2.1 we will explore more in detail the reason behind these results and see in which ways this approach could be improved.

Generalizability Across Diverse Datasets: The systematic evaluation across two distinct datasets (ETDD70 and Dysfluent Readers of German), differing in language, stimulus characteristics, and data collection environments, confirms the generalizability of the proposed framework. The consistent identification of distinct eye-movement patterns in individuals with learning disorders across these varied setups highlights the robustness of the findings.

5.2. Future Works and Improvements

5.2.1. CNN Problem and Improvement

The CNN model presented seemed like a very promising method, but unfortunately, in our study, we found results that were significantly inferior to the others. This performance gap is likely due to the total reading duration of the slowest readers, which constrains the final length of the input signal. In the original paper[8], the magnitude spectrum of time domain interpolated signal contains 1000 point. Instead in the ETDD70 Dataset, we have maximum 140, and in the Dysfluent Readers of German Dataset only 40, as also seen in Figure 3.9. This is a rather significant difference and probably the reason why, by replicating the same structure proposed in the paper, we were unable to achieve performance comparable to the other methods tested in this study.

We should also note that the initial performance on the Dysfluent Readers of German dataset was much lower than reported as the preliminary tests considered only single stimuli (using the same model architecture applied to the ETDD70 dataset). As mentioned above, individual stimuli contains little data and therefore accuracy was very low. For this reason, we modified the code (compared to that used for the ETDD70 Dataset), using all the stimuli from each subject as input, significantly improving the results obtained. This demonstrates once again how the model is extremely dependent on the amount of data provided and struggles to achieve good results with tasks that are too short.

In conclusion, this method requires further examination, in particular about the possibility of managing datasets with shorter reading tasks. Future studies could consider the possibility of redesigning the CNN architecture to make it more generalizable and test it again on the ETDD70 Dataset to compare the performance.

5.2.2. Feature Selection

In machine learning, feature selection is a well-known issue that has attracted increasing interest in research over the years[56, 57]. A subset of input variables (features) that optimizes predictive performance while minimizing dimensionality, redundancy, and noise is what it seeks to identify. Nevertheless, this task is far from simple: the ideal subset is frequently reliant on the data and model, and the number of potential feature subsets increases exponentially with dimensionality. Because of this, there is not a single feature selection technique that works best for everyone. Several strategies such as filter, wrapper, and embedding methods have inherent advantages and disadvantages. Feature selection is therefore a constant problem in both theory and real-world applications.

Our work aimed to compare features present in previous studies, with the integration of additional new features (including similarity measures). Therefore, our interest was not to study and eliminate specific features right away, but rather to evaluate and compare a series of well-defined subsets.

However, as said in the key findings, we noticed that there were interesting differences between the results obtained from statistical tests and the actual performance of some features.

To inspect more in details how inputs impact the classification result we used Integrated Gradients[58], a feature attribution method used to attribute a models prediction back to its input features. This is done to better understand why certain subsets performed better and which features were most important for the classification.

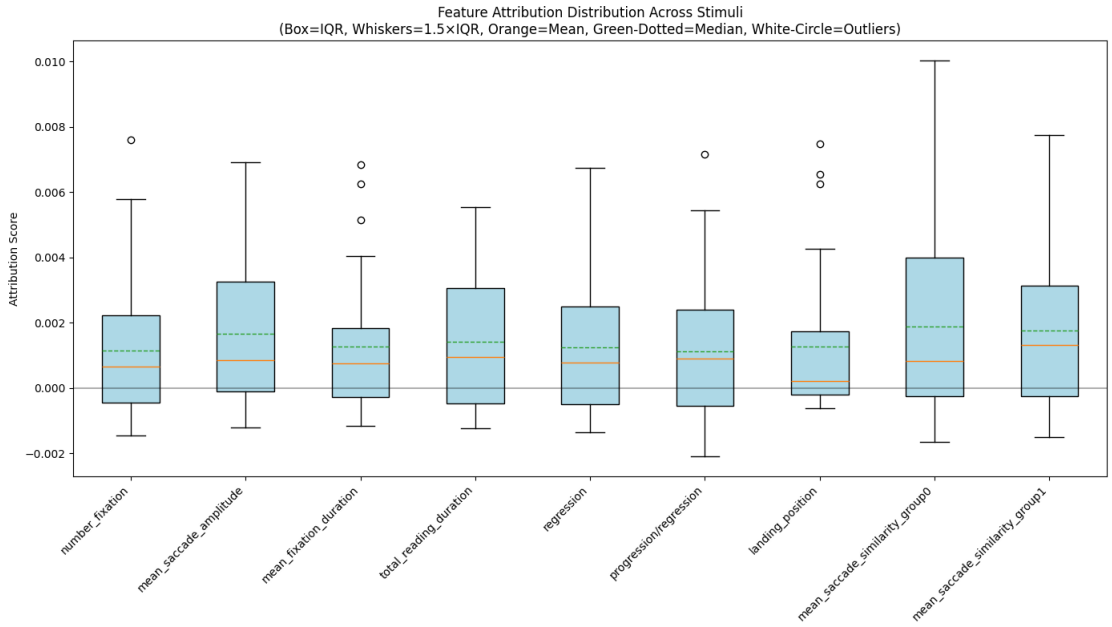
In Figure 5.4 we can see two results of the Integrated Gradients for the different inputs features. Since the input of the MLP for each subject is equal to the number of features multiplied by the number of stimuli, the Attribution Score of a single feature corresponds as the average across the stimuli used.

From Figure 5.4a we can see that on average, all features guide the model's decision towards the correct group. The component with the most negative attribution is the rate of progression/regression and the one with the mean closest to zero is the landing position of the first coordinate. Then, in Figure 5.4b, using a different subset (removing mean saccade amplitude) the model actually increase the performance, even if the feature removed was completely positive in the previous test. Moreover, while some attribution scores (e.g., saccade similarity, number of fixations) were similar between the two tests, others (e.g., landing position, progression/regression ratio) **changed significantly**.

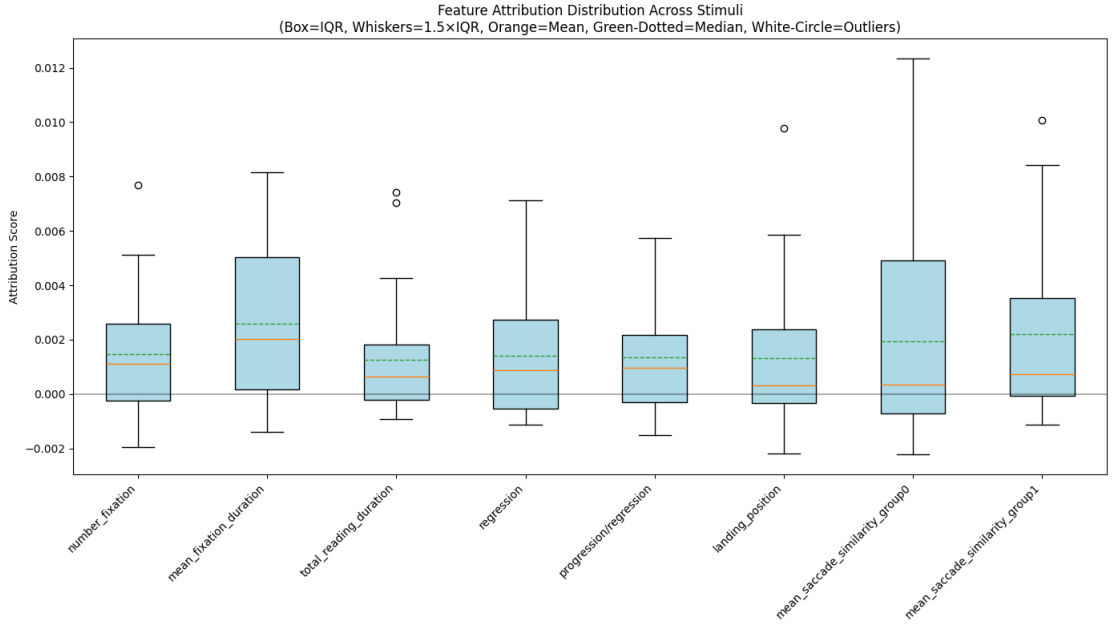
In a similar way for the Graz subjects, in Figure 5.5, we notice that total reading duration has a negative mean attribution score and is also very concentrated around 0. Then, in Figure 5.5b, we can notice how the attribution of this feature has completely changed.

In this study, we evaluated the performance of our neural network based on preliminary statistical tests and by creating groups of metrics that are comparable between the two datasets. However, we have not found any clear factors that allow us to determine a priori which are the best features to use as inputs for our model. Even the mean saccadic amplitude that from all the test was showing no statistical difference between the two groups, not in all the test has reported lower performance compared to the subsets without this feature.

Future work should consider studying in greater depth all the metrics that can be extracted from these eye-tracking systems and consider a more systematic way to evaluate and understand which features have the greatest impact on accuracy and which subsets are the most generalisable for this task.

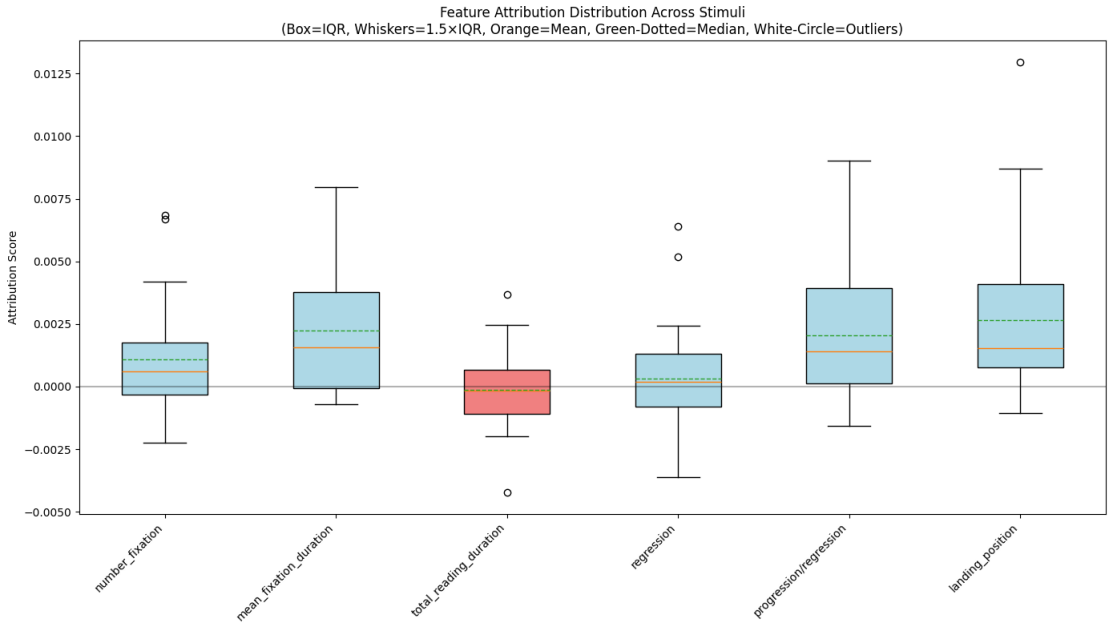


(a) Original features and Similarity Measure

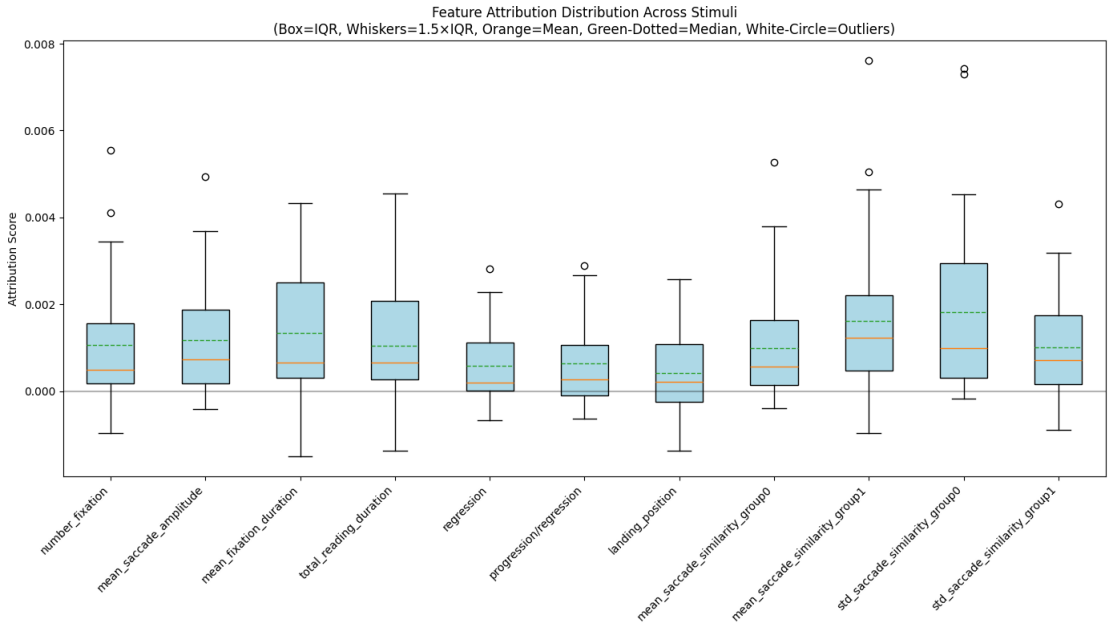


(b) Original features and Similarity Measure, **without** mean saccade amplitude

Figure 5.4.: Dysfluent Readers of German: Munich Subjects - Integrated Gradients of two subset of features



(a) Original features without mean saccade amplitude and similarity measure



(b) Original features and similarity measure

Figure 5.5.: Dysfluent Readers of German: Graz Subjects - Integrated Gradients of two subset of features

5.2.3. Similarity Measures Test & Develop

The Saccade-Angle Similarity measure proposed in this work is still a novel contribution and requires further validation to assess the actual performance of each method. In particular, the *Peak 180 with match sequence* method was specifically developed within the context of this thesis. Although the results obtained on both databases are very promising, it is necessary to test this approach on additional datasets to determine whether it outperforms the Cosine method or the original Peak 180. As discussed in Section 5.2.1, other datasets contain longer scanpath sequences [8], which could potentially influence the performance of the similarity measures.

This thesis has shown that similarity measures are valuable metrics for improving the classification of dyslexic subjects. Nevertheless, it is possible that alternative measures, or modifications of the ones presented by us, could lead to further improvements in accuracy.

In conclusion, although this study tested and compared several similarity measures, future research should extend the analysis by incorporating new types of stimuli, longer scanpaths, and deeper tests on the similarity measures.

5.2.4. Improving accessibility

Recent research has demonstrated the potential of using neural networks to accurately predict fixation points on a screen, achieving results comparable to those obtained with specialized, high-cost eye-tracking devices. This advancement access to eye-tracking technologies, facilitating not only everyday applications but also expanding possibilities for research experiments traditionally reliant on expensive hardware.

One of the pioneering works in this domain is the iTracker model introduced by Krafka et al.[31]. The GazeCapture project was among the first to demonstrate that, given a sufficiently large dataset, gaze prediction using a standard smartphone could achieve results comparable to those of eye-tracking equipment. After fine-tuning, the reported gaze prediction error was approximately 1.04 cm for smartphones and 1.70 cm for tablets. The disparity in accuracy is attributed to differences in the volume of data recorded with each device, with smartphones contributing 85% of the total dataset and tablets contributing the remaining 15%.

Building on this work, Strobl et al.[59] applied iTracker model in a real-world context, specifically for screening individuals with Autism Spectrum Disorder (ASD). Their findings indicated promising results, suggesting that this approach could be an effective tool for early detection and assessment in clinical settings. The study also mentioned the importance of further research in this field, advocating for the development of accessible, non-invasive eye-tracking solutions.

In recent years, significant improvements have been made to the original iTracker model, enhancing its accuracy and robustness. For instance, Valliappan et al.[3] achieved a substation in the gaze prediction error to just 0.42 cm ($\pm$0.03 cm), indicating the potential of these models for achieving near-professional levels of accuracy on consumer devices (as we can see in Figure 5.6).

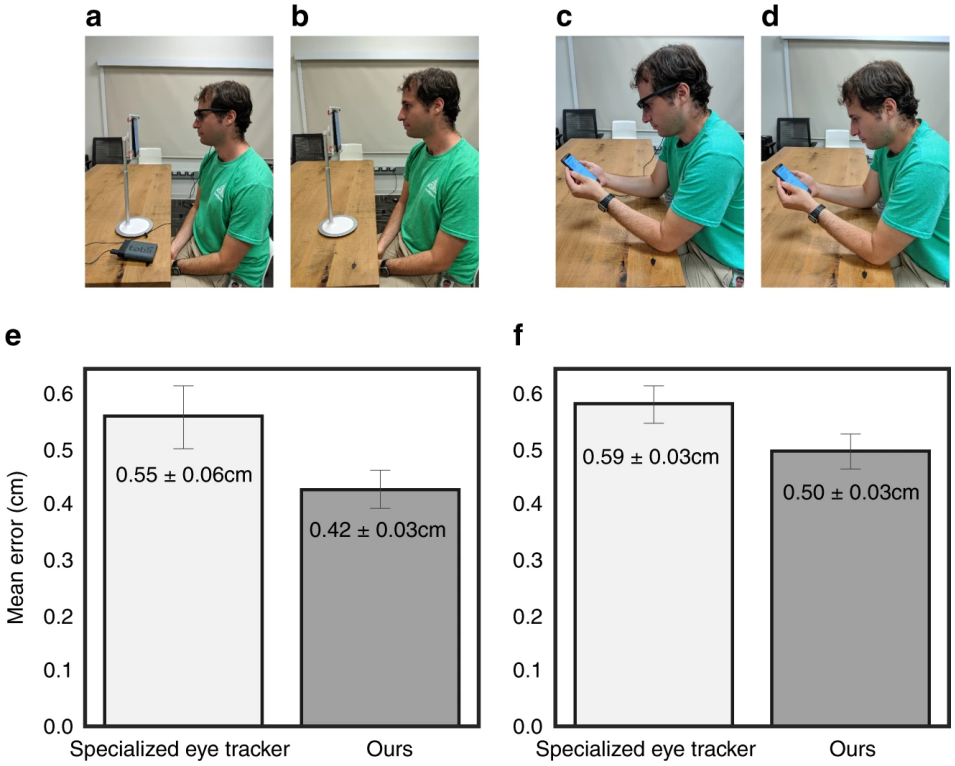


Figure 5.6.: Four experimental conditions showing no significant difference in accuracy across both settings, using Tobii glasses[2] (**a**) and with their smartphone eye tracker (**b**). In **c**, **d** we have similar to the above, but participant holds the phone in the hand. (the result for the test of **a** and **b** are shown in **e** and of **c** and **d** are represented in **f**) - Image from the study of Valliappan et al.[3].

When comparing this system with specialized hardware, the most important limitation to consider is the difference in sampling frequency. Specialized eye-tracking equipment, although expensive, can achieve sampling frequencies of 500 or 1000 Hz (which translates into recording a fixation correspondingly every 2 or 1 milliseconds). However, for the model in analysis we have a sampling frequency of 60 Hz (16.67 milliseconds). This could significantly impact the results of the classification model, especially considering that generally the data available are performed in the laboratory with specialized hardware. Nevertheless, some studies have demonstrated that data recorded at lower sampling rates, between 30 and 60 Hz, can still achieve high accuracy levels[45, 48].

Since these visual patterns are recognizable at lower sampling frequencies, it allows for the development of a cost-effective gaze-tracking test that utilizes widely accessible devices such as laptops, smartphones, and tablets. Furthermore, it will enable the collection of fixation data from a broad range of subjects in a non-invasive way.

Dysfluent Readers of German - Downsampled

From the Dysfluent Readers of German Dataset we have the raw coordinates of each scanpath. Therefore, instead of using the fixation recorded by the eye-tracker software, we can downsample the data and recalculate the new fixation.

Firstly, it is important to note that this type of manipulation is generally not considered as a good practice in computer science. In this particular case, even if we reduce the amount of data fixation and use data from a more accurate sampling device, we cannot be certain that the same data quality would have been achieved if the recording had originally been made at a lower sampling frequency.

The purpose of this test is to evaluate how much does the classification accuracy decreases when the amount of available data is significantly reduced. If the test reveals a drastic drop in accuracy, it is unlikely that good results can be obtained from data recorded at such low frequencies. On the other hand, if accuracy remains acceptable, this approach may open the door to important developments and will need to be examined more thoroughly in the future.

For the downsampling process, we needed to adjust our data from a sampling frequency of 500 Hz to 60 Hz. This would ideally require selecting one point every $8.\overline{3}$ samples. To simplify the procedure, we instead retained every 8th sample, resulting in an effective frequency of 62.5 Hz.

Once we have obtained our new starting dataset, we can proceed to calculate the fixation points. There are several algorithms that allow fixations to be identified. The most important and widely used are divided into three main categories:

- **Velocity-based**
  - I-VT
  - I-HMM
- **Density-based**
  - I-DT
  - I-MST
- **Area-based**
  - I-AOI

From these representative algorithms, Dispersion-Threshold Identification (I-DT) offers reliable and accurate fixation identification by integrating sequential data to facilitate interpretation[60].

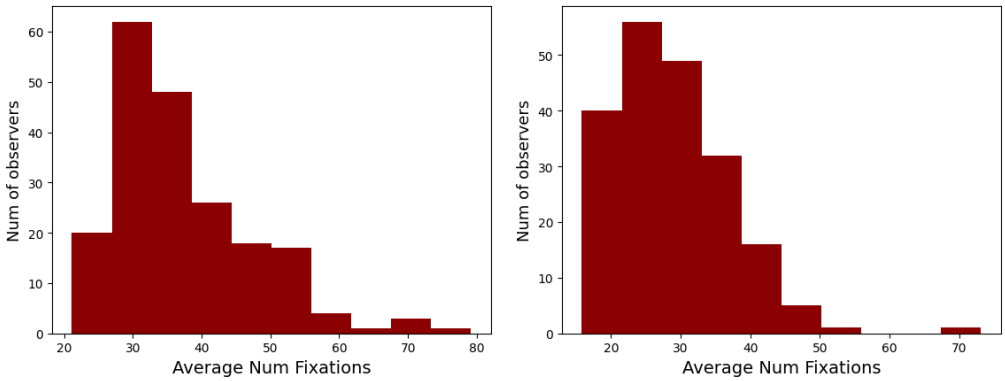
Therefore, we decide to use the I-DT algorithm, using a dispersion threshold of 50 pixels (around 2ř of visual angle) and duration threshold of 80 ms. Starting from the original data for each subject and for each stimulus, we reduced the sampling rate and then applied the following algorithm.

Algorithm 2 I-DT Fixation Detection

Require: Gaze data D, dispersion threshold θ_d, minimum duration θ_t
Ensure: List of fixations F

$F \leftarrow$ empty list
$i \leftarrow 0$
while $i < |D|$ **do**
    $j \leftarrow i + 1$
    while $j < |D|$ **and** $D[j].t - D[i].t < \theta_t$ **do**
        $j \leftarrow j + 1$
    end while
    if $j \geq |D|$ **then**
        break
    end if
    $W \leftarrow D[i : j]$ ▷ Initial window covering θ_t
    Compute $dispersion(W)$
    if $dispersion(W) \leq \theta_d$ **then**
        while $j < |D|$ **do**
            $W' \leftarrow D[i : j + 1]$
            Compute $dispersion(W')$
            if $dispersion(W') > \theta_d$ **then**
                break
            end if
            $j \leftarrow j + 1$
        end while
        $W \leftarrow D[i : j]$
        $duration \leftarrow W[-1].t - W[0].t$
        if $duration \geq \theta_t$ **then**
            Compute centroid (x_{fix}, y_{fix}) of W
            Append fixation $(W[0].t, W[-1].t, duration, x_{fix}, y_{fix})$ to F
        end if
        $i \leftarrow j$ ▷ Remove window points
    else
        $i \leftarrow i + 1$ ▷ Slide window by one point
    end if
end while
return F

After applying the algorithm, we obtained a reduction of approximately 21% of fixation points. In Figure 5.7 we can observe that the global average of number of fixations shifted from 30/40 to 20/30, with also an important decreasing of subjects with fixations higher than 40.



(a) Original Data (b) After Downsampling

Figure 5.7.: Dysfluent Readers of German - Comparison of Number of subjects by Average Number of Fixations, before and after downsampling

In Table 5.1 we report the results obtained on the subsets of Graz and Munich by applying the MLP model over 100 folds (only best the best results) in comparison with the best one obtained with the original datasets. In the appendix Table B.7 and Table B.8, is possible to see the complete results for each subset.

Subjects	**Original**	**Downsampled**
Graz	91.89 ± 8.59	91.56 ± 10.08
Munich	91.44 ± 9.54	93.00 ± 8.41

Table 5.1.: Dysfluent Readers of German - Percentage of accuracy and standard deviation between Original and Downsampled dataset (Graz and Munich subjects)

We can see that the final result is rather unexpected. Given this significant reduction in the quality and quantity of data, we would have assumed a decrease in the accuracy of the model. Instead, we observe similar performance and even an increase for the Munich subjects.

It is not straightforward to attribute the observed improvement to a single factor. However, the choice of algorithm for computing fixations appears to play a decisive role. A notable difference is that the original dataset included several extremely short fixations (approximately 10-20 ms), whereas in our case, even without applying any threshold, the shortest fixation duration was 80 ms. This suggests that fixation time parameters are an important factor that may have contributed to an increase in fixation

duration. Furthermore, from the complete results table, we observe that the original features, such as the number of fixations and fixation duration, gained greater relevance compared to features like mean saccade amplitude and similarity measures. Consequently, recalculating fixations with this specific configuration may have increased the importance of certain features, thereby enabling a clearer separation between classes. The chosen parameters remain consistent with the characteristics of the data. Adopting more restrictive configurations (e.g., by increasing both thresholds) could have resulted in an excessive reduction in the number of detected fixations, thereby limiting the usefulness of the data for subsequent analysis.

However, this remains a preliminary simulation. The key takeaway from this test is that a significant reduction in data did not lead to a drop in performance and this indicates that there are considerable differences between dyslexic and Control subjects in the data.

Overall, these findings are encouraging, as they indicate that diagnostic or research applications relying on eye-tracking may remain feasible even when constrained to lower-cost hardware. More importantly, the acceptable trade-off between accuracy and accessibility paves the way for scalable, non-invasive screening tools that can be deployed outside of laboratory settings. Future studies should validate these results with datasets originally recorded at lower sampling frequencies and consider multiple fixation identification algorithms to better approximate real-world performance and ensure robustness across diverse populations and device types. There might also be some fixation identification algorithms that are particularly suited to Dyslexic subjects.

5.2.5. Dataset Limitations

As outlined in the introduction, a major challenge of this work was the availability of suitable datasets for analysis and testing. The datasets employed in this study are valuable due to their diversity, particularly with respect to stimulus type, number of stimuli, and participant coverage. However, both datasets are relatively new and have only been used in a limited number of studies. Consequently, there is a lack of prior work against which the present results can be directly compared.

To strengthen the generalizability of the findings, it would be necessary to replicate the experiments proposed in this thesis using well-established datasets that have been extensively analysed in the literature, such as the Swedish dataset employed by Nilsson Benfatto et al. [12]. Moreover, as highlighted in the previous section, other datasets recorded with lower sampling frequencies already exist [45, 48]. Utilizing such datasets would enable a more robust evaluation and facilitate a deeper comparison of the present results with established findings.

5.3. Conclusion

This thesis demonstrates that automated dyslexia detection through scanpath analysis is not only feasible but achievable with high accuracy across diverse experimental conditions. By systematically evaluating multiple machine learning approaches (MLP, CNN, and GHMM) on two distinct datasets with different languages and stimulus characteristics, we have established a foundation for generalizable diagnostic frameworks that can adapt to varying experimental setups.

The introduction of novel Saccade-Angle Similarity measures represents a significant methodological contribution. These measures consistently enhanced classification performance across all tested configurations, demonstrating that capturing the geometric properties of eye movement patterns provides discriminative power beyond traditional metrics. The MLP architecture emerged as the most robust and versatile solution, achieving accuracies that exceed 90% across both datasets when augmented with similarity-based features. However, the GHMM also demonstrates strong performance on short stimuli, achieving comparable accuracy with minimal feature engineering and offering superior interpretability. These two methods present interesting prospective for new research and for future practical application.

Beyond the technical findings, this work illuminates a critical path toward dyslexia screening. The performance demonstrated with downsampled data, combined with recent advances in smartphone-based eye-tracking, suggests that effective screening tools could rely on more affordable equipment. This accessibility is crucial: early detection of learning disorders remains a privilege largely confined to well-resourced educational systems. Developing screening tools that can operate on commodity devices like smartphones, tablets, or standard webcams could transform early intervention by making initial assessment accessible to families and educators worldwide.

However, making possible this vision requires acknowledging current limitations. The feature selection challenge remains unresolved. While statistical tests identify group differences, it do not always translate into improved predictive utility. Future work must develop more principled approaches to feature engineering and selection.

The field would benefit from three parallel research directions. First, establishing standardized benchmark datasets with diverse languages, age groups, and recording conditions would enable more rigorous cross-study comparisons. Second, validating these approaches with data originally collected at lower sampling frequencies would confirm their viability for real-world deployment. Third, developing multilingual screening applications to integrate these methods could provide immediate practical value while generating the diverse datasets needed for further refinement.

In conclusion, this thesis shows that the technological and methodological barriers to automated dyslexia screening can be overcome. What remains is the translational challenge: moving from laboratory validation to practical tools that can serve as a first-line screening mechanism in schools, homes, and community settings. Future studies should focus on systems that improve the overall robustness while considering accessibility, making these methods available to those who need them most.

Bibliography

[1] Nicol Dostalova, Roman Svaricek, Jan Sedmidubsky, Wolf Culemann, Cenek Sasinka, Pavel Zezula, and Jiri Cenek. Etdd70: Eye-tracking dyslexia dataset, 2024.

[2] Anneli Olsen. The tobii i-vt fixation filter. *Tobii Technology*, 21:4–19, 2012.

[3] Nachiappan Valliappan, Na Dai, Ethan Steinberg, Junfeng He, Kantwon Rogers, Venky Ramachandran, Pingmei Xu, Mina Shojaeizadeh, Li Guo, Kai Kohlhoff, and Vidhya Navalpakkam. Accelerating eye movement research via accurate and affordable smartphone eye tracking. *Nature Communications*, 11(1), September 2020.

[4] Valentina Levantini, Pietro Muratori, Emanuela Inguaggiato, Gabriele Masi, Annarita Milone, Elena Valente, Alessandro Tonacci, and Lucia Billeci. Eyes are the window to the mind: Eye-tracking technology as a novel approach to study clinical characteristics of adhd. *Psychiatry Research*, 290:113135, August 2020.

[5] Nicholas J. Wade and Benjamin W. Tatler. *Origins and applications of eye movement research*. Oxford University Press, August 2011.

[6] Robert J.K. Jacob and Keith S. Karn. *Eye Tracking in Human-Computer Interaction and Usability Research*, page 573605. Elsevier, 2003.

[7] Dong Yun Lee, Yunmi Shin, Rae Woong Park, Sun-Mi Cho, Sora Han, Changsoon Yoon, Jaheui Choo, Joo Min Shim, Kahee Kim, Sang-Won Jeon, and Seong-Ju Kim. Use of eye tracking to improve the identification of attention-deficit/hyperactivity disorder in children. *Scientific Reports*, 13(1), September 2023.

[8] Boris Neruil, Jaroslav Polec, Juraj Skunda, and Juraj Kaur. Eye tracking based dyslexia detection using a holistic approach. *Scientific Reports*, 11(1), August 2021.

[9] Rosa Sancarlo, Zoya Dare, Jozsef Arato, and Raphael Rosenberg. Does pictorial composition guide the eye? investigating four centuries of last supper pictures. *Journal of Eye Movement Research*, 13(2), August 2020.

[10] Antoine Coutrot, Janet H. Hsiao, and Antoni B. Chan. Scanpath modeling and classification with hidden markov models. *Behavior Research Methods*, 50(1):362379, April 2017.

[11] Jan Sedmidubsky, Nicol Dostalova, Roman Svaricek, and Wolf Culemann. *ETDD70: Eye-Tracking Dataset for Classification of Dyslexia Using AI-Based Methods*, page 3448. Springer Nature Switzerland, October 2024.

[12] Mattias Nilsson Benfatto, Gustaf Öqvist Seimyr, Jan Ygge, Tony Pansell, Agneta Rydberg, and Christer Jacobson. Screening for dyslexia using eye tracking during reading. *PLOS ONE*, 11(12):e0165508, December 2016.

[13] Léon Franzen, Zoey Stark, and Aaron P. Johnson. Individuals with dyslexia use a different visual sampling strategy to read text. *Scientific Reports*, 11(1), March 2021.

[14] Zicai Liu, Zhen Yang, Yueming Gu, Huiyu Liu, and Pu Wang. The effectiveness of eye tracking in the diagnosis of cognitive disorders: A systematic review and meta-analysis. *PLOS ONE*, 16(7):e0254059, July 2021.

[15] B. Cassin, S. Solomon, and M.L. Rubin. *Dictionary of Eye Terminology*. Triad Publishing Company, 1990.

[16] Dale Purves, George J. Augustine, David Fitzpatrick, Lawrence C. Katz, Anthony-Samuel LaMantia, James O. McNamara, and S. Mark Williams. *Neuroscience*. Sinauer Associates, 2nd edition, 2001.

[17] H. B. Barlow. Eye movements during fixation. *The Journal of Physiology*, 116(3):290306, March 1952.

[18] Richard J. Krauzlis, Laurent Goffart, and Ziad M. Hafed. Neuronal control of fixation and fixational eye movements. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1718):20160205, February 2017.

[19] D A Robinson. The mechanics of human smooth pursuit eye movement. *The Journal of Physiology*, 180(3):569591, October 1965.

[20] Karl R. Gegenfurtner. The interaction between vision and eye movements. *Perception*, 45(12):13331357, September 2016.

[21] L. E. Mays. Neural control of vergence eye movements: convergence and divergence neurons in midbrain. *Journal of Neurophysiology*, 51(5):10911108, May 1984.

[22] Dora E. Angelaki and Kathleen E. Cullen. Vestibular system: The many facets of a multimodal sense. *Annual Review of Neuroscience*, 31(1):125150, July 2008.

[23] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5)*. American Psychiatric Publishing, 2013.

[24] James E. Harrison, Stefanie Weber, Robert Jakob, and Christopher G. Chute. Icd-11: an international classification of diseases for the twenty-first century. *BMC Medical Informatics and Decision Making*, 21(S6), November 2021.

[25] INSERM Collective Expertise Centre et al. Dyslexia dysorthography dyscalculia. 2007.

[26] Sally E. Shaywitz, Jonathan E. Shaywitz, and Bennett A. Shaywitz. Dyslexia in the 21st century. *Current Opinion in Psychiatry*, 34(2):8086, December 2020.

[27] Linda Zuppardo, Francisca Serrano, Concetta Pirrone, and Antonio Rodriguez-Fuentes. More than words: Anxiety, self-esteem, and behavioral problems in children and adolescents with dyslexia. *Learning Disability Quarterly*, 46(2):7791, August 2021.

[28] Neil Alexander-Passe. How dyslexic teenagers cope: an investigation of self-esteem, coping and depression. *Dyslexia*, 12(4):256275, 2006.

[29] Jihwan Park, Youngsun Kong, and Yunyoung Nam. A low-cost video-oculography system for vestibular function testing. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, page 40784081. IEEE, July 2017.

[30] Pouya Barahim Bastani, Shervin Badihian, Vidith Phillips, Hector Rieiro, Jorge Otero-Millan, Nathan Farrell, Max Parker, David Newman-Toker, and Ali Saber Tehrani. Smartphones versus goggles for video-oculography: current status and future direction. *Research in Vestibular Science*, 23(3):6370, September 2024.

[31] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[32] Melanie Gangl, Kristina Moll, Manon W. Jones, Chiara Banfi, Gerd Schulte-Körne, and Karin Landerl. Lexical reading in dysfluent readers of german. *Scientific Studies of Reading*, 22(1):2440, July 2017.

[33] Norberto Pereira, Maria Armanda Costa, and Manuela Guerreiro. Effects of word length and word frequency among dyslexic, adhd-i and typical readers. *Journal of Eye Movement Research*, 15(1), June 2022.

[34] M. L. Wijnants, F. Hasselman, R. F. A. Cox, A. M. T. Bosman, and G. Van Orden. An interaction-dominant perspective on reading fluency and dyslexia. *Annals of Dyslexia*, 62(2):100119, March 2012.

[35] Pavel Zikl, Iva Koek Bartoová, Kateina Josefová Víková, Klára Havlíková, Alice Kuírková, Jolana Navrátilová, and Barbora Zetková. The possibilities of ict use for compensation of difficulties with reading in pupils with dyslexia. *Procedia - Social and Behavioral Sciences*, 176:915922, February 2015.

[36] Heiner Böttger, Julia Dose, and Tanja Müller. Contrast and font affect reading speeds of adolescents with and without a need for language-based learning support. *Training Language and Culture*, 1(4):3955, December 2017.

[37] Jessie Rae Ramsey. *The effects of the font dyslexie on oral reading fluency skills in students grades 8 through 12 with and without reading disabilities.* Western Carolina University, 2014.

[38] Otto Loberg, Jarkko Hautala, Jarmo A. Hämäläinen, and Paavo H.T. Leppänen. Influence of reading skill and word length on fixation-related brain activity in school-aged children during natural reading. *Vision Research*, 165:109122, December 2019.

[39] Christoforos Christoforou, Argyro Fella, Paavo H.T. Leppänen, George K. Georgiou, and Timothy C. Papadopoulos. Fixation-related potentials in naming speed: A combined eeg and eye-tracking study on children with dyslexia. *Clinical Neurophysiology*, 132(11):27982807, November 2021.

[40] Guangyao Zhang, Binke Yuan, Huimin Hua, Ya Lou, Nan Lin, and Xingshan Li. Individual differences in first-pass fixation duration in reading are related to resting-state functional connectivity. *Brain and Language*, 213:104893, February 2021.

[41] Ivan Vajs, Vanja Kovic, Tamara Papic, Andrej M. Savic, and Milica M. Jankovic. Dyslexia detection in children using eye tracking data based on vgg16 network. In *2022 30th European Signal Processing Conference (EUSIPCO)*, page 16011605. IEEE, August 2022.

[42] Luz Rello and Miguel Ballesteros. Detecting readers with dyslexia using machine learning with eye tracking measures. In *Proceedings of the 12th International Web for All Conference*, W4A 15, page 18. ACM, May 2015.

[43] Peter Raatikainen, Jarkko Hautala, Otto Loberg, Tommi Kärkkäinen, Paavo Leppänen, and Paavo Nieminen. Detection of developmental dyslexia with machine learning using eye movement data. *Array*, 12:100087, December 2021.

[44] Alae Eddine El Hmimdi, Lindsey M Ward, Themis Palpanas, and Zoï Kapoula. 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, October 2021.

[45] Ivan Vajs, Tamara Papi, Vanja Kovi, Andrej M. Savi, and Milica M. Jankovi. Accessible dyslexia detection with real-time reading feedback through robust interpretable eye-tracking features. *Brain Sciences*, 13(3):405, February 2023.

[46] Soroosh Shalileh, Dmitry Ignatov, Anastasiya Lopukhina, and Olga Dragoy. Identifying dyslexia in school pupils from eye movement and demographic data using artificial intelligence. *PLOS ONE*, 18(11):e0292047, November 2023.

[47] Zbigniew Gomolka, Ewa Zeslawska, Barbara Czuba, and Yuriy Kondratenko. Diagnosing dyslexia in early school-aged children using the lstm network and eye tracking technology. *Applied Sciences*, 14(17):8004, September 2024.

[48] Roman Svaricek, Nicol Dostalova, Jan Sedmidubsky, and Andrej Cernek. Insight: Combining fixation visualisations and residual neural networks for dyslexia classification from eyetracking data. *Dyslexia*, 31(1), January 2025.

[49] John P. Rack, Margaret J. Snowling, and Richard K. Olson. The nonword reading deficit in developmental dyslexia: A review. *Reading Research Quarterly*, 27(1):28, 1992.

[50] Beth A. OBrien, Guy C. Van Orden, and Bruce F. Pennington. Do dyslexics misread a rows for a rose? *Reading and Writing*, 26(3):381402, April 2012.

[51] Roy S. Hessels, Diederick C. Niehorster, Chantal Kemner, and Ignace T. C. Hooge. Noise-robust fixation detection in eye movement data: Identification by two-means clustering (i2mc). *Behavior Research Methods*, 49(5):18021823, October 2016.

[52] Melanie Gangl, Kristina Moll, Chiara Banfi, Stefan Huber, Gerd Schulte-Körne, and Karin Landerl. Reading strategies of good and poor readers of german with different spelling abilities. *Journal of Experimental Child Psychology*, 174:150169, October 2018.

[53] G. S. WATSON. Goodness-of-fit tests on a circle. *Biometrika*, 48(12):109114, 1961.

[54] Lukas Landler, Graeme D. Ruxton, and E. Pascal Malkemper. Advice on comparing two independent samples of circular data in biology. *Scientific Reports*, 11(1), October 2021.

[55] Hayward J. Godwin, Charlotte E. Lee, and Denis Drieghe. A multiverse analysis of cleaning and analyzing procedures of eye movement data during reading. *Behavior Research Methods*, 57(6), May 2025.

[56] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[57] Dipti Theng and Kishor K. Bhoyar. Feature selection techniques for machine learning: a survey of more than two decades of research. *Knowledge and Information Systems*, 66(3):15751637, December 2023.

[58] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 33193328. JMLR.org, 2017.

[59] Maximilian A. R. Strobl, Florian Lipsmeier, Liliana R. Demenescu, Christian Gossens, Michael Lindemann, and Maarten De Vos. Look me in the eye: evaluating the accuracy of smartphone-based eye tracking for potential application in autism spectrum disorder research. *BioMedical Engineering OnLine*, 18(1), May 2019.

[60] Dario D. Salvucci and Joseph H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the Eye Tracking Research and Applications Symposium*, ETRA 00, page 7178. ACM Press, 2000.

Acronyms

ADHD Attention-Deficit/Hyperactivity Disorder. 6

ASC ASCII file. 21

ASD Autism Spectrum Disorder. 6

CNN Convolutional Neural Network. 1

DCT Discrete Cosine Transform. 57

DFT Discrete Fourier Transform. 58

DSM-5 Diagnostic and Statistical Manual of Mental Disorders, 5th Edition. 4

EDF EyeLink Data File. 21

HMM Hidden Markov Model. 1

ICD-11 International Classification of Diseases, 11th Edition. 4

LDs Learning Disorders. 4

MLP Multi-Layer Perceptron. 1

ROI Region of Interest. 127

SGD Stochastic Gradient Descent. 57

SLD Specific Learning Disorder. 2

Glossary

Entropy The entropy of a random variable is a measure of the average level of uncertainty or lack of information associated with its potential states or possible outcomes. In the context of fixation points, entropy quantifies the unpredictability of where a viewer looks, capturing the randomness or concentration of their visual attention across a scene. Higher entropy corresponds to more unpredictable and dispersed fixations, whereas lower entropy indicates more predictable and focused viewing behavior. 10

Likelihood The probability of the observed data given the model parameters. 50

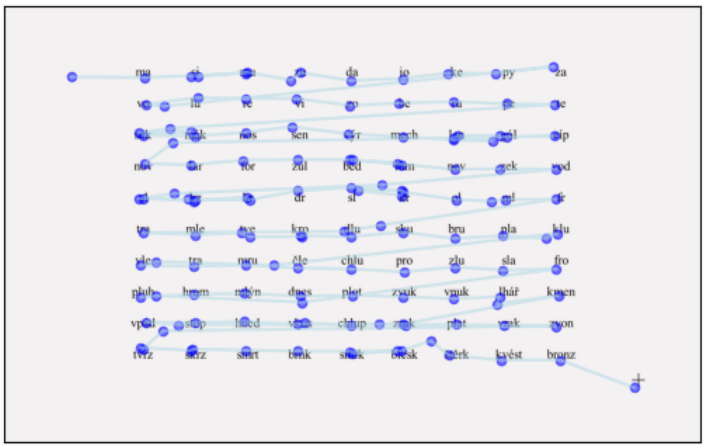
Region of Interest A Region of Interest (ROI) is a portion of an image that has been manually selected by specialists and is suitable for the desired analysis. 12

Scanpath A person's eye movements over a period of time, commonly represented as a series of alternating fixations and saccades. 1

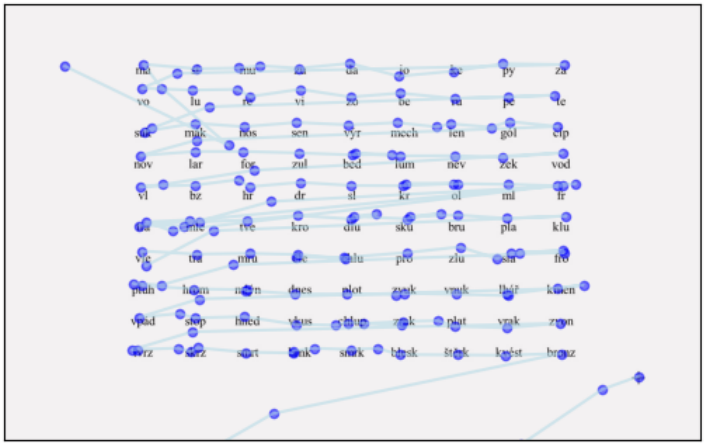
A. ETDD70 Dataset - Additional Resources

A.1. Stimuli Visualization

(a) Task 1: Syllables

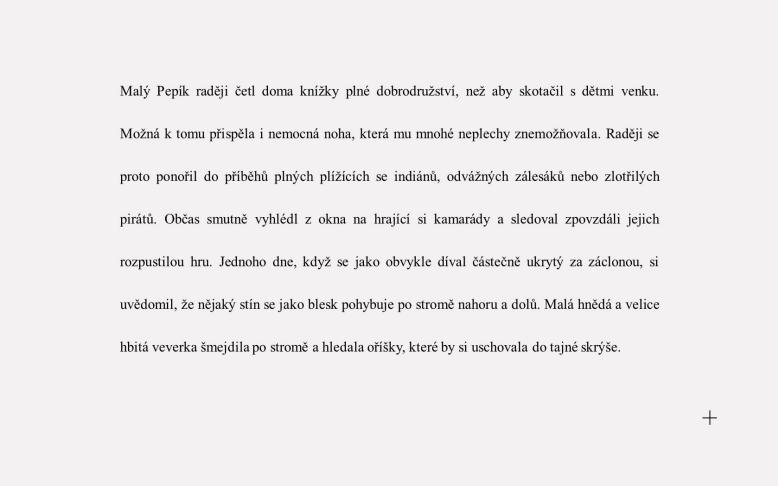


(b) Scanpath for the stimulus of a Non-Dyslexic Subject

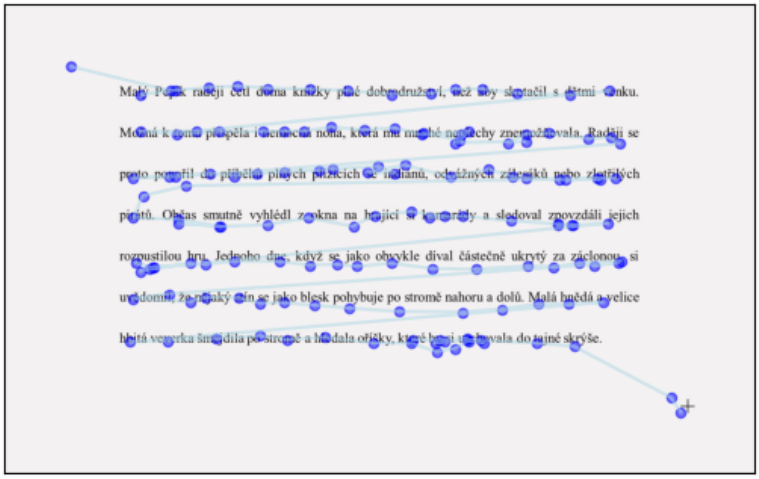


(c) Scanpath for the stimulus of a Dyslexic Subject

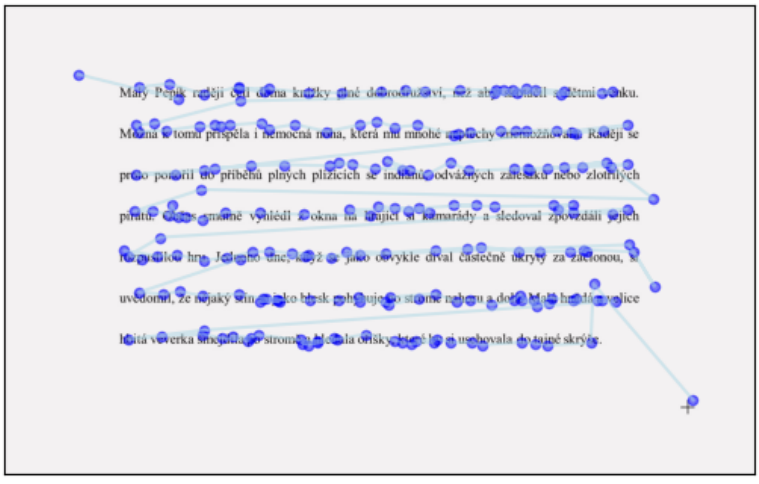
Figure A.1.: ETDD70 - Task 1 stimulus with example scanpath for the two groups



(a) Task 2: Meaningful Text

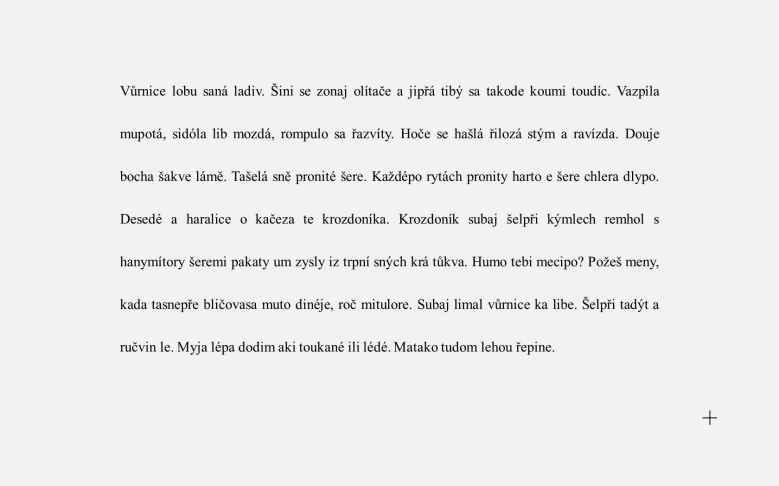


(b) Scanpath for the stimulus of a Non-Dyslexic Subject

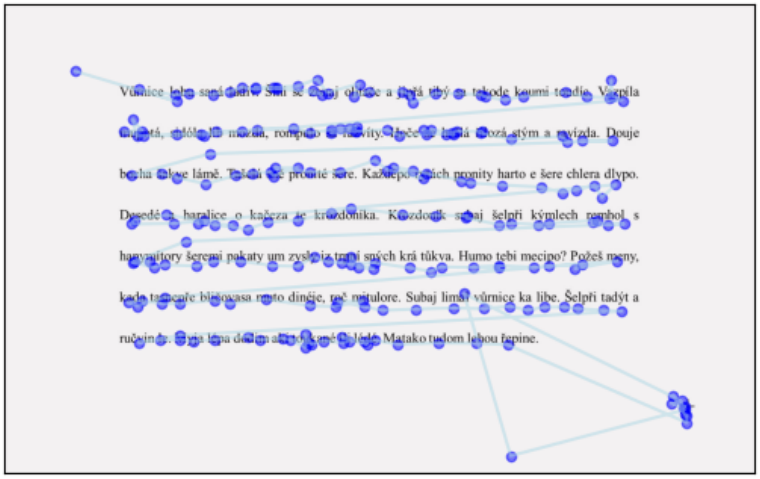


(c) Scanpath for the stimulus of a Dyslexic Subject

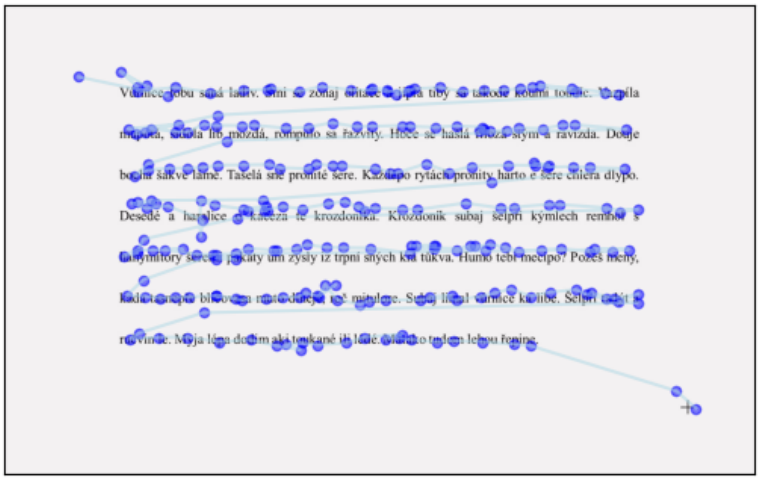
Figure A.2.: ETDD70 - Task 2 stimulus with example scanpath for the two groups



(a) Task 3: Pseudo-Text



(b) Scanpath for the stimulus of a Non-Dyslexic Subject



(c) Scanpath for the stimulus of a Dyslexic Subject

Figure A.3.: ETDD70 - Task 3 stimulus with example scanpath for the two groups

A.2. MANOVA Result Tables

Multivariate Linear Model Statistics						
Effect	**Test**	**Value**	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Intercept	Wilks' Lambda	0.8353	2.0000	12148.0000	1197.6290	< 0.0001
	Pillai's Trace	0.1647	2.0000	12148.0000	1197.6290	< 0.0001
	Hotelling-Lawley Trace	0.1972	2.0000	12148.0000	1197.6290	< 0.0001
	Roy's Greatest Root	0.1972	2.0000	12148.0000	1197.6290	< 0.0001
Label	Wilks' Lambda	0.9975	2.0000	12148.0000	15.3789	< 0.0001
	Pillai's Trace	0.0025	2.0000	12148.0000	15.3789	< 0.0001
	Hotelling-Lawley Trace	0.0025	2.0000	12148.0000	15.3789	< 0.0001
	Roy's Greatest Root	0.0025	2.0000	12148.0000	15.3789	< 0.0001

Table A.1.: ETDD70: Task 1 - MANOVA Intercept and Label Results

Multivariate Linear Model Statistics						
Effect	**Test**	**Value**	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Intercept	Wilks' Lambda	0.8477	2.0000	15888.0000	1426.7754	< 0.0001
	Pillai's Trace	0.1523	2.0000	15888.0000	1426.7754	< 0.0001
	Hotelling-Lawley Trace	0.1796	2.0000	15888.0000	1426.7754	< 0.0001
	Roy's Greatest Root	0.1796	2.0000	15888.0000	1426.7754	< 0.0001
Label	Wilks' Lambda	0.9951	2.0000	15888.0000	39.3709	< 0.0001
	Pillai's Trace	0.0049	2.0000	15888.0000	39.3709	< 0.0001
	Hotelling-Lawley Trace	0.0050	2.0000	15888.0000	39.3709	< 0.0001
	Roy's Greatest Root	0.0050	2.0000	15888.0000	39.3709	< 0.0001

Table A.2.: ETDD70: Task 2 - MANOVA Intercept and Label Results

Multivariate Linear Model Statistics						
Effect	**Test**	**Value**	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Intercept	Wilks' Lambda	0.8305	2.0000	20387.0000	2080.8257	< 0.0001
	Pillai's Trace	0.1695	2.0000	20387.0000	2080.8257	< 0.0001
	Hotelling-Lawley Trace	0.2041	2.0000	20387.0000	2080.8257	< 0.0001
	Roy's Greatest Root	0.2041	2.0000	20387.0000	2080.8257	< 0.0001
Label	Wilks' Lambda	0.9945	2.0000	20387.0000	56.6821	< 0.0001
	Pillai's Trace	0.0055	2.0000	20387.0000	56.6821	< 0.0001
	Hotelling-Lawley Trace	0.0056	2.0000	20387.0000	56.6821	< 0.0001
	Roy's Greatest Root	0.0056	2.0000	20387.0000	56.6821	< 0.0001

Table A.3.: ETDD70: Task 3 - MANOVA Intercept and Label Results

B. Dysfluent Readers of German - Additional Resources

B.1. Stimuli Details

Stimulus	Text	Words	Pseudohom.	Nonwords	Short	Long	Total
8881	laud breit Note lilla	0	0	0	0	0	4
8882	Blau balt oft Beine	0	0	0	0	0	4
8883	Fus faul dann Tase	0	0	0	0	0	4
8884	Glas fier trei lang	1	0	0	1	0	4
1	ihm Safft Farbe erfunden Köhnig Wassa Blätter fahren Tühr Wolke	4	4	0	4	4	10
2	Ball Hut Somma Zättel Bäume FüSSe nimals schlächt Haare üben	4	4	0	4	4	10
3	Glück arm täuer Blitz Mädchen schtreiten falsch Glas stoltz leer	5	3	0	5	3	10
4	treu schparen Hand Zoh froh Butter schprechen ferlaufen Schule Tiger	4	4	0	3	5	10
5	Name sellbst unmöklich Familie Bild kranck Film HoSSe Ferien dick	4	3	0	4	3	10
6	Kopf Arzt Kint Zimma Lehrerin Wonung Katze Stain Leiter Fuchs	5	4	0	5	4	10
7	Berg Munt Tire gestern wonen Gras Welder Geheimnis Rahd Buch	3	5	0	4	4	10
8	süSS teglich fragen Lufft Wint Schnee wünnschen Hals Opst FuSS	3	5	0	4	4	10
9	gelb Fuchs wahrm Mutter schwimmen Fater Brot Fenster zehlen ganz	5	3	0	4	4	10
10	rot Jaar wehlen fehlen nur runt Kleid Schwantz Mittwoch Stück	4	4	0	4	4	10
8885	Rall kelo Kirm Reile fode kurl mell Sero	0	0	0	0	0	8
11	pelo groh Sanilia Hond engunsen fulsch Kleud Fäne Schree rele	0	0	8	4	4	10
12	Mane Folm Laale Müdchel Lut Fehrasin Sutten fahmen Kals Ropp	0	0	8	4	4	10
13	Brog Delien Littmoch Glap Meiker Glitz Flätten Alzt Fonstel Lotz	0	0	7	3	4	10
14	Aro lur Buld tesbern Kapfe Scheke dahnen Beleimnes Furde glab	0	0	8	4	4	10
15	sira Lutten Käune Fechs schwillen Krot Gres flaben orm Muro	0	0	8	5	3	10
16	gose Hind miefals Summel Rak Pälger ventauben Naft Zommel Def	0	0	8	3	5	10
17	Orel sund urdögnich Muft Schwunz Staum gohlen stelz Kofe Rebo	0	0	8	5	3	10
18	kamme zühnen Nind schlocht Kiene Zot tünschan sähren fägnich surd	0	0	8	3	5	10
19	soku Kaber Mahr Abst Zutten streufen beuel solbst Nind Mot	0	0	8	5	3	10
20	fott Tar Kohnang krink spracker Käbin spolen Munk Dassel Bloma	0	0	7	4	3	10
8886	offt Sonne bunt Node	0	0	0	0	0	4
21	Sohn täglich Butta sparen Rad ahrm Geheimniss warm Graas Blume	4	4	0	4	4	10
22	Apfel König Hud Wohnung faren Hose wählen Artzt Mitwoch bald	4	4	0	4	4	10
23	Fluss Hant Jahr krank Glaas schlecht Fehrien Lererin Wälder gut	4	4	0	4	4	10
24	gern Kind Füse verlaufen Beume erfunnden Wasser Stein Katse Land	4	4	0	5	3	10
25	Post frahgen Tür Plitz stolz fro Wind felen wohnen Korb	4	4	0	5	3	10
26	drei sprechen Schuhle teuer Medchen Hare zählen Schneh Luft jung	4	4	0	3	5	10
27	Keks Fux Zettel gesstern nuhr Tiere Mund fallsch streiten Hexe	4	4	0	4	4	10
28	Kanne wünschen Klaid Obst Zimmer Mutta Vater Fänster Vilm Ohr	4	4	0	4	4	10
29	Wald rund Broht schwimen selbst Fabe niemals Zoo Famielie Zahl	4	4	0	4	4	10
30	nett Billd Sommer Laiter Saft Schwanz Halls Bletter unmöglich April	4	4	0	3	5	10

Table B.1.: Dysfluent Readers of German - Stimuli Analysis

(a) Typical Reader Subject

(b) Subject with Isolated Spelling Deficit

(c) Subject with Combined Reading and Spelling Deficit

(d) Subject with Isolated Reading Deficit

Figure B.1.: Dysfluent Readers of German: Stimulus 1 - Example scanpath for the four group

B.2. MANOVA Result Tables

Multivariate Linear Model Statistics						
Effect	**Test**	**Value**	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Intercept	Wilks' Lambda	0.8582	2.0000	257860.0000	21300.2710	< 0.0001
	Pillai's Trace	0.1418	2.0000	257860.0000	21300.2710	< 0.0001
	Hotelling-Lawley Trace	0.1652	2.0000	257860.0000	21300.2710	< 0.0001
	Roy's Greatest Root	0.1652	2.0000	257860.0000	21300.2710	< 0.0001
Label	Wilks' Lambda	0.9982	2.0000	257860.0000	238.3810	< 0.0001
	Pillai's Trace	0.0018	2.0000	257860.0000	238.3810	< 0.0001
	Hotelling-Lawley Trace	0.0018	2.0000	257860.0000	238.3810	< 0.0001
	Roy's Greatest Root	0.0018	2.0000	257860.0000	238.3810	< 0.0001

Table B.2.: Dysfluent Readers of German - MANOVA results across all the stimuli

Multivariate Linear Model Statistics						
Effect	**Test**	**Value**	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Intercept	Wilks' Lambda	0.9361	2.0000	22359.0000	763.2098	< 0.0001
	Pillai's Trace	0.0639	2.0000	22359.0000	763.2098	< 0.0001
	Hotelling-Lawley Trace	0.0683	2.0000	22359.0000	763.2098	< 0.0001
	Roy's Greatest Root	0.0683	2.0000	22359.0000	763.2098	< 0.0001
Label	Wilks' Lambda	0.9998	2.0000	22359.0000	2.1854	0.1125
	Pillai's Trace	0.0002	2.0000	22359.0000	2.1854	0.1125
	Hotelling-Lawley Trace	0.0002	2.0000	22359.0000	2.1854	0.1125
	Roy's Greatest Root	0.0002	2.0000	22359.0000	2.1854	0.1125

Table B.3.: Dysfluent Readers of German - MANOVA results (real words only)

Multivariate Linear Model Statistics						
Effect	**Test**	**Value**	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Intercept	Wilks' Lambda	0.8488	2.0000	235498.0000	20977.7709	< 0.0001
	Pillai's Trace	0.1512	2.0000	235498.0000	20977.7709	< 0.0001
	Hotelling-Lawley Trace	0.1782	2.0000	235498.0000	20977.7709	< 0.0001
	Roy's Greatest Root	0.1782	2.0000	235498.0000	20977.7709	< 0.0001
Label	Wilks' Lambda	0.9978	2.0000	235498.0000	261.9562	< 0.0001
	Pillai's Trace	0.0022	2.0000	235498.0000	261.9562	< 0.0001
	Hotelling-Lawley Trace	0.0022	2.0000	235498.0000	261.9562	< 0.0001
	Roy's Greatest Root	0.0022	2.0000	235498.0000	261.9562	< 0.0001

Table B.4.: Dysfluent Readers of German - MANOVA results (only pseudowords and nonwords)

Multivariate Linear Model Statistics						
Effect	**Test**	**Value**	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Intercept	Wilks' Lambda	0.8482	2.0000	145490.0000	13015.0351	< 0.0001
	Pillai's Trace	0.1518	2.0000	145490.0000	13015.0351	< 0.0001
	Hotelling-Lawley Trace	0.1789	2.0000	145490.0000	13015.0351	< 0.0001
	Roy's Greatest Root	0.1789	2.0000	145490.0000	13015.0351	< 0.0001
Label	Wilks' Lambda	0.9980	2.0000	145490.0000	148.3100	< 0.0001
	Pillai's Trace	0.0020	2.0000	145490.0000	148.3100	< 0.0001
	Hotelling-Lawley Trace	0.0020	2.0000	145490.0000	148.3100	< 0.0001
	Roy's Greatest Root	0.0020	2.0000	145490.0000	148.3100	< 0.0001

Table B.5.: Dysfluent Readers of German - MANOVA results (excluding stimuli containing only nonwords and real words)

Multivariate Linear Model Statistics						
Effect	**Test**	**Value**	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Intercept	Wilks' Lambda	0.8496	2.0000	90005.0000	7969.2938	< 0.0001
	Pillai's Trace	0.1504	2.0000	90005.0000	7969.2938	< 0.0001
	HotellingLawley Trace	0.1771	2.0000	90005.0000	7969.2938	< 0.0001
	Roy's Greatest Root	0.1771	2.0000	90005.0000	7969.2938	< 0.0001
Label	Wilks' Lambda	0.9974	2.0000	90005.0000	115.6169	< 0.0001
	Pillai's Trace	0.0026	2.0000	90005.0000	115.6169	< 0.0001
	HotellingLawley Trace	0.0026	2.0000	90005.0000	115.6169	< 0.0001
	Roy's Greatest Root	0.0026	2.0000	90005.0000	115.6169	< 0.0001

Table B.6.: Dysfluent Readers of German - MANOVA results (only nonwords)

B.3. Supplementary Classification Result Tables

- **Subset 1**: Original and Similarity (Average & Standard Deviation)
- **Subset 2**: Original and Similarity (Average & Standard Deviation, no mean saccade amplitude)
- **Subset 3**: Original and Average Similarity
- **Subset 4**: Original and Average Similarity (no mean saccade amplitude)

Features	**Momentum = 0.8**	**Momentum = 0.5**
Original	89.67 ± 8.64	90.44 ± 8.89
Original (without mean saccade amplitude)	88.44 ± 9.42	89.78 ± 10.14
Original and Additional	89.78 ± 9.90	90.56 ± 9.35
Subset 1	90.22 ± 9.97	**91.56 ± 10.08**
Subset 2	89.44 ± 9.21	90.11 ± 8.73
Subset 3	88.67 ± 10.18	89.22 ± 9.74
Subset 4	89.88 ± 10.43	90.78 ± 9.94
All	90.88 ± 8.37	91.11 ± 8.60

Table B.7.: Dysfluent Readers of German: Graz Subjects (with sampling rate of 60 Hz) - MLP classification accuracy and standard deviations, tested on 100 folds using Peak 180 with match similarity. (best result in **bold**)

Features	**Momentum = 0.8**	**Momentum = 0.5**
Original	91.89 ± 9.27	91.67 ± 9.08
Original (without mean saccade amplitude)	**93.00 ± 8.41**	92.78 ± 8.66
Original and Additional	92.89 ± 8.54	91.89 ± 8.87
Subset 1	91.78 ± 9.11	91.33 ± 9.38
Subset 2	91.56 ± 9.82	91.33 ± 9.76
Subset 3	92.00 ± 9.70	92.44 ± 9.28
Subset 4	91.66 ± 9.21	91.56 ± 9.18
All	91.78 ± 9.11	91.78 ± 9.38

Table B.8.: Dysfluent Readers of German: Munich Subjects (with sampling rate of 60 Hz) - MLP classification accuracy and standard deviations, tested on 100 folds using Cosine with threshold 5° similarity . (best result in **bold**)