Developing a Dyslexia Indicator Using Eye Tracking

Kevin Cogan[1‡], Vuong M. Ngo[2 ✉‡], and Mark Roantree[1]

[1] Insight Centre, School of Computing, Dublin City University, Ireland
[2] Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam
kevin.s.cogan@gmail.com, vuong.nm@ou.edu.vn, mark.roantree@dcu.ie

‡These authors designated as co-first authors.

Abstract. Dyslexia, which affects 10% to 20% of the global population, poses significant challenges to learning, underscoring the need for accessible diagnostic tools. This study explores the use of eye-tracking technology combined with machine learning as a cost-effective and non-invasive approach for early dyslexia detection. By analyzing key eye movement patterns—such as prolonged fixations and erratic saccades—we proposed an enhanced feature framework and achieved 88.58% accuracy using a Random Forest Classifier. Hierarchical clustering was also applied to uncover varying dyslexia severity levels. The results, validated across diverse populations and settings, highlight the method's scalability and potential for identifying even borderline dyslexia traits, offering a promising advancement in clinical diagnostics.

Keywords: Reading Difficulties · Early Diagnosis · Machine Learning

1 Introduction

Dyslexia is a specific learning disability characterized by difficulties in word recognition, spelling, and decoding, often impairing reading fluency and comprehension [6]. It is linked to atypical brain function, particularly in regions like the left temporal-parietal cortex involved in phonological processing. Early diagnosis is essential but often limited by high costs, specialist shortages, and reliance on teacher judgment. Eye-tracking technology offers a non-invasive alternative by capturing distinct eye movement patterns—such as longer fixations, frequent short saccades, and increased regressions—common among dyslexic readers [7].

Recent studies have leveraged eye-tracking and machine learning to advance dyslexia detection. Franzen et al. [1] analyzed eye movements during standardized reading tasks using IReST texts, revealing distinct visual sampling strategies in dyslexic readers. Raatikainen et al. [5] applied Random Forest for feature selection and SVM for classification on eye-tracking data from 165 participants, demonstrating effective dimensionality reduction and dyslexia identification. Vajs et al. [8] introduced a novel approach by converting gaze data into 2D color-coded time series graphs for input into a VGG16 model, enhancing classification accuracy. Nerusil et al. [3] employed CNNs to process raw eye-tracking

data from 185 subjects with minimal preprocessing, analyzing signals in both time and frequency domains for robust dyslexia detection. Lastly, Smyrnakis et al. [7] combined traditional and novel features, such as Fixation Intersection Coefficient and Fixation Fractal Dimension, with multiple machine learning classifiers to capture complex gaze dynamics and enrich the understanding of dyslexic reading behavior.

In summary, the reviewed studies lacked a sufficiently comprehensive set of eye-tracking-based dyslexia features, which may have limited their classification performance. Moreover, none explored correlations between features and clusters—an overlooked step that could reveal meaningful relationships between eye-tracking metrics and dyslexia severity. Incorporating a broader feature set alongside correlation analysis has the potential to enhance model accuracy and offer deeper insights into the underlying patterns of dyslexia.

Contribution. This paper explores eye-tracking technology as an innovative, non-invasive, and cost-effective approach to dyslexia detection. By capturing precise metrics such as saccades and fixations, we propose an enhanced framework for developing eye-tracking-based dyslexia features. Leveraging machine learning techniques, including Random Forest and Agglomerative Hierarchical Clustering, the study analyzes dyslexia-related patterns to evaluate accuracy and effectiveness. It also highlights the scalability and adaptability of eye-tracking across diverse populations and educational settings, underscoring its potential for large-scale screening. Overall, this work supports the integration of advanced technologies to enable more inclusive, timely, and efficient dyslexia assessments.

2 Feature Engineering

Dataset Description: The Provo Corpus (78MB) contains detailed eye-tracking data from 84 native English speakers at Brigham Young University [2], with 230,412 entries across 63 features. Recorded using the EyeLink 1000 Plus, it captures fixations, saccades, and regressions during natural reading. The dataset includes Predictability Norms from a Qualtrics survey of 470 participants and Eye-Tracking Texts based on 55 short passages. Words (2,689 total; 1,197 unique) were POS-tagged with CLAWS and analyzed via Latent Semantic Analysis, supporting rich exploration of reading behavior.

Data Processing: To streamline the model and enhance performance, redundant columns with high missing rates (e.g., `Ia_Regression_Path_Duration`) and highly correlated features (e.g., `Ia_First_Saccade_Start_Time`) were removed. Missing values were handled using advanced techniques: the MissForest algorithm imputed missing `Ia_Skip` values, and sequential fields like `Sentence_Number` and `Word_In_Sentence_Number` were forward-filled. Semi-structured data containing both categorical and continuous variables were integrated. A new feature, `Saccade_Duration`, was created by calculating the time difference between `Ia_First_Saccade_End_Time` and `Ia_First_Saccade_Start_Time`. Additionally, the `Word_Cleaned` column was transformed using the TF-IDF technique.

Basic Dyslexia Features: Building on prior research, our study examines key eye-tracking metrics to distinguish individuals with dyslexia from non-dyslexic readers. Using statistical analysis and visualizations across percentiles, we identify significant patterns that reveal cognitive and behavioral traits associated with reading difficulties. Metrics such as `Ia_First_Saccade_Amplitude`, `Ia_Dwell_Time`, `Ia_Regression_In_Count`, `Ia_Regression_Out_Count`, `Ia_Fixation_Count` and `Saccade_Duration` highlight inefficiencies in text scanning, increased cognitive effort, and disrupted reading fluency among dyslexic readers. To label reading difficulty levels, we applied the 95th percentile threshold for each metric—assigning a value of 1 to entries above the threshold (indicating higher difficulty) and 0 to those below—recorded under the column `Reading_Difficulties`. This approach provides both statistical and intuitive insights into reading behavior and dyslexia-related patterns.

Enhancing Dyslexia Features: To identify key contributors to the `Reading_Difficulty` feature, a RandomForestClassifier was applied to a balanced dataset, where the majority class was resampled to address class imbalance and ensure fair model evaluation. To prevent data leakage, features directly used in constructing the `Reading_Difficulty` target—such as `Ia_First_Saccade_Amplitude`, `Ia_Dwell_Time`, and `Ia_Fixation_Count`—were excluded. Bayesian hyperparameter tuning optimized the model, and feature importance was visualized via a bar plot to highlight influential predictors. The most impactful features included: `Ia_Regression_In_Count`, `Ia_First_Run_Fixation_.`, `Saccade_Duration`, `Ia_Right`, `Ia_First_Fixation_Time`, `Ia_Regression_Out_Count`, `Ia_First_Fixation_Index`, `Ia_First_Fixation_Y`, `Ia_First_Fixation_X`, `Word_Number`, `Ia_Skip`, and `Word_Length`. Together with the basic features above, these form an enhanced set of eye-tracking-based dyslexia features to improve the detection and understanding of reading difficulties.

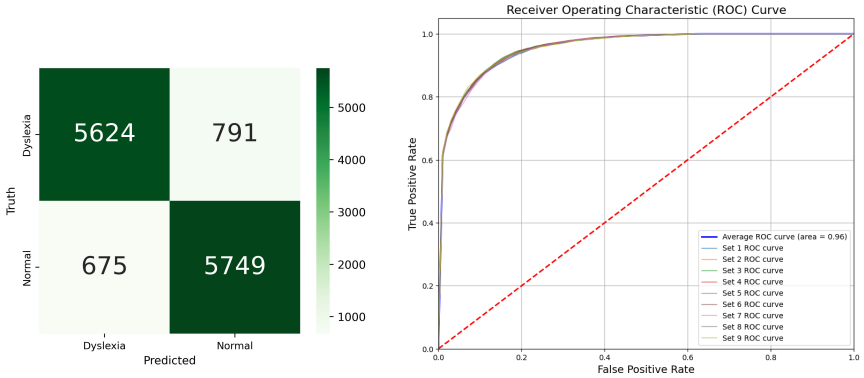
3 Algorithm and Analysis

3.1 Classification

Random Forest Classifier: builds an ensemble of decision trees, each trained on a bootstrap sample of the data. At each split, a random subset of features is considered to promote diversity and reduce overfitting. Trees make independent predictions, and final classification is determined by majority vote. Hyperparameters are optimized using BayesSearchCV, focusing on features selected via forward selection to maximize cross-validation performance. The refined model, trained on key features, is benchmarked against a version using all features, with both employing tuned hyperparameters. To enhance robustness and address class imbalance, training was performed on multiple resampled subsets of the data.

Evaluation and Discussion: The Random Forest Classifier was evaluated using 9-fold cross-validation to ensure robustness and generalizability across different data subsets. This approach established a reliable performance baseline for real-world applications. The average results show strong and consistent metrics: 88.58% accuracy, 87.91% precision, 89.49% recall, and an F1-score of 88.67%.

Additionally, Figure 1(a) presents the average confusion matrix, highlighting the Random Forest Classifier's consistent balance between sensitivity and specificity across validation folds. This reliability makes it suitable for high-stakes tasks. The average ROC curve in Figure 1(b) shows an AUC of 0.96, indicating excellent class discrimination. Consistent ROC curves across folds further confirm the model's robustness and applicability, especially for tasks like diagnosing reading difficulties.



(a) Average confusion matrix (b) ROC curves

Fig. 1. Experimental Results on 9-Fold Cross-Validation

3.2 Cluster

Agglomerative Hierarchical Clustering (AHC) is a bottom-up method where each data point begins as its own cluster, and pairs are merged iteratively using Euclidean distance and Ward linkage [4]. It requires no predefined number of clusters, making it ideal for exploratory analysis. Its interpretability and ability to capture nested structures make it useful in domains like bioinformatics and text analysis. Principal Component Analysis reduces the data to two components for clear 2D cluster visualization.

Figure 2 presents clustering results based on enhanced eye-tracking features, revealing three

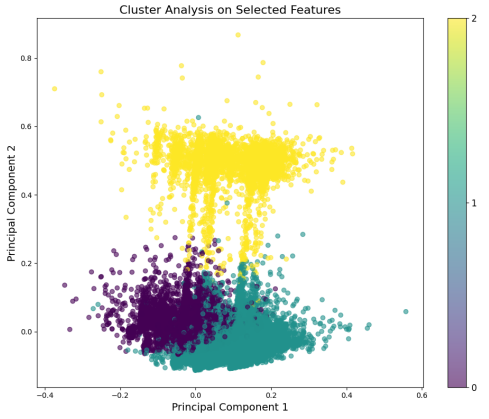


Fig. 2. Cluster Analysis on Our Enhancing Dyslexia Features

distinct groups. Cluster 0, with low dwell times, short saccades, and moderate regression counts, represents proficient readers exhibiting efficient eye movements. Cluster 1 shows higher dwell times, lower regression counts, and average saccade durations, indicating moderately proficient readers with occasional difficulties. Cluster 2, characterized by high saccade amplitudes, increased dwell times, and longer saccade durations, suggests poor readers, possibly including individuals with dyslexia, who experience significant challenges in processing text. These findings highlight a spectrum of reading abilities, from proficient to poor readers, emphasizing the potential of eye-tracking features in diagnosing dyslexia and developing targeted interventions.

4 Conclusion

The paper presents a non-invasive, objective, and cost-effective method for detecting dyslexia using eye-tracking data and machine learning, focusing on metrics such as saccade amplitude, dwell time, and fixation count. At the 95th percentile, the approach minimizes reliance on subjective teacher assessments and enables early identification of students needing support. A Random Forest classifier, optimized through hyperparameter tuning and selection of the top 12 features, achieved strong performance while streamlining data collection. Clustering analysis revealed diverse reading behaviors—from proficient to struggling readers, including those with dyslexic traits—emphasizing the need for further research into personalized interventions based on shared group characteristics.

Acknowledgment

This research has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland under Grant number 12/RC/2289_P2.

References

1. Franzen, L., Stark, Z., Johnson, A.: Individuals with dyslexia use a different visual sampling strategy to read text. Scientific Reports **11**(1) (2021)
2. Luke, S., Christianson, K.: The provo corpus: A large eye-tracking corpus with predictability norms. Behavior Research Methods **50**(2), 826–833 (2017)
3. Nerušil, B., Polec, J., Škunda, J., Kačur, J.: Eye tracking based dyslexia detection using a holistic approach. Scientific Reports **11**(1) (2021)
4. Ngo, V.M., Helmer, S., Le-Khac, N.A., Kechadi, M.T.: Structural textile pattern recognition and processing based on hypergraphs. Information Retrieval Journal **24**(2), 137–173 (2021)
5. Raatikainen, P., et al.: Detection of developmental dyslexia with machine learning using eye movement data. Array **12** (2021)
6. Reid, G.: Dyslexia: A Practitioner's Handbook, 5th Edition. Wiley-Blackwell (2016)
7. Smyrnakis, I., et al.: Silent versus reading out loud modes: An eye-tracking study. Journal of Eye Movement Research **14**(2) (2021)
8. Vajs, I., et al.: Dyslexia detection in children using eye tracking data based on vgg16 network. In: 2022 30th European Signal Processing Conference (EUSIPCO) (2022)