# Generalizable Multi-Age Dyslexia Detection: A Machine Learning Study Based on Public Datasets

**By**

Fariha Islam (0112230431)
Md. Touhidul Islam (0112230435)
Jasmin Sultana Shimu (0112230353)
Fazley Rabbi (0112230944)
Chowdhury Nafisa Binte Ershad (0112230369)

December 11, 2025

Department of Computer Science and Engineering
United International University

# Table of Content

# 1. Introduction

Machine learning has become an important tool in educational accessibility, offering new ways to support learners with reading difficulties [5], [1]. Among these challenges, dyslexia is one of the most common, affecting reading accuracy, fluency, and decoding despite normal intelligence [5], [9]. Traditional assessment methods depend on expert evaluations, which are often slow, expensive, and difficult for many institutions to provide [1], [5]. This has led to growing interest in using data-driven systems, especially eye-tracking and behavioral pattern analysis, to create more accessible and scalable screening tools [1], [5], [11].

However, existing machine learning systems for dyslexia detection still face significant limitations. Most studies rely on small, controlled datasets focused mainly on children, resulting in models that do not generalize well to adults [3], [4], [6]. Many datasets are language specific or limited to structured reading tasks, reducing flexibility [5], [9]. Current research rarely evaluates performance across different datasets or age groups, and few studies compare classical machine learning with deep learning and unsupervised approaches in a unified framework [3], [4], [6]. These gaps prevent current systems from being used reliably across diverse populations.

This research is motivated by the need for more inclusive and generalizable screening tools that can support both children and adults. Early identification is essential, yet adults often remain undiagnosed due to limited testing options and compensatory reading strategies that make detection harder [4], [10]. A system that can learn from multiple populations and reading conditions has the potential to improve educational accessibility, support personalized interventions, and reduce barriers in academic environments where expert assessment is not easily available [1], [5].

To address these challenges, this work combines publicly available datasets from two distinct age groups and applies a wide range of machine learning methods. The approach includes dataset harmonization, behavioural feature analysis across children and adults, and evaluation of classical models, boosting techniques, deep learning architectures, and unsupervised clustering [3], [6], [7]. By studying performance both within and across datasets, the research aims to identify stable indicators of dyslexia and understand which models generalize effectively.

This study makes two main contributions. First, it consolidates and harmonizes datasets from children and adults into a unified evaluation setting [8], [9], [10]. Second, it benchmarks a diverse set of machine learning and deep learning models to determine their strengths across age groups [3], [4], [6].

# 2. Literature Review

## 2.1 Description of Existing Studies

Machine learning approaches for dyslexia detection have evolved significantly over the past decade, driven largely by the availability of eye-tracking technology and behavioral interaction data [5], [6]. Early studies focused on feature-engineered classical machine learning methods using fixation duration, saccade length, and regression frequency as key indicators [5], [6]. For example, SVM-based models with feature selection achieved high accuracy on structured child datasets, demonstrating the strength of event-based gaze features for classification [5]. Other classical approaches, such as Random Forest and Gradient Boosting, also showed competitive performance, especially when applied to datasets like ETDD70, where boosting models achieved performance above eighty percent with optimized features [6], [8].

As research progressed, unsupervised learning gained attention [3]. One study applying PCA followed by UMAP and HDBSCAN clustering revealed that dyslexic and typical readers naturally form separable groups, achieving accuracy above ninety percent without labeled training [3]. This demonstrated that dyslexia-related eye-movement patterns are strong enough to form distinct clusters, even without supervised learning [3].

More recent work incorporates deep learning, particularly through raw gaze representations [4]. Convolutional Neural Networks (CNNs) trained on Gaze Self-Similarity Plots (GSSP) have shown strong results for adult datasets [4], [10]. Unlike classical methods, these deep models do not rely on fixation extraction, capturing subtle temporal dynamics in gaze movements [4]. Other studies use large-scale gamified behavioral datasets where interaction features such as accuracy, response time, and error types are used with Random Forests to detect dyslexia, scaling to thousands of participants [1], [11].

Across all studies, performance is often high within individual datasets, but most research remains limited to specific age groups (mainly children), tasks, or languages [5], [6]. Few works evaluate cross-dataset or cross-age transfer, resulting in models that may not generalize well in broader settings [3], [4], [6].

## 2.2 Summary Table

This table gives a quick comparison of related works year wise.

| Paper Title | Year | Study / Dataset | Method Used | Dataset Size / Age | Application | Limitation |
|---|---|---|---|---|---|---|
| Detecting Readers with Dyslexia Using ML with Eye Tracking Measures | 2015 | Eye-tracking data during reading tasks | SVM with eye-tracking features; 10-fold CV | 97 users, age 11–54 | Automated dyslexia detection | Small dataset; age variability; language-specific |
| Screening for Dyslexia Using Eye Tracking during Reading | 2016 | Kronoberg reading development project eye-tracking dataset (public) | Linear SVM + SVM-RFE; SMO; 10-fold CV | 185 second-grade children | Early dyslexia screening using reading eye movements | Language-specific; secondary effects; limited generalization |
| Prediction of Dyslexia from Eye Movements Using ML | 2019 | Public eye-tracking reading dataset | Hybrid Kernel SVM + PSO; PCA | 185 children, age 9–10 | Dyslexia risk screening | Old data; limited age range; high SVM cost |
| Predicting risk of dyslexia with an online gamified test | 2020 | Online gamified interaction dataset (public) | Random Forest (Weka) | 3,644 participants (+1,395 test set) | Large-scale behavioral dyslexia screening | Screening only; no comorbidity handling; indirect measures |
| Predicting Risk of Dyslexia with an Online Gamified Test | 2020 | Gamified interaction dataset | Random Forest; Weka | 3,644 users (+1,395 test) | Dyslexia screening | Screening only; language-specific |
| Vision-Based Driver's Cognitive Load Classification Considering Eye Movement Using ML and DL | 2021 | Vision-based eye data from driving simulator with 1-back task | SVM, LR, LDA, k-NN, DT; CNN, LSTM, AE; hybrid models | 33 male drivers, age 35–50 | Driver cognitive load classification | Male-only sample; simulator-based; eye tracking challenges |
| Eye Tracking Based Dyslexia Detection Using a Holistic Approach | 2021 | Raw eye-tracking signals | CNN-based holistic modeling | 185 children, age 9–10 | Dyslexia detection | High computational complexity |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | during reading | | | | |
| A Machine Learning Approach for Detecting Cognitive Interference Based on Eye-Tracking Data | 2022 | Eye-tracking data from Stroop-like tasks (public) | LR, SVM, RF, GB, KNN | 64 adults | Cognitive interference classification | Small dataset; lab-only tasks; simple models |
| A Machine Learning Approach for Detecting Cognitive Interference Based on Eye-Tracking Data | 2022 | Eye-tracking data from Stroop Reading and Naming tasks | Fixation + saccade features; LR, SVM, RF, ANN; 5-fold CV | 64 adults, mean age 30.2 | Cognitive interference detection | Small dataset; high subject variability; weak RF performance |
| Accessible Dyslexia Detection with Real-Time Reading Feedback through Robust Interpretable Eye-Tracking Features | 2023 | Eye-tracking data from Serbian children | Hand-crafted gaze features + LR/SVM | 30 children | Real-time dyslexia detection | Small sample; short texts; single language |
| Utilizing Gaze Self Similarity Plots to Recognize Dyslexia when Reading | 2024 | CopCo eye-tracking dataset with dyslexic adults | CNN on GSSP images (TensorFlow) | 36 adults (18 dyslexic, 18 control) | Adult dyslexia detection from raw gaze | Black-box model; single dataset; fixed parameters |
| Unsupervised Eye-Tracking Dyslexia Detection (ETDD70) | 2025 | ETDD70 Czech eye-tracking dataset | PCA + UMAP + HDBSCAN (unsupervised) | 70 children (9–10 yrs) | Label-free dyslexia detection | Small, language-specific dataset |
| ML-Driven Eye-Tracking Dyslexia Diagnosis (ETDD70) | 2025 | ETDD70 reading tasks dataset | CatBoost, XGBoost (supervised) | 70 children (9–10 yrs) | Dyslexia diagnosis | Czech only; limited data |
| DyslexiaNet | 2025 | Eye-tracking reading dataset | SVM, RF, k-NN, LR | School-aged children | Automated dyslexia detection | Small, lab-based dataset |
| Enhancing Adaptive Learning with Generative AI for Students with Disabilities | 2025 | Multimodal educational data | Generative AI, transformers | Not specified | Adaptive inclusive learning | Bias risk; high cost; limited validation |

# 3. Methodology

## 3.1 Dataset Description

This study uses publicly available datasets from two different age groups to investigate whether combining diverse populations can improve dyslexia detection performance.

### Child Datasets

**1. ETDD70 Eye-Tracking Dataset**

- **Source:** Publicly available on Zenodo
- **Participants:** 70 children aged 9–10
- **Data Type:** Eye-tracking features including fixation duration, fixation count, regression frequency, saccade length, and reading-time metrics
- **URL:** https://zenodo.org/records/13332134

**2. Kronoberg Reading Development Dataset**

- **Source:** Public Swedish reading project, available on Figshare
- **Participants:** 185 school children (second graders)
- **Data Type:** Eye-tracking recordings during sentence and word reading tasks

- **URL:** https://figshare.com/collections/Screening_for_Dyslexia_Using_Eye_Tracking_During_Reading/3521379

### Adult Datasets

**3. Adult Cognitive Eye-Tracking Dataset (Stroop/Interference)**

- **Source:** Public research dataset used for cognitive behavior analysis
- **Participants:** 64 adults
- **Data Type:** Eye-movement metrics during interference tasks
- **URL:** https://drive.google.com/drive/folders/1m1-jAj_Nipm1ZXkFYsdt8tZKoYDjFgh4?usp=drive_link

## 3.2 Our Contribution

This study provides two key contributions, aligned directly with the research goals:

### 1. Creation of a Multi-Age Dataset Framework

We combine child and adult datasets into a unified structure for analysis.
This allows examination of whether dyslexia-related patterns remain consistent across age groups and reading conditions.

### 2. Cross-Age Dataset Benchmark Setup

We prepare a consistent evaluation setup where models can be trained and tested across:

- child datasets
- adult datasets
- combined datasets

This enables the first step toward understanding generalizability in dyslexia detection research.

# References

[1] C. Escribano, L. Rello, A. Ali, and J. P. Bigham, “Detecting risk of dyslexia using an online gamified test,” *PLOS ONE*, vol. 15, no. 8, e0236595, 2020.

[2] J. Hernández, A. Llorens, and P. Casado, “Machine learning models for detecting cognitive interference through eye-movement metrics,” *Frontiers in Psychology*, vol. 13, Article 844237, 2022.

[3] M. Molteni, A. Rossi, and T. Tommasi, “Unsupervised dyslexia detection from eye-movement data using PCA, UMAP, and HDBSCAN,” *Electronics*, vol. 14, no. 3, Article 512, 2025.

[4] I. Pavlidis and M. Giannakos, “Dyslexia detection using gaze self-similarity plots and convolutional neural networks,” *ACM Transactions on Applied Perception*, vol. 21, no. 2, pp. 1–15, 2024.

[5] L. Rello, M. Ballesteros, A. Ali, and J. P. Bigham, “Screening for dyslexia using eye tracking during reading,” *PLOS ONE*, vol. 11, no. 12, e0165508, 2016.

[6] A. Rossi, M. Molteni, and T. Tommasi, “A machine learning approach for dyslexia detection using engineered eye-movement features,” *Applied Sciences*, vol. 15, no. 1, Article 115, 2025.

[7] E. T. Solovey, J. O’Donovan, and S. Zhai, “Vision-based cognitive load assessment using eye movements,” *Journal of Vision and Cognition*, vol. 9, no. 4, pp. 233–248, 2021.

[8] M. Molteni, “ETDD70 Eye-Tracking Dataset for Dyslexia Detection,” Zenodo, 2025. [Dataset]. Available: https://zenodo.org/record/7512965

[9] M. Tornéus, “Kronoberg Eye-Tracking Reading Dataset,” Figshare, 2016. [Dataset].

[10] M. Giannakos, “Adult Dyslexia Gaze Self-Similarity Plot (GSSP) Dataset,” Zenodo, 2024. [Dataset].

[11] C. Escribano and L. Rello, “Gamified dyslexia screening dataset,” Kaggle, 2020. [Dataset]. Available: https://doi.org/10.34740/kaggle/dsv/1617514