UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Neuroscience Supervisors:

Dr. Samy Rima

Prof. Michael Schmid

Machine Learning Supervisor:

Prof. Bastian Grossenbacher-Rieck

No Reading Required: Using Gamified Eye-Tracking to Predict RAN Scores and Screen for Developmental Dyslexia

Master Thesis for the academic degree

Specialized Master of Science in Digital Neuroscience

at the Faculty of Science and Medicine, University of Fribourg

Olivia Lecomte

20-219-002

Fribourg, 4 September 2025

Acknowledgements

I would like to sincerely thank my supervisor, Dr. Samy Rima, for his invaluable support over the past year and a half. His advice to discover our passions early and pursue them wholeheartedly has stayed with me, and I feel incredibly fortunate that it was his guidance that inspired me to explore my love for diagnostics through this project. I would also like to thank both Samy and Prof. Michael Schmid for their trust in me and for allowing me the freedom to pursue my curiosities. The independence I was granted throughout this thesis was deeply appreciated.

I'm grateful to Niloufar for her help with the object detection models and pre-processing scripts, and to everyone who contributed to the data collection, without whom this thesis would not have been possible. Special thanks to Prof. Bastian Grossenbacher-Rieck for his continued guidance and support in helping me further my understanding of machine learning and deep learning. I feel very fortunate to have been part of such responsive and supportive teams.

I'd also like to thank the SSUDK for supporting my transition from dance to neuroscience. Without their help over the past five years, I would not have been able to make such a meaningful and life-changing career shift.

On a personal note, I would like to thank my family. Especially my mum, Julie, my partner, Christian, and my sister, Kai, for their unwavering support. From dance to psychology to digital neuroscience, they've been with me every step of the way. Thank you especially to Christian for cooking the best meals and for our unmatched banter.

To my wonderful university friends Claire and Jessi: thank you for the train chats, cinnamon bun tastings, Zurich fun days, and years of companionship through our Bachelor's and Master's degrees. I'll cherish our library days forever.

To my dance community in Lucerne, thank you for bringing light and joy into my life. I am especially grateful to my friends, students, and my exceptional colleagues at Tanzhaus Luzern, particularly Kiara and Debora, who took on extra responsibilities to support me while I completed my studies. Your professionalism and teamwork made it possible for me to pursue my education—something I'll always be grateful for.

Contents

Abstract

Eye-tracking combined with machine learning shows strong potential as a diagnostic tool for developmental dyslexia, particularly when decoupled from reading-based assessments. However, few studies have explored the use of gamified, non-reading tasks suitable for pre-readers. This pilot study investigates whether eye-tracking metrics recorded during gameplay (*Fruit Ninja*) can predict Rapid Automatized Naming scores and classify individuals into "good" versus "poor" performance groups. A total of 23 eye-tracking features were extracted from 57 participants (children and adults) and analysed using correlation, regression, and classification models. Correlation and regression analyses identified several significant predictors of Rapid Automatized Naming performance, particularly features related to blink behaviours, saccade length, and gaze complexity, while classification models yielded the highest accuracy with blink and fixation-based features. These findings suggest that oculomotor behaviour during a non-reading task contains information relevant to reading ability and may reflect magnocellular processing differences. This approach shows promise for early, literacy-independent risk detection, but further research with larger and clinically diverse samples is needed to validate and extend these results.

Keywords: developmental dyslexia, eye-tracking, machine learning, non-reading assessment, rapid automatized naming

Introduction

Developmental dyslexia (DD) is the most common neurodevelopmental disorder, affecting approximately one in five people, with around 7% experiencing it severely enough to meet diagnostic criteria (*Dyslexia FAQ*, n.d.; Leung et al., 2022; Tooze, 2022; Vajs et al., 2022). The International Dyslexia Association (IDA) defines DD as a neurobiological learning disability primarily rooted in phonological deficits, leading to difficulties with word recognition, spelling and decoding (IDA, 2002). However, recent research suggests that this definition no longer captures the full complexity of the disorder. A growing body of evidence suggests that DD involves a broader set of impairments, particularly in visual processing, echoing theories that were once mainstream nearly a century ago (Stein, 2018).

The impact of DD extends far beyond reading, affecting academic achievement, emotional well-being, and overall quality of life (Gomolka et al., 2024; Leung et al., 2022; Vajs et al., 2022). Children with DD often face stigmatisation despite average or above-average intelligence (Leung et al., 2022). They are six times more likely to drop out of school, and learning disorders such as DD or ADHD are estimated to underlie roughly 65% of academic failures (Bartolomé et al., 2012; Daniel et al., 2006; Leung et al., 2022). Alarmingly, individuals with DD experience higher rates of mental health issues, such as depression and have a suicide rate three times that of their neurotypical peers (Leung et al., 2022; Mammarella et al., 2016). One study reported significant spelling and handwriting impairments consistent with DD in 89% of teen suicide notes, underscoring the importance of early DD intervention (McBride & Siegel, 1997).

Currently, diagnosis typically occurs only after reading difficulties become evident, leading to delayed interventions (Vajs et al., 2022). Historically, DD assessments relied primarily on a combination of family history and the observation of a discrepancy between a child's poor reading abilities and their comparatively high oral and non-verbal skills (Stein, 2022). However, with the rise of the phonological theory, which focuses on phoneme segmentation and awareness, diagnostic assessments have shifted towards literacy-dependent measures (Stein, 2018). This reliance on reading-based tests presents a paradox: they demand proficiency in the very skill impaired by DD, thereby placing undue stress on individuals already struggling with the task (Gomolka et al., 2024; Stein, 2022). In addition to this conceptual contradiction, DD diagnostic procedures are time-consuming, expensive, and resource-intensive, often requiring the involvement of licensed professionals and spanning

over a year (Asvestopoulou et al., 2019; Gomolka et al., 2024; Stein, 2022, 2023). Hurdles which disproportionately affect families from lower socioeconomic backgrounds.

These limitations have catalysed a movement toward early, objective, and scalable screening methods that do not rely on reading. Emerging research supporting the magnocellular theory of DD, suggests that deficits in visual motion perception and eye movement control precede reading failure and may reflect a deeper neurocognitive disruption underlying DD (Mascheretti et al., 2018). Magnocellular (M) cells, comprising roughly 10% of the retinal ganglion cells, are specialised for rapid motion detection, visual attention, and eye movement coordination (Stein, 2019). Once these signals reach the primary visual cortex (V1), their signal is propagated along the dorsal stream, a pathway frequently implicated in DD (Laycock et al., 2008; Stein, 2022).

Although not all individuals with DD exhibit a measurable magnocellular deficit, and some with such deficits learn to read, this does not rule out the visual system's importance. (Stein, 2022). M-cells can only be reliably differentiated in the early visual tract (retina, LGN, and L4 of V1), making direct assessment a challenge (Stein, 2019). Nevertheless, a study found that M-cell sensitivity, as measured through motion tasks like the Random Dot Kinematogram (RDK), has shown strong predictive value for reading ability in all neurotypical and neurodivergent participants (Talcott et al., 2002). The reduced performance on the RDK observed in individuals with DD is likely attributable to decreased sensitivity in the V5/MT area, which is responsible for motion detection and receives its primary input from the magnocellular dorsal stream (Leung et al., 2022; Livingstone et al., 1991; Mascheretti et al., 2018; Stein & Walsh, 1997; Werth, 2021b). However, due to high inter-individual variability, motion sensitivity alone remains a suboptimal diagnostic marker. Nonetheless, Mascheretti et al. (2018) argue that visual motion processing meets the criteria for an endophenotype of DD, offering predictive value even before formal reading education. This strengthens the magnocellular theory, which is further supported by histological (Giraldo-Chica et al., 2015; Rao et al., 1997; Stein, 2022; Werth, 2021) and neuroimaging evidence (Demb et al., 1997, 1998; Eden et al., 1996; Huettig et al., 2018; Jednoróg et al., 2011; Klistorner et al., 1997; Leppänen et al., 2010; Livingstone et al., 1991; Premeti et al., 2022; Schulte-Körne et al., 1998). Crucially, several intervention studies have shown that training rapid visual processing through tasks such as action video games, significantly improves magnocellular function, and as a result, reading ability (Bavelier & Davidson, 2013; Peters et al., 2019; Werth, 2021a).

Magnocellular dysfunction also affects oculomotor control, particularly affecting the two basic eye movements, fixations (stable maintenance of gaze on a point), and saccades (rapid shifts between fixation points) (Asvestopoulou et al., 2019; Raatikainen et al., 2021; Stein, 2019). Individuals with DD often display unstable fixations and erratic saccades even during non-reading tasks (Asvestopoulou et al., 2019; Premeti et al., 2022; Stein, 2022). These disruptions are consistent with anecdotal reports of "moving letters", possibly due to the M-cells' impaired ability to suppress unwanted saccades (Stein, 2022; Werth, 2021b).

However, the impact of M-cell dysfunction extends across a wide range of oculomotor behaviours. Individuals with DD often struggle to maintain smooth pursuit of moving objects, instead showing saccadic intrusions and other atypical pursuit patterns (Eden et al., 1994; Stein, 2019). Broader abnormalities in gaze control have also been reported, including excessive blinking (Tooze, 2022) and poor binocular coordination (Premeti et al., 2022). These impairments are consistent with the consequences of magnocellular dysfunction and together contribute to inefficient visual scanning.

Several eye-tracking metrics have been developed to quantify these behaviours. Readers with DD typically exhibit longer fixations, greater fixation counts, and reduced fixation stability (Asvestopoulou et al., 2019; Pavlidis, 1981; Premeti et al., 2022). More recently developed metrics such as the Fixation Intersection Coefficient (FIC) and Fixation Fractal Dimension (FFD) have been proposed to capture the spatial complexity and irregularity of gaze behaviour (Vajs et al., 2022). These features quantify self-intersections and variability within fixation paths, and have been shown to increase in individuals with DD, reflecting a more chaotic and less efficient visual scanning strategy (Vajs et al., 2022). Saccade profiles are similarly affected—readers with DD often produce more saccades of smaller amplitude than neurotypicals (Asvestopoulou et al., 2019; Premeti et al., 2022). These finding are widely supported by the literature and have been reported for both reading and non-reading tasks, suggesting they are not merely a consequence of poor reading ability but may in fact contribute to it (Asvestopoulou et al., 2019; Premeti et al., 2022; Raatikainen et al., 2021; Stein, 2022).

In addition to fixations and saccades, blink behaviour has also been implicated. Although the evidence on blink duration remains mixed—with some studies reporting longer durations in adults with DD but no significant differences in children—higher blink frequency has been associated with increased cognitive load or task difficulty (Tooze, 2022). This

suggests that blinking, like other oculomotor behaviours, may serve as a biomarker of visual or attentional strain in DD. Beyond standard blink metrics, the present study introduces a novel measure, the blink ratio, defined as the proportion of task time spent blinking, to test whether overall blink activity provides additional predictive value. Together, these oculomotor abnormalities form a distinct behavioural profile of DD that can be detected independently of literacy. The affordability and portability of eye-tracking technology make it a promising tool for large-scale, non-invasive screening in young children.

For nearly a decade, researchers have developed machine learning (ML) models capable of classifying DD using eye-tracking data. Most studies report accuracies ranging from 80% to 97.7% using features extracted during reading tasks and classified using ML models such as Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbours (KNN), Random Forest (RF), and increasingly deep learning (Gomolka et al., 2024; Usman et al., 2021; Vajs et al., 2023). For example, *DysLexML* achieved 97% accuracy using only four basic eye movement features during a reading task, even under noisy conditions (Asvestopoulou et al., 2019).

Only a handful of studies have explored non-reading tasks, and most report comparatively lower classification accuracy. A notable exception is a 2024 study by Gomolka et al., which achieved 97.7% accuracy using a Long Short-Term Memory (LSTM) model on spatio-temporal eye-tracking data from a short visual task, the Benton Visual Retention Test (BVRT). The task lasted just ten minutes and involved no reading, significantly reducing stress and testing time. This demonstrates the feasibility of screening for DD by targeting underlying cognitive differences rather than reading performance.

However, identifying pre-reading children with DD remains a challenge, as formal diagnosis is still dependent on reading failure. One solution is to use Rapid Automatized Naming (RAN) as a proxy. In particular, the object-naming subtask requires children to quickly name common objects and has been shown to reliably predict DD across languages and orthographies (Ji & Bi, 2020; Lervåg & Hulme, 2009; Pan et al., 2013). RAN performance has also been shown to correlate strongly with age in children, with older children typically naming more quickly and accurately (Lervåg & Hulme, 2009; Mascheretti et al., 2018; Pan et al., 2013). This makes age control essential in studies involving developmental samples (Mascheretti et al., 2018). Although it does not meet the criteria for an endophenotype, RAN remains one of the strongest behavioural predictors of DD (Mascheretti et al., 2018).

Numerous articles have demonstrated that individuals with DD have slowed RAN compared to neurotypicals (Araújo et al., 2015). This is likely due to the processes underlying successful RAN completion, including rapid temporal processing and the sequencing of auditory and visual cues, two functions that are often linked to the magnocellular system (Stein, 2019, 2023).

This thesis contributes to the understanding of DD by examining the relationship between eye-tracking features from a non-reading task and RAN scores—widely recognised as strong predictors of reading ability. Specifically, the pilot study investigates the correlation between a range of conventional and novel eye-tracking features and age-controlled RAN scores. To identify the most predictive features, the study applies Random Forest Importance Ranking combined with Shapley Additive Explanations (SHAP), providing interpretable insights into the model's decision-making process. In addition, individuals are classified based on a median split of their age-controlled RAN scores, demonstrating the potential of game-based, non-reading eye-tracking tasks to predict reading-related difficulties without directly relying on reading performance. By leveraging machine learning to predict RAN scores from eye-tracking data collected during a visual game, this study aims to identify accessible, non-reading-based indicators of dyslexia risk for future research and screening applications.

Methods

Data Collection and Materials

Participants

A total of 61 participants, including children and adults, were recruited to take part in this study. Data from three participants were excluded from analysis due to errors encountered during data collection or feature extraction. Of the 57 remaining participants, 17 were children aged 4-6 years (M = 5.56, SD = 0.63). One child's age was missing and imputed using the mean age of six. The remaining adult participants ranged from 19-34 years old (M = 24.00, SD = 4.15). Age data for 29 adults were missing and were similarly imputed using the group mean. Inclusion criteria required participants to have a normal IQ based on school records and teacher evaluations, normal or corrected visual acuity, no neurological or psychiatric disorders, no sensory visual impairments, and no prescribed medications. At the time of testing, none of the participants had a confirmed DD diagnosis.

Children were recruited via a direct connection within a local school, whereas adults were recruited through a Google Form circulated in the University of Fribourg. Written consent was obtained from the parents of child participants, and verbal consent was obtained from all participants prior to testing.

Experiment

The study was conducted in two to three phases. For children, the session took place in an empty classroom, while adult sessions were held in a lab at the University of Fribourg. All participants wore Mobile Pupil Labs Neon eye tracking glasses (Pupil Labs, Berlin, Germany), connected via USB-C to an Android phone equipped with the Pupil Cloud Companion App. These glasses recorded both eye movement data and a video of the visual scene throughout the experiment.

In the first phase, participants completed the RAN object subtest from the Dyslexia Adult Screening Test (DAST), presented on a 12.9-inch iPad Pro (2048 x 2732 px, 120Hz), in German and/or French. Participants were instructed to name a sequence of 20-line drawings as quickly and accurately as possible. After a brief explanation and practice trial, eye tracking was initiated, and the RAN task was administered. The RAN score was calculated as the time to complete the task in seconds, plus a 5-second penalty per error. This score was then

converted to objects per second. The task was repeated at least twice and, for some participants, in multiple languages depending on their language background. For 43 participants, the RAN task was repeated again at the end of the experiment (phase 3). Although the best RAN score across all trials was retained for analysis, we noted that 79.07% of participants performed better during the initial phase. The language of the best RAN trial was German for 56.14% of participants. This phase lasted approximately 10 minutes.

In the second phase, participants played the game *Fruit Ninja* on the same iPad for 20 minutes. Upon completion, child participants received a small treat, and adult participants were compensated at a rate of CHF 20/hour. This study was approved by the Business Administration System for Ethics Committees (BASEC) and conducted by the Department of Neuroscience, Faculty of Science and Medicine, University of Fribourg.

Modelling and Analysis

Object Detection in Game Frames

To compare participants' eye movements with *Fruit Ninja* game events, two YOLOv8s (Ultralytics) object detection models were trained to detect fruits and bombs. The first model, pre-trained on the COCO dataset, was fine-tuned on a custom Fruit Ninja dataset composed of 7,770 images from three different sources, with over half originating from our data collection. The images sampled from this study were automatically annotated using Roboflow's integrated auto-labelling system, followed by manual inspection for correction. Only whole fruits were labelled; once sliced, their fragments were excluded. Fruit rotation made intact status sometimes ambiguous, requiring additional review.

Despite achieving a mAP50 of 97.40% on a 202-image test set, the first model struggled with partially occluded objects. A second model was fine-tuned by continuing training from the first model's weights and adding 7,924 additional frames sampled at 3 FPS from recordings of three children and four adults under varying lighting conditions. Though this model achieved a lower mAP50 (85.10%), it detected partially occluded objects more reliably, providing more accurate onset times. Both models were trained using PyTorch 2.6.0 with CUDA acceleration. Detailed implementation and hardware configurations are provided in Appendix A and B.

Inference and Postprocessing

Object detection was applied frame-by-frame to all screen recordings. Postprocessing included assigning persistent object IDs, filtering out short-lived detections (< 5 frames), and timestamping. These timestamps were then used to align object detection with the eye-tracking data (including blinks, fixations, saccades, gaze, and IMU data).

Feature Extraction and Engineering

A total of 23 eye-tracking features were extracted from active *Fruit Ninja* gameplay segments and were grouped into defined feature sets (Table 1). Nine standard metrics were extracted from Pupil Cloud including mean and median blink duration (ms), mean and median fixation duration (ms), mean and median saccade length (px), and the number of blinks, fixations, and saccades per minute.

The Basic feature set included mean and median fixation duration, as well as mean and median saccade length (Asvestopoulou et al., 2019; Pan et al., 2013; Premeti et al., 2022; Raatikainen et al., 2021; Strandberg, 2019; Tooze, 2022; Vajs et al., 2022, 2023; Werth, 2021b). The Fixation and Saccade sets included the three respective duration/count features. In addition, a set of engineered features captured Area of Interest (AOI) measures, blink behaviours, fixation latency, and spatial complexity.

AOI features included the mean and median Euclidean distances between the gaze point and the nearest detected object (fruit or bomb), and fruit only. Additional AOI metrics were: percentage of fixations outside object bounding boxes and percentage of fruits that were never fixated.

The Blink feature set included mean and median blink duration (ms), blink frequency (blinks per minute), and a novel feature, blink ratio, defined as the proportion of task time spent blinking. For each gameplay trial, blink ratio was calculated by dividing the total blink duration by the total trial duration. Participant-level scores were obtained by averaging across trials, with values closer to 1 indicating more time spent blinking and values closer to 0 indicating more time spent with the eyes open. Unlike frequency or duration measures considered in isolation, blink ratio provides a global index of visual availability, capturing the overall extent to which visual input is interrupted during task performance.

Five latency features were derived by measuring the time from fruit onset to the first fixation within the object's bounding box: mean, median, standard deviation, minimum, and

maximum. Since the minimum latency was consistently 0 ms across all participants, reflecting cases where the gaze was already positioned on the object at onset, this feature was excluded from further analysis.

Three spatial features were computed following the methodology outlined by Vajs et al. (2022). The FIC measures how often a fixation path intersects itself. The mean and standard deviation of intersection counts across all fixations were included as features. The FFD, on the other hand, quantifies the spatial complexity of the gaze path within a fixation using the box-counting method. Higher FFD values reflect greater spatial irregularity and complexity. Both FIC and FFD were calculated only for fixations occurring when fruits were present on the screen. For more details regarding the implemented formulas, please see the appendices (Appendix C).

Table 1
Defined Feature Sets

Feature Subset	Variable
All (24 features)	Age (only for prediction task)
	Blink Ratio
	Fixation Fractal Dimension (FFD)
	Maximum Latency
	Mean Blink Duration
	Mean Fixation Intersection Coefficient (FIC)
	Mean Fixation Distance to Fruit
	Mean Fixation Distance to Object
	Mean Fixation Duration
	Mean Latency
	Mean Saccade Length
	Median Blink Duration
	Median Fixation Distance to Fruit
	Median Fixation Distance to Object
	Median Fixation Duration
	Median Latency
	Median Saccade Length
	Number of Blinks per Minute
	Number of Fixations per Minute
	Number of Saccades per Minute
	Percent of Fixations Outside the Bounding Boxes
	Percent Un-Fixated Fruits
	SD Fixation Intersection Coefficient (FIC)
	SD Latency

Feature Subset	Variable
AOI (6 features)	Mean Fixation Distance to Fruit [px]
	Mean Fixation Distance to Object [px]
	Median Fixation Distance to Fruit [px]
	Median Fixation Distance to Object [px]
	Percent of Fixations Outside the Bounding Boxes
	Percent Un-Fixated Fruits
Basic (4 features)	Mean Fixation Duration [ms]
	Median Fixation Duration [ms]
	Mean Saccade Length [px]
	Median Saccade Length [px]
Blinks (4 features)	Blink Ratio
	Mean Blink Duration [ms]
	Median Blink Duration [ms]
	Number of Blinks per Minute
Fixation (3 features)	Mean Fixation Duration [ms]
	Median Fixation Duration [ms]
	Number of Fixations per Minute
Latency (4 features)	Maximum Latency [ms]
	Mean Latency [ms]
	Median Latency [ms]
	SD Latency [ms]
Most Important (5 features)	Blink Ratio
	Fixation Fractal Dimension (FFD)
	Mean Blink Duration [ms]
	Median Blink Duration [ms]
	Percent Un-Fixated Fruits
Saccade (3 features)	Mean Saccade Length
	Median Saccade Length
	Number of Saccades per Minute
Significant (12 features)	Age (only for prediction task)
	Fixation Fractal Dimension (FFD)
	Mean Blink Duration [ms]
	Mean Fixation Distance to Fruit [px]
	Mean Fixation Distance to Object [px]
	Mean Saccade Length [px]
	Median Blink Duration [ms]
	Median Fixation Distance to Object [px]
	Median Saccade Length [px]
	Number of Blinks per Minute
	Percent Un-Fixated Fruits
	SD Fixation Intersection Coefficient (FIC)
Spatial (3 features)	Fixation Fractal Dimension (FFD)
	Mean Fixation Intersection Coefficient (FIC)
	SD Fixation Intersection Coefficient (FIC)

Feature Selection

To control for age effects in classification, residuals from a linear regression on age were computed and used in subsequent analyses. Participants were then divided into “good” and “poor” RAN groups via a median split. A RF classifier with an 80/20 train-test-split was used to rank feature importance. Visual inspection of the RF feature importance plot revealed a clear drop-off in scores, which served as an initial criterion for selecting key predictors. Shapley Additive Explanations (SHAP) were subsequently used to interpret the RF model. SHAP assigns additive values to each feature’s contribution to the prediction, with positive values pushing classifications towards the positive class (“good” RAN scorers). By comparing the RF rankings with the SHAP value distributions, a consistent set features was identified as the most informative across both approaches. It should be noted that while SHAP provides interpretability at the level of individual feature contributions, it does not capture interactions between features.

Pearson correlations were computed between all eye-tracking features and age with the raw RAN scores. Features significantly correlated with RAN were assigned to the "Significant" feature set.

Model Selection

To assess the predictive power of the 23 eye-tracking features and age on RAN score, a linear regression model was trained using 5-fold cross-validation. Statistical evaluation involved one-sample t-tests comparing R-squared values to zero.

Four of the most used classification models for DD detection—LR, SVM, KNN, and RF—were selected to classify participants into "good" vs. "poor" RAN groups. Children represented 20% of the “good” group and 40% of the “poor” group, reflecting their 30% representation in the total sample.

A total of 40 models (4 classifiers x 10 feature sets) were trained. For each model-feature pair, hyperparameters were optimised using Randomized Search CV (Appendix D). Final models were trained using Leave-One-Out Classification. All features were scaled (except for RF) to avoid bias. Performance was evaluated using accuracy, precision, recall, and F1 score. A binomial test assessed whether overall classification accuracy significantly exceeded chance level (50%). All modelling was conducted using the sci-kit learn Python library.

Results

Correlation Between Extracted Features and the RAN Score

Of the 24 features examined, including both eye-tracking measures from Fruit Ninja gameplay and age, 12 showed significant correlations with the raw RAN score (Figure 1). Age was strongly and positively correlated with RAN performance (r = .79, p < .001), confirming that older participants generally completed the RAN task more efficiently.

Three blink-related features were significantly associated with RAN performance. Both mean (r = .35, p = .007) and median blink durations (r = .36, p = .006) were positively correlated with RAN score, while the number of blinks per minute showed a negative correlation (r = -.40, p = .002). Blink ratio did not show a significant association with RAN performance.

Fixation Fractal Dimension (FFD), a measure of gaze path complexity, was negatively correlated with RAN score (r = -.32, p = .015), indicating that more chaotic fixations were associated with poorer RAN performance. Similarly, the percentage of un-fixated fruits (r = -.45, p < .001), was strongly and negatively associated with RAN scores, implying that efficient target selection plays a key role in rapid object naming.

Saccade-related metrics also showed significant associations. Mean (r = -.55, p < .001) and median (r = -.42, p < .001) saccade lengths were negatively correlated with RAN score, indicating that participants who made longer saccades tended to have lower RAN performance.

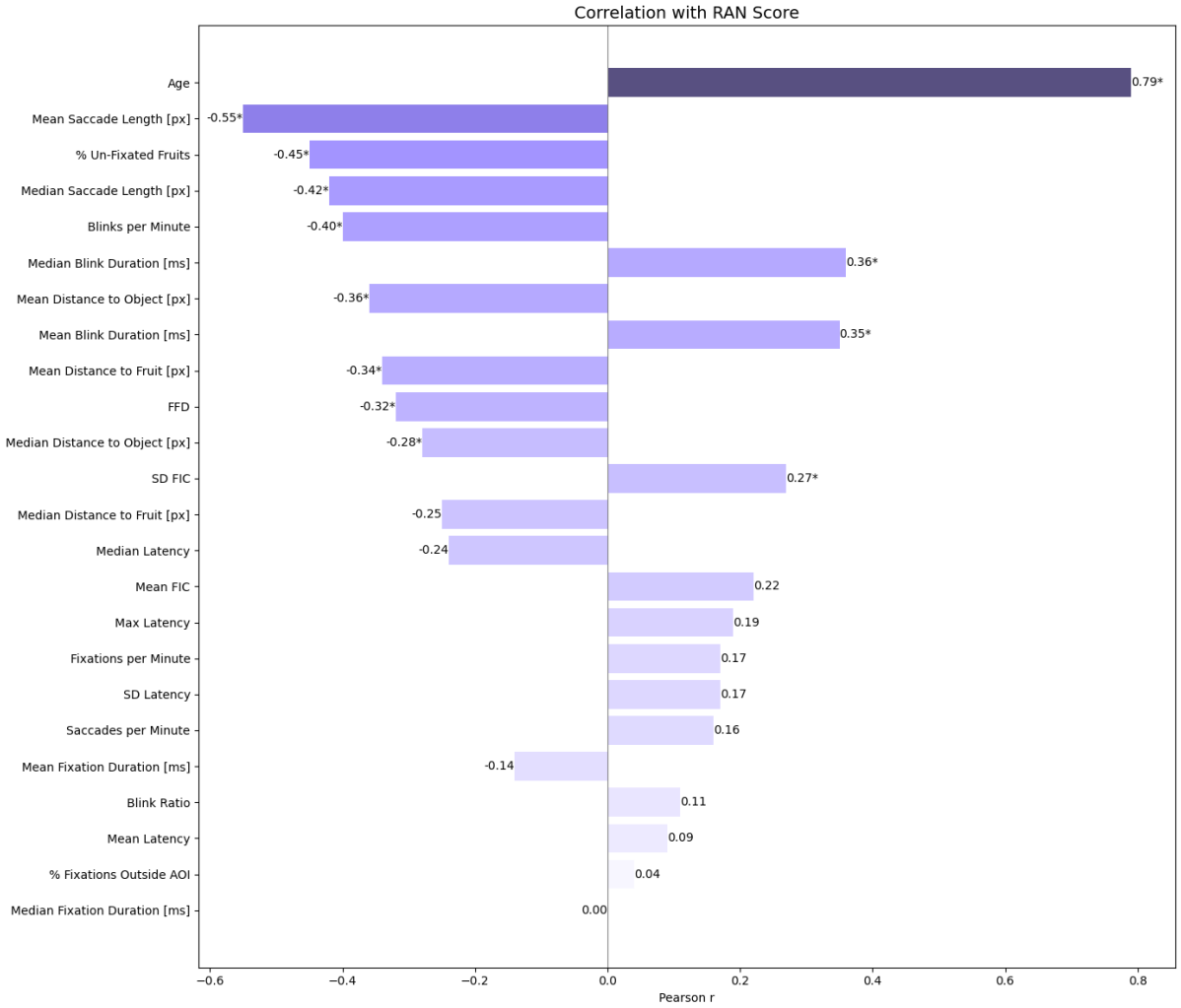
The standard deviation of the FIC—which captures variability in gaze self-intersections—was positively correlated with RAN score (r = .27, p = .044), indicating more variation in self-intersecting behaviours in better-performing participants.

Finally, spatial fixation features revealed that better RAN performance was associated with closer fixations to on-screen targets. Specifically, mean fixation distance to the nearest fruit (r = -.35, p = .011), mean fixation distance to the nearest object (r = -.36, p = .007), and median fixation distance to the nearest object (r = -.28, p = .035) all showed significant negative correlations with RAN score.

The remaining features did not show significant correlations with RAN performance (see Appendix E).

Figure 1

Pearson Correlations Between Features and RAN Score



Note. Statistically significant correlations with *p* values < .05 are marked with an asterisk (*).

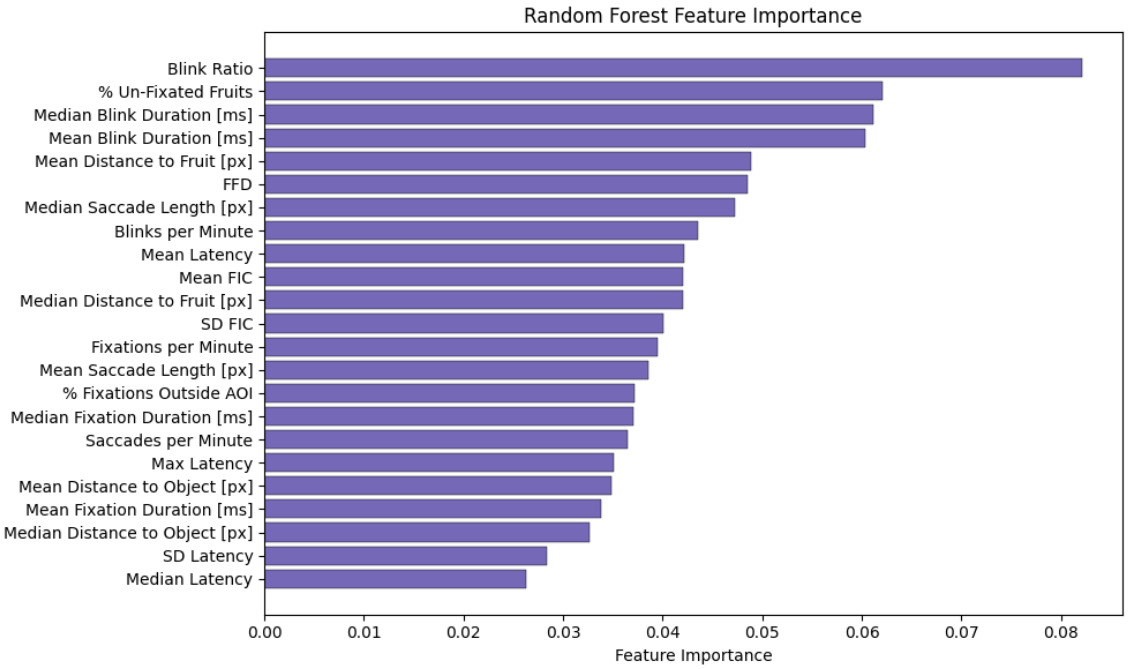
Feature Importance and Explainability

After visual inspection of the RF feature importance plot, four features with importance scores above 0.06 were identified as meaningful contributors to the model, with a noticeable drop in importance among the remaining features, suggesting diminishing marginal contributions (Figure 2). Notably, three of the four top-ranked features were blink-related, highlighting the potential relevance of blink behaviour in predicting RAN performance.

While Random Forest importance scores showed a clear drop after the top four predictors, SHAP values indicated a more gradual decline (Figure 3). To prioritise features that were consistently informative, emphasis was placed on predictors highlighted by both methods. The final feature set therefore included the four features ranked highest in both RF

Figure 2

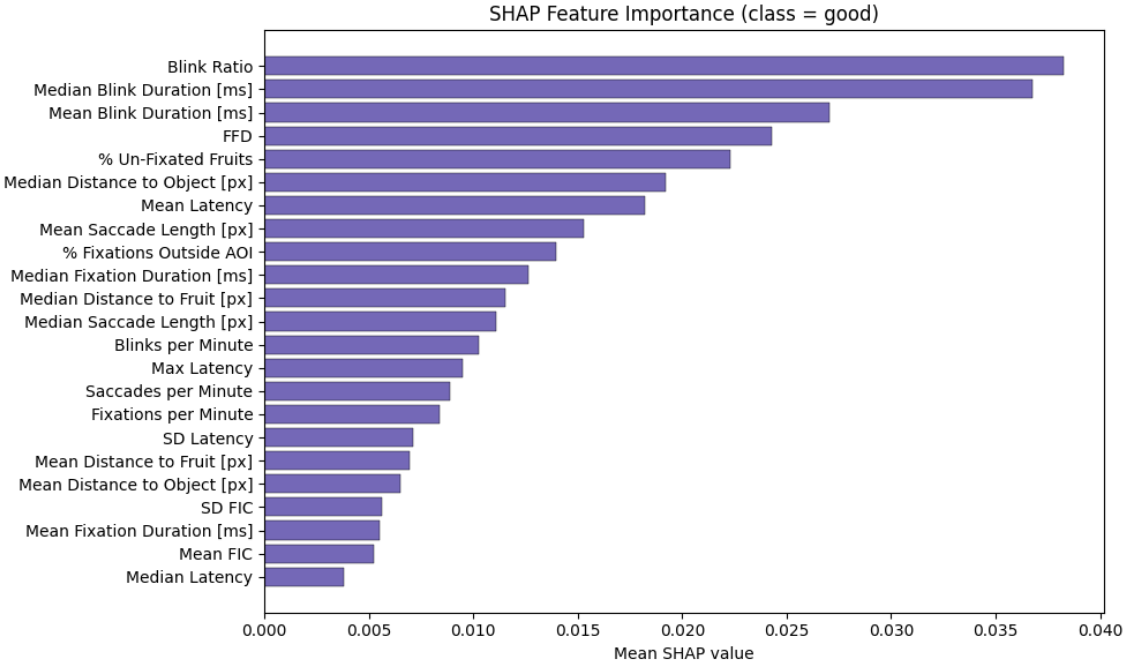
Random Forest Feature Ranking Importance Results



Note. The top four features with an importance greater than 0.06 were selected for the Most Important feature set in the analyses.

Figure 3

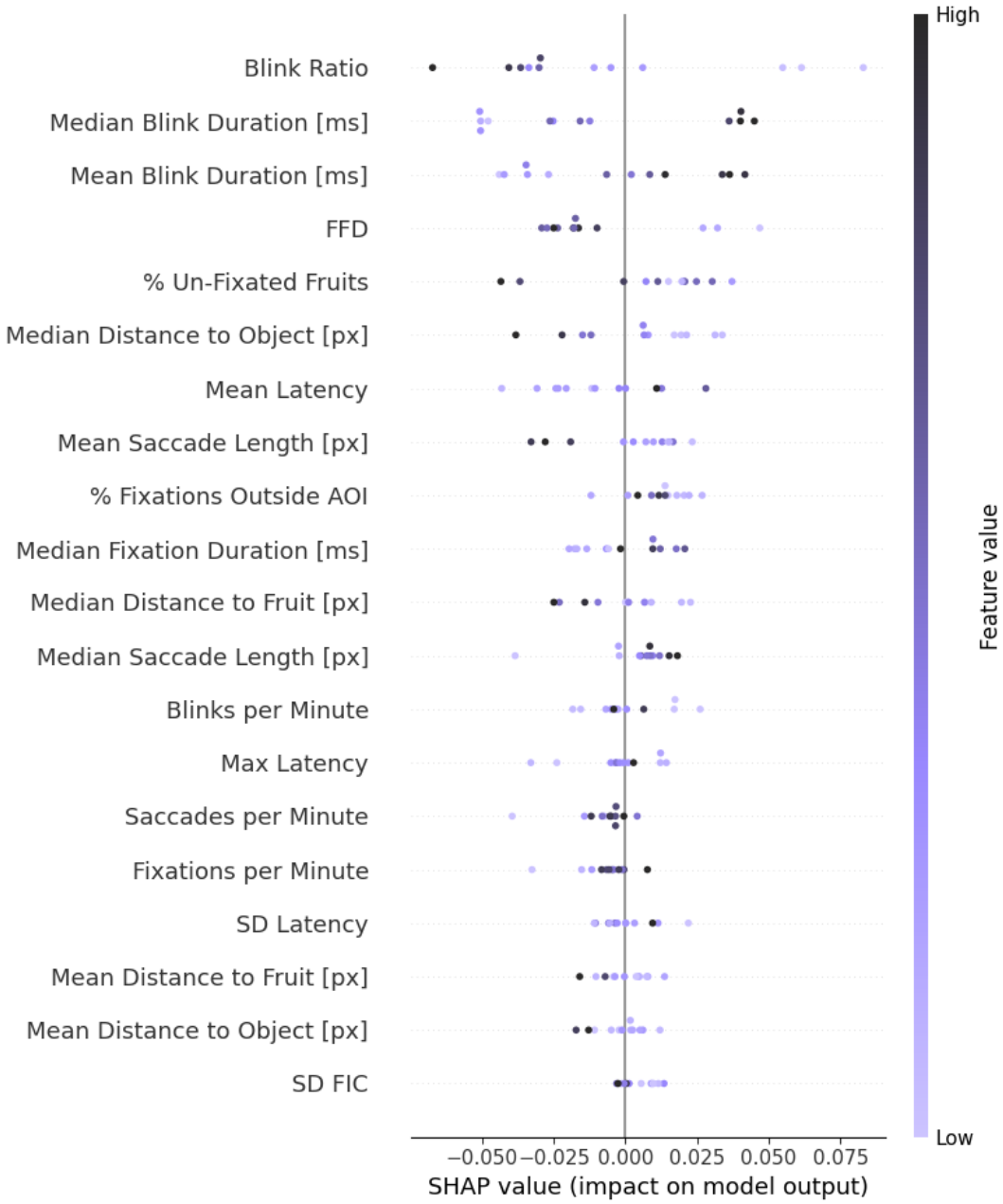
SHAP Feature Values



Note. The five features with the highest SHAP values were retained for further analysis.

Figure 4

SHAP on Random Forest Model



Note. Each point represents the SHAP value for one participant for a given feature; the x-axis shows how much that feature contributed to the prediction, and the colour indicates the actual feature value.

and SHAP, along with fixation fractal dimension (FFD), which fell just below the RF cutoff but ranked fourth in the SHAP analysis. This selection strategy ensured a robust set of predictors supported by multiple importance criteria, while avoiding overreliance on a single ranking method.

The SHAP analysis aligned with the directions observed in the correlation analysis (Figure 4). Participants with longer mean and median blink durations were more often classified into the "good" RAN group. In contrast, higher FFD, indicating more irregular fixation patterns, and a greater percentage of un-fixated fruits were linked to the "poor" RAN group. Together, these results highlight how SHAP values clarify the contribution of individual features to model predictions. Individual SHAP plots for all 12 RF test participants are included in Appendix F.

Predicting the RAN Score

Among the ten tested feature sets, two models significantly predicted RAN score: the Basic and Saccade feature sets. The Basic feature set demonstrated the strongest performance, with an R^2 mean of .30, MSE of .25, and p = .01 (Figure 5). The Saccade feature set followed closely with an R^2 mean of .26, MSE of .27, and p = .04 (Figure 6). These results indicate that multiple subsets of eye-tracking features offer statistically significant predictive power for RAN score. Detailed evaluation metrics for all ten feature sets are available in the appendices (Appendix G).

Figure 5

Prediction of the RAN Score Using the Basic Features

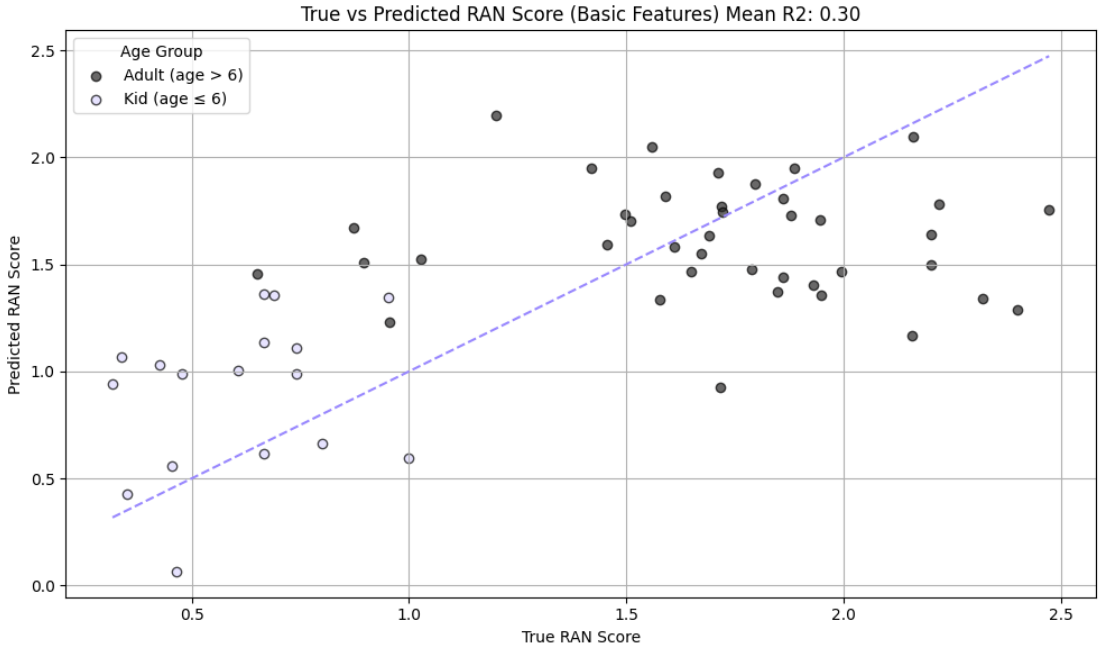
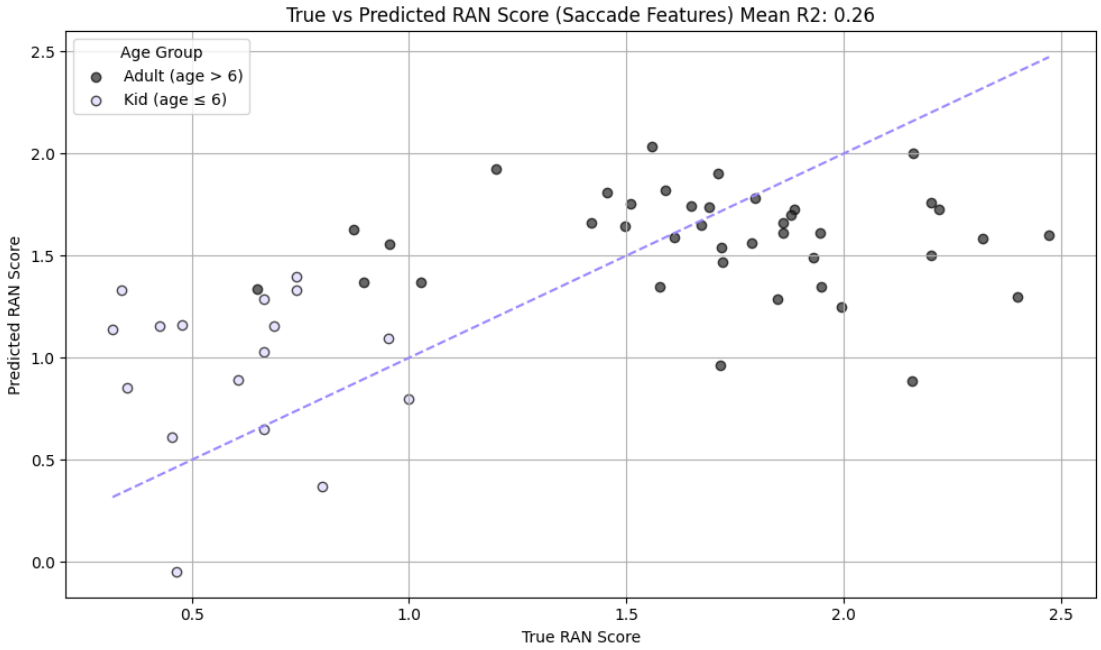


Figure 6

Prediction of the RAN Score Using the Saccade Features



Classification Results

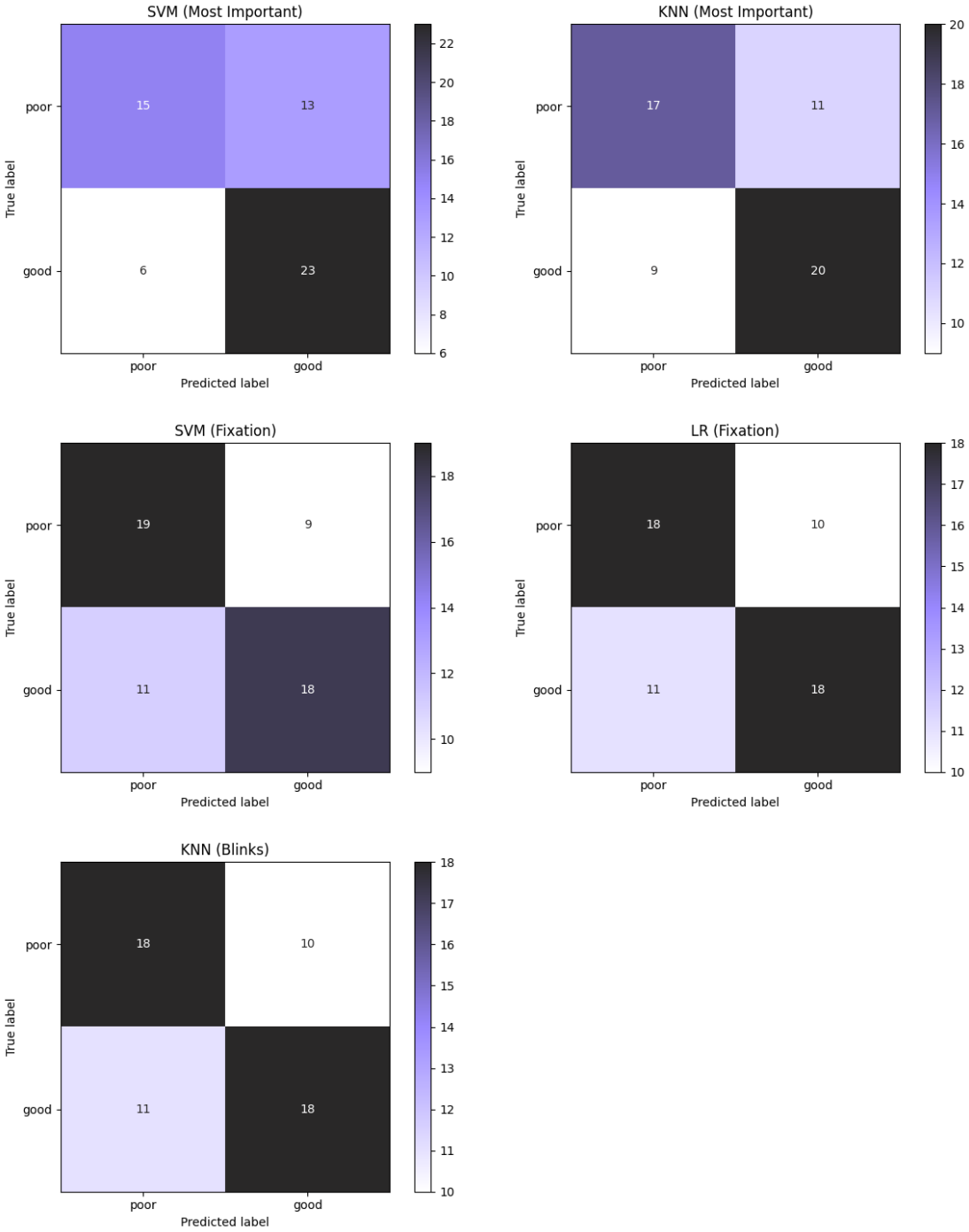
In total, five out of forty classification models achieved accuracies significantly above chance level. The best-performing model was an SVM trained on the most important feature set, reaching an accuracy of 66.67%, a recall of .54 for the "poor" RAN group, and a weighted F1 score of .66 (p = .01). A KNN trained on the same feature set performed comparably, with 64.91% accuracy, a recall of .61 for the "poor" group, and a weighted F1 of .65 (p = .02) (Figure 7). Two models trained on the fixation feature set also performed significantly above chance. The SVM achieved 64.91% accuracy, a recall of .68 for the "poor" group, and a weighted F1 of .65 (p = .02), while the LR model achieved 63.16% accuracy, a recall of .64 for the "poor" group, and a weighted F1 of .63 (p = .03). Finally, a KNN trained on the blinks feature set reached 63.16% accuracy, with a recall of .62 for the "poor" group and a weighted F1 of .63 (p = .03) (Figure 7). Across classifiers, the most important feature set produced the highest mean accuracy (61.50%), with a mean recall of .56 for the "poor" group and a weighted F1 of .61.

By contrast, most other models performed around or below chance level, despite the use of carefully selected features (Appendix H). The weakest results were obtained with the Basic feature set and the full feature set. The Basic set, despite overlapping substantially with the stronger fixation set, achieved only 40.50% mean accuracy, a recall of .27 for the "poor"

group, and a weighted F1 of .39. Similarly, the full feature set yielded a mean accuracy of 40.00%, a recall of .22 for the “poor” group, and a weighted F1 of .37. A complete summary of mean classification results for all feature sets is provided in Appendix I.

Figure 7

Classification of “good” vs. “poor” RAN Scores



Note. Confusion matrices are shown only for models that performed significantly above chance level, $p < .05$

Discussion

This pilot study investigated whether eye-tracking features from a non-reading visual task (*Fruit Ninja*) could predict RAN scores and classify individuals as "good" vs. "poor" RAN performers using various machine learning models and feature sets. Of the 24 features, 12 were significantly correlated with RAN scores, particularly those related to fixations, saccades, and blinks—suggesting a link between oculomotor control and reading-related abilities. Several of these findings align with previous literature, most notably the strong positive correlation between age and RAN score performance (Lervåg & Hulme, 2009; Mascheretti et al., 2018; Pan et al., 2013). Blink behaviour was particularly informative for the prediction of RAN scores. While the number of blinks per minute was negatively correlated with RAN score—potentially reflecting task difficulty and engagement—mean and median blink durations were positively correlated. These findings support the hypothesis that blink behaviour can serve as a proxy for cognitive load or magnocellular processing difficulty (Tooze, 2022). However, the literature has reported mixed results regarding blink duration in individuals with DD, and this remains an area requiring further investigation (Tooze, 2022).

The FFD, which captures gaze path complexity, was negatively correlated with RAN score, confirming that participants with more erratic gaze patterns performed worse. In contrast, the standard deviation of the FIC was positively correlated with RAN score, a finding that diverges from prior reports suggesting greater FIC variability in individuals with DD (Vajs et al., 2022). This discrepancy could stem from task differences or the subclinical nature of the current sample. Additionally, the percentage of un-fixated fruits was negatively correlated with RAN score, indicating that participants who relied less on peripheral vision and more on direct fixations were more likely to achieve better RAN scores. Mean and median saccade lengths were also negatively correlated, suggesting that shorter saccades were linked to better RAN performance, a finding that contrasts some literature on DD saccade profiles (Asvestopoulou et al., 2019; Premeti et al., 2022).

Each feature set, except for latency, included at least one significantly correlated variable. The latency features, which measured the time to first fixation following object onset, showed no significant correlations with RAN score and did not contribute meaningfully to any predictive or classification models. This suggests that visual orienting speed may not capture reading-relevant processes in this context, or that peripheral processing during the game may have masked such deficits.

The strongest predictive performance for RAN scores was achieved using the Basic feature set. This model included age, the most strongly correlated individual variable, and explained 30.40% of the variance in RAN scores, yielding statistically significant predictions. The Saccade feature set, which did not include age, also produced statistically significant predictions and explained 26.26% of the variance. This demonstrates that age was an important predictor in the Basic model, but that saccade-related features alone still accounted for a substantial proportion of variance in RAN performance. Both sets included saccade-based measures, consistent with previous work linking atypical saccadic control to reading difficulties. However, the pattern observed here diverges from much of the existing literature, where dyslexic readers are typically characterised by more frequent and shorter saccades (Asvestopoulou et al., 2019; Premeti et al., 2022). In the present study, longer saccades during Fruit Ninja gameplay were associated with poorer RAN performance, suggesting that in this fast-paced, non-reading task, saccade length does not serve as a distinguishing predictor of RAN scores in the same way it has been shown to differentiate performance in reading-based classification tasks. The consistent inclusion of saccade length across both significant models nonetheless indicates that the spatial precision and efficiency of eye movements may provide a useful proxy for reading-related oculomotor control. The success of the Basic model, despite its simplicity, further reinforces the value of established eye-tracking features in capturing core components of visual attention and control relevant to reading performance.

Despite the success in predicting RAN scores, these features did not support accurate binary classification. Of the 40 classification models tested, only five achieved statistically significant accuracies (63-67%), while most others performed at or near chance. The best-performing feature set was the subset selected by RF feature ranking and SHAP analysis, which highlighted blink ratio, mean and median blink duration, FFD, and the percentage of un-fixated fruits. Together, these features reflect two interrelated domains of gaze behaviour, temporal continuity of visual input and spatial organisation of fixations, both of which are hypothesised to involve the magnocellular visual stream (Vajs et al., 2022).

Blink-related features dominated the RF rankings, suggesting that blink behaviour is a sensitive marker of magnocellular inefficiency. SHAP analysis refined this interpretation: higher blink ratios (greater overall time spent with the eyes closed) predicted poorer RAN performance, while longer average blink durations were associated with better outcomes.

This asymmetry may reflect a distinction between maladaptive frequent short blinks, which fragment the visual stream, and adaptive longer blinks, which could stabilise visual processing by reducing oculomotor noise.

The importance of blink ratio in the SHAP analysis indicates that extended interruptions to vision may initiate a cascade of downstream effects. Periods with the eyes closed fragment the continuity of scene representation, increasing the likelihood of un-fixated targets and irregular fixation distributions, as reflected in higher FFD values. This helps explain why blink features were so dominant in the RF rankings and why fixation features still contributed to classification performance despite weaker linear associations with RAN scores.

A plausible explanation is a self-reinforcing blink–saccade loop. Frequent saccades elevate ocular strain and increase the need for lubrication, which prompts additional blinks. These blinks disrupt perception and elicit corrective saccades, and the resulting oculomotor load further increases blink frequency and duration. This loop could explain both the dominance of blink metrics and their association with inefficient scanning patterns. Notably, one of the blink-only feature sets produced a significant classification model, reinforcing the diagnostic potential of blink behaviour in literacy-independent tasks. Unlike in reading, where blinks have not consistently emerged as strong predictors, the present task required continuous tracking of multiple moving objects, which may have increased visual strain and elicited more blinks. Future studies could incorporate additional blink-related features, such as average interblink latency, to assess the temporal regularity of blink generation under different task demands.

Another well-performing feature set was the fixation set. The success of fixation features in classification may reflect their sensitivity to categorical differences in attentional control and visual strategy. Even if individual fixation metrics do not track RAN performance linearly, their combined patterns may capture discrete differences between “good” and “poor” performers, particularly in fast-paced visual tasks such as Fruit Ninja. SHAP analysis revealed that longer median fixation durations and a higher frequency of fixations pushed the models towards classifying participants as good RAN performers. This pattern is consistent with the idea that more stable and controlled fixations reflect stronger attentional allocation and more efficient scene sampling (Ebrahimi et al., 2022; Stein, 2019, 2022).

The relevance of fixation features is further supported by prior findings in dyslexia research, where readers with DD typically show longer fixations, greater fixation counts, and

reduced fixation stability (Asvestopoulou et al., 2019; Pavlidis, 1981; Premeti et al., 2022). Such abnormalities have been linked to magnocellular dysfunction, which disrupts the temporal and spatial precision of attentional shifts (Stein, 2018, 2019, 2022). The present results suggest that fixation metrics can provide similarly valuable information outside of reading contexts: in dynamic, non-reading tasks, they appear to capture aspects of visual control that differentiate between better and poorer RAN performers.

In both significant fixation-based models, recall for the “poor” RAN group exceeded overall accuracy, an encouraging property for diagnostic applications where minimising false negatives is essential. Moreover, fixation features, like blink features, are not only informative but also practical. They can be extracted reliably from standard eye-tracking data streams, without the need for complex preprocessing, and thus are well suited for integration into real-time diagnostic tools. Their demonstrated utility in both reading and non-reading tasks highlights their potential for large-scale, literacy-independent screening applications.

While these models do not yet meet clinical standards, the broader implications are clear. Meaningful signals related to reading ability can be extracted from gaze behaviour in a non-reading task, even in the absence of a formal diagnosis. Unlike many prior studies that focus on distinguishing between neurotypical and diagnosed individuals, this study addresses a more subtle but potentially more impactful question: whether risk can be detected before a diagnosis is established. This framing likely contributes to the more moderate classification accuracies observed, as identifying at-risk individuals in a mixed population is inherently more challenging than separating already-diagnosed groups. By contrast, studies using clearly defined diagnostic categories often report higher accuracies, but at the cost of reduced generalisability for early screening.

A key limitation of this study, as in much of the field, is the modest sample size. Building robust, generalisable models will require larger and more diverse datasets. Future work should aim to expand recruitment efforts or leverage open-access eye-tracking datasets where appropriate.

One of the study's major strengths lies in its use of a gamified, non-reading task. This approach enhances participant engagement, especially among children, and reduces the stress commonly associated with literacy assessments. This is particularly beneficial for younger or pre-reading populations, where maintaining motivation and attention is critical. The *Fruit Ninja* task likely encouraged more naturalistic gaze behaviour and aligns with

broader efforts to develop accessible, low-stigma tools for early screening. However, the novelty of the task also presents a limitation. Most previous research relies on standardised reading tasks, making direct comparisons difficult. While the results show predictive utility within the context of a visual game, further studies are needed to test whether these findings generalise across different tasks and populations, including diagnosed, at-risk, and neurotypical individuals.

Conclusion

This pilot study contributes to the development of early DD screening by demonstrating the feasibility of using eye-tracking and machine learning to model reading-related cognitive traits during a non-reading, gamified task. As the first study to apply this approach to RAN score prediction and classification, it offers a literacy-independent, enjoyable, and accessible alternative to traditional screening tools—particularly valuable for pre-readers and young children.

Although classification accuracy was moderate, it is consistent with expectations in a non-clinical, developmentally diverse sample. Importantly, the findings reveal meaningful variation in oculomotor behaviour that correlates with RAN performance, suggesting that attentional control and visual efficiency, especially those linked to magnocellular-dorsal function, can be assessed even before reading failure becomes apparent.

This work highlights the potential of gamified eye-tracking tools for scalable, low-stigma screening. Such tools could be embedded in school settings or educational apps, enabling widespread, early identification of DD and other attentional disorders. A single, enjoyable task could offer rich, interpretable data useful for screening for learning and attentional disorders, helping educators and clinicians better understand each child's profile and tailor interventions accordingly.

With continued research and validation, this method could become a cost- and time-efficient screening alternative. Eye-tracking capabilities are now found in mobile devices, and with the right models and feature selection, testing could occur at home or in classrooms, requiring less clinician oversight. This would not only save hours of diagnostic time but also improve access to early support, particularly in under-resourced settings.

Future studies should focus on validating this approach in clinically diagnosed samples, longitudinal outcomes, and exploring temporally dynamic models such as LSTMs or transformers to better capture the temporal progressing of gaze behaviour. This study lays the groundwork for such advances, opening the door to equitable, scalable screening technologies for developmental disorders like DD.

References

Araújo, S., Reis, A., Petersson, K. M., & Faísca, L. (2015). Rapid automatized naming and reading performance: A meta-analysis. *Journal of Educational Psychology*, *107*(3), 868–883. https://doi.org/10.1037/edu0000006

Asvestopoulou, T., Manousaki, V., Psistakis, A., Smyrnakis, I., Andreadakis, V., Aslanides, I. M., & Papadopouli, M. (2019). *DysLexML: Screening Tool for Dyslexia Using Machine Learning* (arXiv:1903.06274). arXiv. https://doi.org/10.48550/arXiv.1903.06274

Bartolomé, N. A., Zorrilla, A. M., & Zapirain, B. G. (2012). Dyslexia diagnosis in reading stage though the use of games at school. *2012 17th International Conference on Computer Games (CGAMES)*, 12–17. https://doi.org/10.1109/CGames.2012.6314544

Bavelier, D., & Davidson, R. J. (2013). Games to do you good. *Nature*, *494*(7438), 425–426. https://doi.org/10.1038/494425a

Daniel, S. S., Walsh, A. K., Goldston, D. B., Arnold, E. M., Reboussin, B. A., & Wood, F. B. (2006). Suicidality, School Dropout, and Reading Problems Among Adolescents. *Journal of Learning Disabilities*, *39*(6), 507–514. https://doi.org/10.1177/00222194060390060301

Demb, J. B., Boynton, G. M., Best, M., & Heeger, D. J. (1998). Psychophysical evidence for a magnocellular pathway deficit in dyslexia. *Vision Research*, *38*(11), 1555–1559. https://doi.org/10.1016/S0042-6989(98)00075-3

Demb, J. B., Boynton, G. M., & Heeger, D. J. (1997). Brain activity in visual cortex predicts individual differences in reading performance. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(24), 13363–13366.

Dyslexia FAQ. (n.d.). Yale Dyslexia. Retrieved 9 March 2024, from https://dyslexia.yale.edu/dyslexia/dyslexia-faq/

Ebrahimi, L., Pouretemad, H., Stein, J., Alizadeh, E., & Khatibi, A. (2022). Enhanced reading abilities is modulated by faster visual spatial attention. *ANNALS OF DYSLEXIA*, *72*(1), 125–146. https://doi.org/10.1007/s11881-021-00245-x

Eden, G. F., Stein, J. F., Wood, H. M., & Wood, F. B. (1994). Differences in eye movements and reading problems in dyslexic and normal children. *Vision Research*, *34*(10), 1345–1358. https://doi.org/10.1016/0042-6989(94)90209-7

Eden, G. F., VanMeter, J. W., Rumsey, J. M., & Zeffiro, T. A. (1996). The Visual Deficit Theory of Developmental Dyslexia. *NeuroImage*, *4*(3), S108–S117. https://doi.org/10.1006/nimg.1996.0061

Gomolka, Z., Zeslawska, E., Czuba, B., & Kondratenko, Y. (2024). Diagnosing Dyslexia in Early School-Aged Children Using the LSTM Network and Eye Tracking Technology. *Applied Sciences*, *14*(17), Article 17. https://doi.org/10.3390/app14178004

Huettig, F., Lachmann, T., Reis, A., & Petersson, K. M. (2018). Distinguishing cause from effect—Many deficits associated with developmental dyslexia may be a consequence of reduced and suboptimal reading experience. *Language, Cognition and Neuroscience*, *33*(3), 333–350. https://doi.org/10.1080/23273798.2017.1348528

IDA. (2002, November 12). *Definition of Dyslexia—International Dyslexia Association*. https://dyslexiaida.org/definition-of-dyslexia/

Jednoróg, K., Marchewka, A., Tacikowski, P., Heim, S., & Grabowska, A. (2011). Electrophysiological evidence for the magnocellular-dorsal pathway deficit in dyslexia. *Developmental Science*, *14*(4), 873–880. https://doi.org/10.1111/j.1467-7687.2011.01037.x

Ji, Y., & Bi, H.-Y. (2020). Visual Dysfunction in Chinese Children With Developmental Dyslexia: Magnocellular-Dorsal Pathway Deficit or Noise Exclusion Deficit? *Frontiers in Psychology*, *11*, 958. https://doi.org/10.3389/fpsyg.2020.00958

Klistorner, A., Crewther, D. P., & Crewther, S. G. (1997). Separate magnocellular and parvocellular contributions from temporal analysis of the multifocal VEP. *Vision Research*, *37*(15), 2161–2169. https://doi.org/10.1016/S0042-6989(97)00003-5

Laycock, R., Crewther, D. P., & Crewther, S. G. (2008). The advantage in being magnocellular: A few more remarks on attention and the magnocellular system. *Neuroscience and Biobehavioral Reviews*, *32*(8), 1409–1415. https://doi.org/10.1016/j.neubiorev.2008.04.008

Leppänen, P. H. T., Hämäläinen, J. A., Salminen, H. K., Eklund, K. M., Guttorm, T. K., Lohvansuu, K., Puolakanaho, A., & Lyytinen, H. (2010). Newborn brain event-related potentials revealing atypical processing of sound frequency and the subsequent association with later literacy skills in children with familial dyslexia. *Cortex*, *46*(10), 1362–1376. https://doi.org/10.1016/j.cortex.2010.06.003

Lervåg, A., & Hulme, C. (2009). Rapid Automatized Naming (RAN) Taps a Mechanism That Places Constraints on the Development of Early Reading Fluency. *Psychological Science*, *20*(8), 1040–1048. https://doi.org/10.1111/j.1467-9280.2009.02405.x

Leung, T., Cheong, A. M., & Chan, H. H. (2022). Deficits in the Magnocellular Pathway of People with Reading Difficulties. *CURRENT DEVELOPMENTAL DISORDERS REPORTS*, *9*(3), 68–75. https://doi.org/10.1007/s40474-022-00248-2

Livingstone, M. S., Rosen, G. D., Drislane, F. W., & Galaburda, A. M. (1991). Physiological and anatomical evidence for a magnocellular defect in developmental dyslexia.

Proceedings of the National Academy of Sciences, *88*(18), 7943–7947. https://doi.org/10.1073/pnas.88.18.7943

Mammarella, I. C., Ghisi, M., Bomba, M., Bottesi, G., Caviola, S., Broggi, F., & Nacinovich, R. (2016). Anxiety and Depression in Children With Nonverbal Learning Disabilities, Reading Disabilities, or Typical Development. *Journal of Learning Disabilities*, *49*(2), 130–139. https://doi.org/10.1177/0022219414529336

Mascheretti, S., Gori, S., Trezzi, V., Ruffino, M., Facoetti, A., & Marino, C. (2018). Visual motion and rapid auditory processing are solid endophenotypes of developmental dyslexia. *Genes, Brain, and Behavior*, *17*(1), 70–81. https://doi.org/10.1111/gbb.12409

McBride, H. E. A., & Siegel, L. S. (1997). Learning Disabilities and Adolescent Suicide. *Journal of Learning Disabilities*, *30*(6), 652–659. https://doi.org/10.1177/002221949703000609

Pan, J., Yan, M., Laubrock, J., Shu, H., & Kliegl, R. (2013). Eye–voice span during rapid automatized naming of digits and dice in Chinese normal and dyslexic children. *Developmental Science*, *16*(6), 967–979. https://doi.org/10.1111/desc.12075

Pavlidis, G. Th. (1981). Do eye movements hold the key to dyslexia? *Neuropsychologia*, *19*(1), 57–64. https://doi.org/10.1016/0028-3932(81)90044-0

Peters, J. L., De Losa, L., Bavin, E. L., & Crewther, S. G. (2019). Efficacy of dynamic visuo-attentional interventions for reading in dyslexic and neurotypical children: A systematic review. *NEUROSCIENCE AND BIOBEHAVIORAL REVIEWS*, *100*, 58–76. https://doi.org/10.1016/j.neubiorev.2019.02.015

Premeti, A., Bucci, M. P., & Isel, F. (2022). Evidence from ERP and Eye Movements as Markers of Language Dysfunction in Dyslexia. *Brain Sciences*, *12*(1), Article 1. https://doi.org/10.3390/brainsci12010073

Raatikainen, P., Hautala, J., Loberg, O., Kärkkäinen, T., Leppänen, P., & Nieminen, P. (2021). Detection of developmental dyslexia with machine learning using eye movement data. *Array*, *12*, 100087. https://doi.org/10.1016/j.array.2021.100087

Schulte-Körne, G., Deimel, W., Bartling, J., & Remschmidt, H. (1998). Auditory processing and dyslexia: Evidence for a specific speech processing deficit. *NeuroReport*, *9*(2), 337.

Stein, J. (2018). What is Developmental Dyslexia? *Brain Sciences*, *8*(2), Article 2. https://doi.org/10.3390/brainsci8020026

Stein, J. (2019). The current status of the magnocellular theory of developmental dyslexia. *Neuropsychologia*, *130*, 66–77. https://doi.org/10.1016/j.neuropsychologia.2018.03.022

Stein, J. (2022). The visual basis of reading and reading difficulties. *FRONTIERS IN NEUROSCIENCE*, *16*, 1004027. https://doi.org/10.3389/fnins.2022.1004027

Stein, J. (2023). Theories about Developmental Dyslexia. *Brain Sciences*, *13*(2), Article 2. https://doi.org/10.3390/brainsci13020208

Stein, J., & Walsh, V. (1997). To see but not to read; the magnocellular theory of dyslexia. *Trends in Neurosciences*, *20*(4), 147–152. https://doi.org/10.1016/s0166-2236(96)01005-3

Strandberg, A. (2019). Eye movements during reading and reading assessment in swedish school children: A new window on reading difficulties. *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, 1–3. https://doi.org/10.1145/3314111.3322878

Talcott, J. B., Witton, C., Hebb, G. S., Stoodley, C. J., Westwood, E. A., France, S. J., Hansen, P. C., & Stein, J. F. (2002). On the relationship between dynamic visual and auditory

processing and literacy skills; results from a large primary-school study. *Dyslexia (Chichester, England)*, *8*(4), 204–225. https://doi.org/10.1002/dys.224

Tooze, L. (2022). Fixation and Blink Eye Movement Behavior in Dyslexic and Non-Dyslexic English Readers. *Open Journal of Social Sciences*, *10*(12), Article 12. https://doi.org/10.4236/jss.2022.1012001

Usman, O. L., Muniyandi, R. C., Omar, K., & Mohamad, M. (2021). Advance Machine Learning Methods for Dyslexia Biomarker Detection: A Review of Implementation Details and Challenges. *IEEE Access*, *9*, 36879–36897. IEEE Access. https://doi.org/10.1109/ACCESS.2021.3062709

Vajs, I., Ković, V., Papić, T., Savić, A. M., & Janković, M. M. (2022). Spatiotemporal Eye-Tracking Feature Set for Improved Recognition of Dyslexic Reading Patterns in Children. *Sensors*, *22*(13), Article 13. https://doi.org/10.3390/s22134900

Vajs, I., Papić, T., Ković, V., Savić, A. M., & Janković, M. M. (2023). Accessible Dyslexia Detection with Real-Time Reading Feedback through Robust Interpretable Eye-Tracking Features. *Brain Sciences*, *13*(3), Article 3. https://doi.org/10.3390/brainsci13030405

Werth, R. (2021a). Dyslexic Readers Improve without Training When Using a Computer-Guided Reading Strategy. *Brain Sciences*, *11*(5), Article 5. https://doi.org/10.3390/brainsci11050526

Werth, R. (2021b). Is Developmental Dyslexia Due to a Visual and Not a Phonological Impairment? *BRAIN SCIENCES*, *11*(10), 1313. https://doi.org/10.3390/brainsci11101313

Appendix A

Model 1 Implementation and Hardware Specifications

A.1 Device and Software Specifications

- Device: CUDA NVIDIA GeForce RTX 3050 Ti Laptop GPU, 4096MiB
- Python Version: 3.10.16
- PyTorch Version: 2.6.0+cu126
- Ultralytics Version: N/A

Table A1

Training Dataset Composition

Source	Number of Annotated Images	Link
Data Collection 1 (Automatically labelled with Roboflow)	2,373	https://universe.roboflow.com/fruit-ninja-mumar/fruit-ninja-fruits-bombs/dataset/4
Data Collection 2 (Unknown labelling technique)	2,103	
Roboflow Dataset (Unknown labelling technique)	3,294	https://universe.roboflow.com/yolo-j3jjb/merged-fruitninja/dataset/4/images?split=train
Total	7,770	

Note. The number of images reported for the Roboflow dataset reflects only the subset used during training, not the total number available in the linked dataset.

Table A2

Test Dataset Composition

Source	Number of Annotated Images	Link
Data Collection 2 (Automatically labelled with Roboflow)	202	https://universe.roboflow.com/fruit-ninja-mumar/fruit-ninja-fruits-bombs/dataset/4

A.2 Preprocessing and Data Augmentation

Source	Preprocessing	Data Augmentation
Data Collection 1	Auto-Orient: Applied Resize: Stretch to 640x640	Outputs per training example: 3 Flip: Horizontal Crop: 0% Minimum Zoom, 25% Maximum Zoom Rotation: Between -15° and +15° Brightness: Between -5% and +5%
Data Collection 2	Unknown	Unknown
Roboflow Dataset	Resize: Stretch to 640x640 Filter Null: Require at least 90% of images to contain annotations	Outputs per training example: 3 Flip: Horizontal

Note. Preprocessing and data augmentation was handled upon export from Roboflow.

A.3 Training Configuration

- Model: YOLOv8s
- Epochs: 50
- Image Size: 640
- Batch Size: 16
- Cache: True
- Amp: True

Trained 50 epochs in 6.17 hours.

Table A3

Model Performance

Class	Images	Instances	Box(P	R	mAP50	mAP50-95
All	202	148	0.941	0.948	0.974	0.795
Bomb	27	32	0.970	0.998	0.993	0.786
Fruit	77	116	0.912	0.897	0.954	0.803

Appendix B

Model 2 Implementation and Hardware Specifications

B.1 Device and Software Specifications

- Device: CUDA NVIDIA L4, 22693MiB
- Python Version: 3.11.12
- PyTorch Version: 2.6.0+cu124
- Ultralytics Version: 8.3.127

Table B1

Training Dataset Composition

Source	Number of Annotated Images	Link
Data Collection 3 (Automatically labelled with Roboflow)	4,347	https://universe.roboflow.com/fruit-ninja-mumar/fruit-ninja-fruits-bombs-2c/dataset/1
Data Collection 4 (Automatically labelled with Roboflow)	2,373	https://universe.roboflow.com/fruit-ninja-mumar/fruit-ninja-fruits-bombs/dataset/4
Total	6,720	

Note. More images were annotated due to the recruitment of new participants with differing lighting conditions.

Table B2

Test Dataset Composition

Source	Number of Annotated Images	Link
Data Collection 3 (Automatically labelled with Roboflow)	629	https://universe.roboflow.com/fruit-ninja-mumar/fruit-ninja-fruits-bombs-2c/dataset/1
Data Collection 4 (Automatically labelled with Roboflow)	202	https://universe.roboflow.com/fruit-ninja-mumar/fruit-ninja-fruits-bombs/dataset/4
Total	831	

Note. More images were annotated due to the recruitment of new participants with differing lighting conditions.

B.2 Preprocessing and Data Augmentation

Source	Preprocessing	Data Augmentation
Data Collection 3	Auto-Orient: Applied	Outputs per training example: 3
	Resize: Stretch to 640x640	Flip: Horizontal
		Crop: 0% Minimum Zoom, 15% Maximum Zoom
		Rotation: Between -15° and +15°
		Brightness: Between -5% and +5%
Data Collection 4	Auto-Orient: Applied	Outputs per training example: 3
	Resize: Stretch to 640x640	Flip: Horizontal
		Crop: 0% Minimum Zoom, 25% Maximum Zoom
		Rotation: Between -15° and +15°
		Brightness: Between -5% and +5%

Note. Preprocessing and data augmentation was handled upon export from Roboflow.

B.3 Training Configuration

- Model: Resumed training from the best weights of Model 1 YOLOv8s
- Epochs: 300
- Augment: True
- Patience: 50
- Optimizer: AdamW
- Learning Rate: 0.0001
- Image Size: 640
- Batch Size: 16
  Trained 300 epochs in 6.79 hours.

Table B3

Model Performance

Class	Images	Instances	Box(P	R	mAP50	mAP50-95
All	831	810	0.886	0.792	0.851	0.608
Bomb	158	176	0.877	0.812	0.823	0.550
Fruit	383	634	0.896	0.772	0.878	0.665

Appendix C

This study included two spatial features, FIC and FFD, to quantify the complexity and irregularity of gaze behaviour, following the methodology described by Vajs et al. (2022). Both features were computed only for fixations occurring when at least one fruit was visible on the screen during gameplay.

Fixation Intersection Coefficient (FIC)

The FIC measures the degree to which gaze paths within individual fixations intersect themselves, reflecting the disorganization or looping in gaze trajectories. Each fixation is represented as a time-ordered sequence of (x, y) gaze coordinates. The algorithm checks for intersections between all non-adjacent line segments formed by consecutive gaze points within a fixation.

The FIC is calculated as the average number of self-intersections per fixation:

$$FIC = \frac{1}{n}\sum_{j=1}^{n} FI_j$$

Where:

- n is the total number of fixations
- FI_j is the number of self-intersections in fixation j, computed by counting how many times line segments formed by consecutive gaze points cross one another in 2D space.

Both the mean and standard deviation of FIC across all valid fixations were included in the final feature set.

Fixation Fractal Dimension (FFD)

The FFD estimates the spatial complexity of gaze paths during fixations using the box-counting method, a common technique for calculating fractal dimensions. Each fixation is first converted into a binary 2D grid in which each cell visited by the gaze is marked as 'True'. Grids of varying sizes are overlaid, and the number of occupied boxes is counted for each grid size.

The FFD is computed as:

$$FFD = \frac{1}{n}\sum_{j=1}^{n} FD_j$$

Where:

- n is the total number of fixations
- FD_j is the fractal dimension of the gaze path during fixation j, estimated using the box-counting method.

Higher FFD values indicate more irregular, space-filling, or scattered gaze trajectories, while lower values suggest more linear or constrained fixations. This feature captures the micro-patterns of gaze behaviour that may not be apparent in traditional fixation metrics.

Appendix D

Table D1

Best Parameters for Classification Models (RandomizedSearchCV)

Model	Feature Set	Parameters
KNN	All	metric: manhattan, n_neighbors: 16, weights: uniform
	AOI	metric: euclidean, n_neighbors: 17, weights: uniform
	Basic	metric: euclidean, n_neighbors: 23, weights: uniform
	Blinks	metric: euclidean, n_neighbors: 12, weights: distance
	Fixations	metric: euclidean, n_neighbors: 6, weights: distance
	Latency	metric: euclidean, n_neighbors: 4, weights: distance
	Most Important	metric: euclidean, n_neighbors: 26, weights: distance
	Saccades	metric: euclidean, n_neighbors: 3, weights: uniform
	Significant	metric: euclidean, n_neighbors: 21, weights: uniform
	Spatial	metric: euclidean, n_neighbors: 3, weights: uniform
LR	All	C: np.float64(0.001559747953698376), penalty: l2, solver: saga
	AOI	C: np.float64(0.12173252504194051), penalty: l2, solver: saga
	Basic	C: np.float64(69.58780103230364), penalty: l2, solver: lbfgs
	Blinks	C: np.float64(9.877700294007917), penalty: l2, solver: saga
	Fixations	C: np.float64(69.58780103230364), penalty: l2, solver: lbfgs
	Latency	C: np.float64(83.4298801304735), penalty: l2, solver: saga
	Most Important	C: np.float64(0.1.7718847354806828), penalty: l2, solver: saga
	Saccades	C: np.float64(69.58780103230364), penalty: l2, solver: lbfgs
	Significant	C: np.float64(0.1408146893930583), penalty: l2, solver: lbfgs
	Spatial	C: np.float64(69.58780103230364), penalty: l2, solver: lbfgs
SVM	All	C: np.float64(5.573452302583897), gamma: np.float64(0.0007707278591463597), kernel: rbf
	AOI	C: np.float64(276.03912956530013), gamma: np.float64(0.000655234487829567), kernel: rbf
	Basic	C: np.float64(69.92636148959322), gamma: np.float64(0.04878360603452144), kernel: rbf
	Blinks	C: np.float64(69.92636148959322), gamma: np.float64(0.04878360603452144), kernel: rbf
	Fixations	C: np.float64(793.2047656808546), gamma: np.float64(0.0025135566617708314), kernel: rbf
	Latency	C: np.float64(69.92636148959322), gamma: np.float64(0.04878360603452144), kernel: rbf

Model	Feature Set	Parameters
	Most Important	C: np.float64(23.66154006460317), gamma: np.float64(0.020597335357437203), kernel: rbf
	Saccades	C: np.float64(276.03912956530013), gamma: np.float64(0.000655234487829567), kernel: rbf
	Significant	C: np.float64(51.41096648805744), gamma: np.float64(0.0003972110727381913), kernel: rbf
	Spatial	C: np.float64(276.03912956530013), gamma: np.float64(0.000655234487829567), kernel: rbf
RF	All	max_depth: 20, min_samples_leaf: 3, min_samples_split: 4, n_estimators: 128
	AOI	max_depth: 20, min_samples_leaf: 4, min_samples_split: 6, n_estimators: 181
	Basic	max_depth: 10, min_samples_leaf: 3, min_samples_split: 4, n_estimators: 187
	Blinks	max_depth: 15, min_samples_leaf: 1, min_samples_split: 5, n_estimators: 113
	Fixations	max_depth: 15, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 157
	Latency	max_depth: None, min_samples_leaf: 1, min_samples_split: 7, n_estimators: 127
	Most Important	max_depth: 15, min_samples_leaf: 1, min_samples_split: 5, n_estimators: 113
	Saccades	max_depth: 10, min_samples_leaf: 2, min_samples_split: 6, n_estimators: 101
	Significant	max_depth: 15, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 157
	Spatial	max_depth: 15, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 256

Appendix E

Table F1

Correlations Between Features and RAN Score

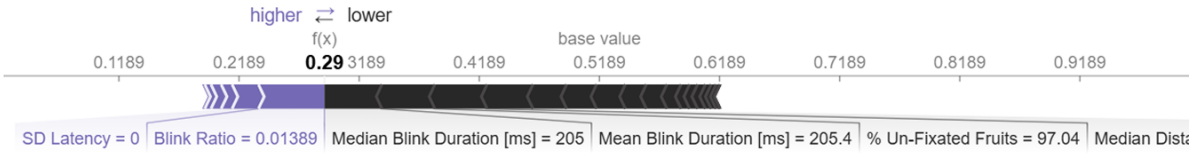
Variable	*r*	*p*
Age	0.79***	< .001
Mean Saccade Length	-0.55***	< .001
Percent Unseen Fruits	-0.45**	.001
Median Saccade Length	-0.42**	.001
Number of Blinks per Minute	-0.40**	.002
Mean Fixation Distance to Object	-0.36**	.007
Median Blink Duration	0.36**	.006
Mean Blink Duration	0.35**	.007
Mean Fixation Distance to Fruit	-0.34*	.011
Fixation Fractal Dimension (FFD)	-0.32*	.015
Median Fixation Distance to Object	-0.28*	.035
SD FIC	0.27*	.044
Median Fixation Distance to Fruit	-0.25	*ns*
Median Latency	-0.24	*ns*
Mean FIC	0.22	*ns*
Maximum Latency	0.19	*ns*
SD Latency	0.17	*ns*
Number of Fixations per Minute	0.17	*ns*
Number of Saccades per Minute	0.16	*ns*
Mean Fixation Duration	-0.14	*ns*
Blink Ratio	0.11	*ns*
Mean Latency	0.09	*ns*
Percent of Fixations Outside the Bounding Boxes	0.04	*ns*
Median Fixation Duration	0.00	*ns*

Note. *r* = Pearson correlation coefficient. *p* values < .05 are considered statistically significant. $p < .05$ = *, $p < .01$ = **, $p < .001$ = ***.

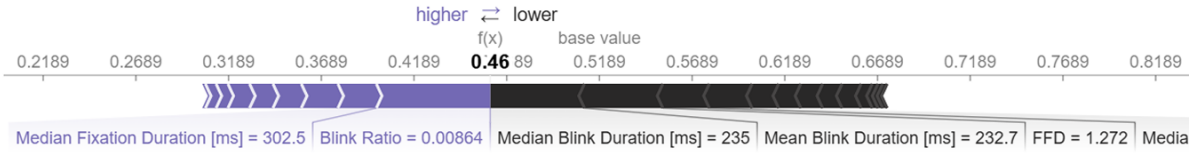
Appendix F

Individual Feature Contributions to SHAP Scores

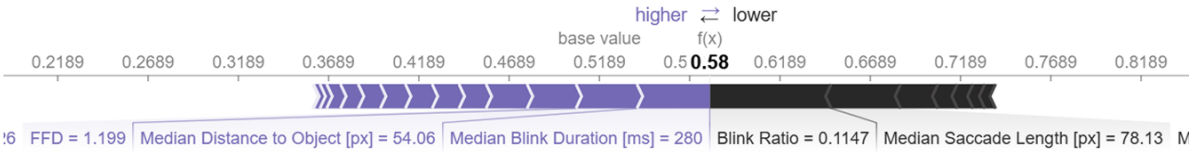
Participant 1 - Class 1 SHAP Force Plot



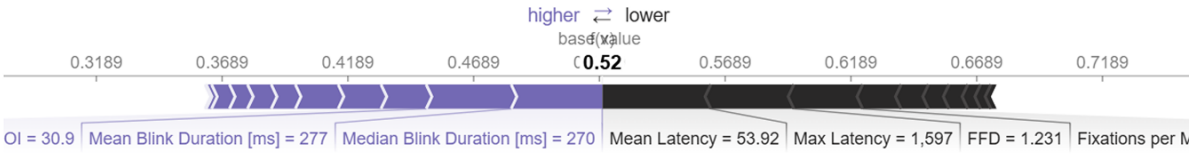
Participant 2 - Class 1 SHAP Force Plot



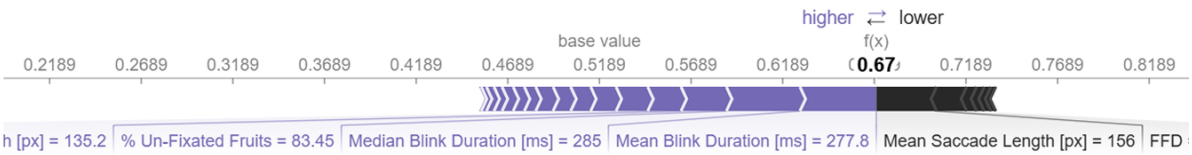
Participant 3 - Class 1 SHAP Force Plot



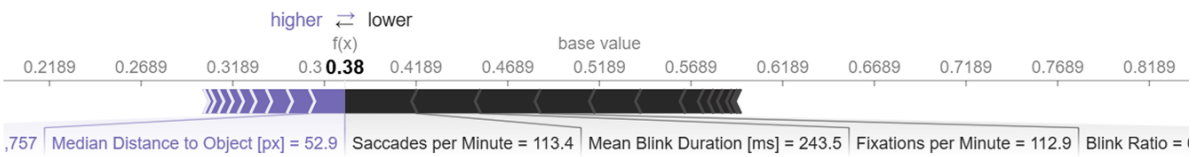
Participant 4 - Class 1 SHAP Force Plot



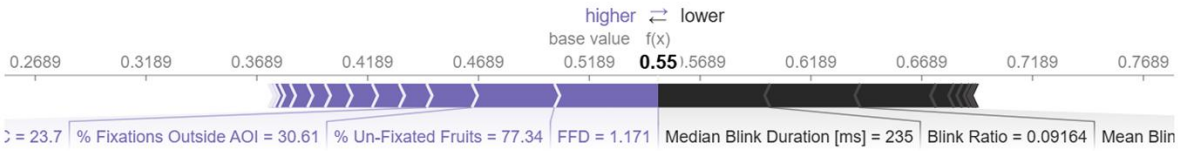
Participant 5 - Class 1 SHAP Force Plot



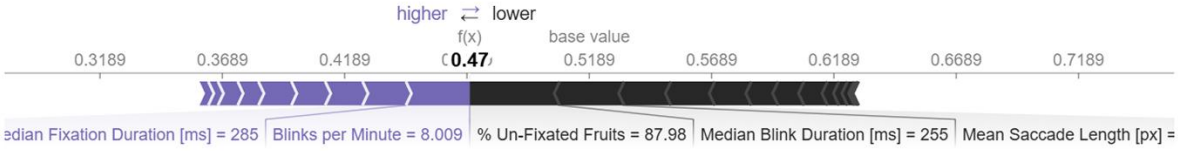
Participant 6 - Class 1 SHAP Force Plot



Participant 7 - Class 1 SHAP Force Plot

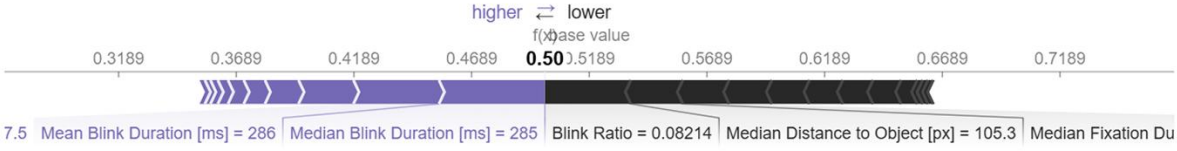


Participant 8 - Class 1 SHAP Force Plot

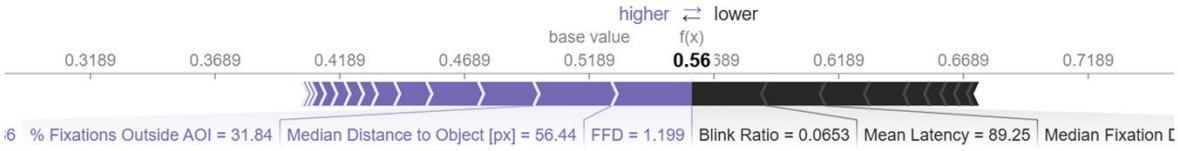


Participant 9 - Class 1 SHAP Force Plot

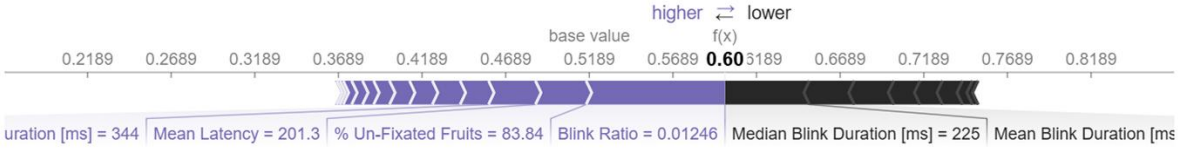
Participant 10 - Class 1 SHAP Force Plot



Participant 11 - Class 1 SHAP Force Plot



Participant 12 - Class 1 SHAP Force Plot



Appendix G

Table G1

Prediction Model Results

Feature Set	R^2 Mean	R^2 Std	MAE	MSE	RMSE	R^2 *p*
AOI	-0.13	0.55	0.51	0.40	0.63	0.66
All	-8.56	17.71	0.57	2.47	1.57	0.39
Basic	**0.30**	**0.13**	**0.41**	**0.25**	**0.50**	**0.01**
Blinks	-0.03	0.04	0.51	0.37	0.61	0.20
Fixation	0.07	0.23	0.50	0.33	0.58	0.56
Latency	-0.08	0.19	0.55	0.38	0.62	0.46
Most Important	0.14	0.18	0.47	0.31	0.55	0.20
Saccade	**0.26**	**0.17**	**0.43**	**0.27**	**0.52**	**0.04**
Significant	-0.05	1.11	0.38	0.32	0.57	0.93
Spatial	0.32	0.24	0.39	0.24	0.49	0.06

Note. R^2 Mean and R^2 Std represent the mean and standard deviation of the coefficient of determination across cross-validation folds. MAE = Mean Absolute Error; MSE = Mean Squared Error; RMSE = Root Mean Squared Error; R^2 *p* = *p*-value from permutation testing of R^2. Bolded rows indicate statistically significant models ($p < .05$).

Appendix H

Table H1

Classification Model Results

	Features	Acc	$Prec_p$	Rec_p	$F1_p$	$Prec_g$	Rec_g	$F1_g$	Mac F1	Wtd F1	*p*
KNN	AOI	0.49	0.48	0.39	0.43	0.50	0.59	0.54	0.49	0.49	0.60
	All	0.44	0.43	0.46	0.45	0.44	0.41	0.43	0.44	0.44	0.86
	Basic	0.49	0.45	0.18	0.26	0.50	0.79	0.61	0.43	0.44	0.60
	Blinks	**0.63**	**0.62**	**0.64**	**0.63**	**0.64**	**0.62**	**0.63**	**0.63**	**0.63**	**0.03**
	Fixation	0.56	0.55	0.61	0.58	0.58	0.52	0.55	0.56	0.56	0.21
	Latency	0.47	0.47	0.54	0.50	0.48	0.41	0.44	0.47	0.47	0.70
	Most Important	**0.65**	**0.65**	**0.61**	**0.63**	**0.65**	**0.69**	**0.67**	**0.65**	**0.65**	**0.02**
	Saccade	0.44	0.43	0.43	0.43	0.45	0.45	0.45	0.44	0.44	0.86
	Significant	0.46	0.41	0.25	0.31	0.48	0.66	0.55	0.43	0.43	0.79
	Spatial	0.60	0.59	0.57	0.58	0.60	0.62	0.61	0.60	0.60	0.09
LR	AOI	0.58	0.60	0.43	0.50	0.57	0.72	0.64	0.57	0.57	0.14
	All	0.28	0.12	0.07	0.09	0.35	0.48	0.41	0.25	0.25	1.00
	Basic	0.46	0.44	0.39	0.42	0.47	0.52	0.49	0.45	0.45	0.79
	Blinks	0.54	0.54	0.46	0.50	0.55	0.62	0.58	0.54	0.54	0.30
	Fixation	**0.63**	**0.62**	**0.64**	**0.63**	**0.64**	**0.62**	**0.63**	**0.63**	**0.63**	**0.03**
	Latency	0.51	0.50	0.50	0.50	0.52	0.52	0.52	0.51	0.51	0.50
	Most Important	0.54	0.54	0.54	0.54	0.55	0.55	0.55	0.54	0.54	0.30
	Saccade	0.47	0.46	0.39	0.42	0.48	0.55	0.52	0.47	0.47	0.70
	Significant	0.51	0.50	0.46	0.48	0.52	0.55	0.53	0.51	0.51	0.50
	Spatial	0.56	0.56	0.50	0.53	0.56	0.62	0.59	0.56	0.56	0.21

	Features	Acc	$Prec_p$	Rec_p	$F1_p$	$Prec_g$	Rec_g	$F1_g$	Mac F1	Wtd F1	*p*
RF	AOI	0.37	0.33	0.29	0.31	0.39	0.45	0.42	0.36	0.36	0.98
	All	0.39	0.35	0.29	0.31	0.41	0.48	0.44	0.38	0.38	0.97
	Basic	0.32	0.30	0.29	0.29	0.33	0.34	0.34	0.31	0.32	1.00
	Blinks	0.60	0.59	0.61	0.60	0.61	0.59	0.60	0.60	0.60	0.09
	Fixation	0.54	0.53	0.57	0.55	0.56	0.52	0.54	0.54	0.54	0.30
	Latency	0.51	0.50	0.46	0.48	0.52	0.55	0.53	0.51	0.51	0.50
	Most Important	0.60	0.60	0.54	0.57	0.59	0.66	0.62	0.59	0.59	0.09
	Saccade	0.40	0.38	0.32	0.35	0.42	0.48	0.45	0.40	0.40	0.94
	Significant	0.46	0.43	0.36	0.39	0.47	0.55	0.51	0.45	0.45	0.79
	Spatial	0.53	0.52	0.50	0.51	0.53	0.55	0.54	0.53	0.53	0.40
SVM	AOI	0.47	0.43	0.21	0.29	0.49	0.72	0.58	0.43	0.44	0.70
	All	0.49	0.40	0.07	0.12	0.50	0.90	0.64	0.38	0.39	0.60
	Basic	0.35	0.29	0.21	0.24	0.39	0.48	0.43	0.34	0.34	0.99
	Blinks	0.51	0.50	0.50	0.50	0.52	0.52	0.52	0.51	0.51	0.50
	Fixation	**0.65**	**0.63**	**0.68**	**0.66**	**0.67**	**0.62**	**0.64**	**0.65**	**0.65**	**0.02**
	Latency	0.60	0.62	0.46	0.53	0.58	0.72	0.65	0.59	0.59	0.09
	Most Important	**0.67**	**0.71**	**0.54**	**0.61**	**0.64**	**0.79**	**0.71**	**0.66**	**0.66**	**0.01**
	Saccade	0.54	0.57	0.29	0.38	0.53	0.79	0.64	0.51	0.51	0.30
	Significant	0.58	0.62	0.36	0.45	0.56	0.79	0.66	0.56	0.56	0.14
	Spatial	0.51	0.50	0.46	0.48	0.52	0.55	0.53	0.51	0.51	0.50

Note. Acc = Accuracy; $Prec_p$ = Precision ("Poor"); Rec_p = Recall ("Poor"); $F1_p$ = F1 score ("Poor"); $Prec_g$ = Precision ("Good"); Rec_g = Recall ("Good"); $F1_g$ = F1 score ("Good"); Mac F1 = Macro-averaged F1; Wtd F1 = Weighted F1; *p* = Binomial test p-value. Bolded rows indicate statistically significant models ($p < .05$)

Appendix I

Table I1

Mean Classification Results Across Feature Sets

Feature Set	Acc	$Prec_p$	Rec_p	$F1_p$	$Prec_g$	Rec_g	$F1_g$	Mac F1	Wtd F1
AOI	0.48	0.46	0.33	0.38	0.49	0.62	0.54	0.46	0.46
All	0.40	0.32	0.22	0.24	0.42	0.57	0.48	0.36	0.36
Basic	0.40	0.37	0.27	0.30	0.42	0.53	0.47	0.38	0.39
Blinks	0.57	0.56	0.55	0.56	0.58	0.59	0.58	0.57	0.57
Fixation	0.60	0.58	0.62	0.60	0.61	0.57	0.59	0.60	0.60
Latency	0.52	0.52	0.49	0.50	0.52	0.55	0.54	0.52	0.52
Most Important	0.62	0.62	0.56	0.59	0.61	0.67	0.64	0.61	0.61
Saccade	0.46	0.46	0.36	0.40	0.47	0.57	0.52	0.46	0.46
Significant	0.50	0.49	0.36	0.41	0.51	0.64	0.56	0.49	0.49
Spatial	0.55	0.54	0.51	0.52	0.55	0.58	0.57	0.55	0.55

Note. Acc = Accuracy; $Prec_p$ = Precision (“Poor”); Rec_p = Recall (“Poor”); $F1_p$ = F1 score (“Poor”); $Prec_g$ = Precision (“Good”); Rec_g = Recall (“Good”); $F1_g$ = F1 score (“Good”); Mac F1 = Macro-averaged F1; Wtd F1 = Weighted F1

Statement of Independence

I hereby certify that I have written this Master thesis independently without the help of third parties and without using any sources or aids other than those indicated.

Olivia Lecomte

4 September 2025

Use of AI Tools

During the preparation of this thesis, AI tools (ChatGPT, GitHub Copilot) were used to assist with language clarity and code debugging. All scientific content, ideas, interpretations, and analyses were independently developed and verified by me.