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 during reading	CNN-based holistic modeling	185 children, age 9–10	Dyslexia detection	High computational complexity
A Machine Learning Approach for Detecting Cognitive Interference Based on Eye-Tracking Data	2022	Eye-tracking data from Stroop-like tasks	LR, SVM, RF, GB, KNN	64 adults	Cognitive interference classification	Small dataset; lab-only tasks; simple

		(public)				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. 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.

4. Methodology

This section presents a comprehensive framework for cross-age dyslexia detection using eye-tracking data. Our approach addresses the fundamental challenge of model generalization across different populations, recording devices, and developmental stages through three complementary experiments combining supervised learning and unsupervised analysis.

4.1 Data Acquisition and Harmonization

We integrated three publicly available eye-tracking datasets representing distinct age groups and experimental paradigms to evaluate cross-population generalization.

4.1.1 Dataset Overview

- **Child Dataset 1 (ETDD70):** Sourced from Zenodo, this dataset contains fixation and saccade events from 66 Czech children (ages 9-10). It includes 31 dyslexic and 35 non-dyslexic subjects recorded at 500 Hz during reading comprehension tasks.

- **Child Dataset 2 (Kronoberg Reading Dataset):** Sourced from Figshare , this dataset includes 185 Swedish second-grade children recorded at ~50 Hz. Diagnostic labels were derived from participant IDs, differentiating between dyslexic and control groups performing sentence reading tasks.

- **Adult Dataset (Cognitive Eye-Tracking):** This dataset contains raw gaze coordinates from 64 adults (ages 18+) performing Stroop interference tasks. While lacking dyslexia labels, it serves as a critical control group for unsupervised cross-age feature analysis.

4.1.2 Multi-Dataset Integration Strategy

The integration of these heterogeneous datasets allows for the evaluation of model generalization across recording devices and protocols. Furthermore, it enables the assessment of feature stability across age groups, establishing whether age-specific or unified models are required for clinical deployment.

4.2 Feature Engineering

4.2.1 Event Detection (I-VT Algorithm)

To standardize feature extraction across datasets with varying formats and sampling rates, we implemented the Velocity-Threshold Identification (I-VT) algorithm. The algorithm computes inter-sample displacement and instantaneous velocity, classifying events based on an adaptive threshold ($\theta = \mu_v + 2\sigma_v$). This dynamic approach ensures robustness to hardware variations and noise levels inherent in different eye-tracking devices.

4.2.2 Domain-Invariant Features

We extracted three core features selected for their computational robustness and established associations with developmental dyslexia literature (Rayner, 1998; Reichle et al., 2003):

1. **Fixation Duration (Mean):** The average time the eye remains stationary (ms). Prolonged fixations typically indicate slower word recognition.

2. **Saccade Amplitude (Mean):** The average distance between fixation points. Shorter saccades often reflect fragmented reading patterns.

3. **Regression Count:** The frequency of backward eye movements, signaling comprehension difficulties and re-reading behavior.

4.2.3 Addressing Domain Shift: Quantile Mapping

A primary technical challenge in cross-dataset generalization was the significant discrepancy in spatial scales caused by varying screen resolutions (e.g., 1000px vs. 1920px). Conventional normalization techniques, such

as StandardScaler, proved insufficient for aligning these heterogeneous feature spaces.

To address this, we implemented **Quantile Mapping** (QuantileTransformer) to map the feature distributions of each independent dataset to a standard Gaussian (normal) distribution. This mechanism aligns the statistical distributions across datasets while strictly preserving the relative ranking and variance of individual data points. This domain adaptation step was critical for model performance; preliminary analysis indicated that its implementation increased cross-dataset accuracy from 20% to over 60%, effectively mitigating the impact of hardware-induced domain shifts.

4.3 Machine Learning Models and Pattern Recognition Algorithms

We employed a dual-approach modeling strategy, utilizing supervised classifiers for dyslexia detection and unsupervised algorithms for cross-age feature analysis.

4.3.1 Supervised Classifiers (Dyslexia Detection)

We selected three algorithms representing distinct inductive biases to benchmark detection performance:

- **Support Vector Machine (SVM):** Configured with an RBF kernel, C=1.0, and scaled gamma. SVMs were selected for their ability to maximize margin separation, providing robustness in high-dimensional spaces with limited sample sizes.
- **Random Forest:** Implemented with 100 trees and bootstrap aggregation. This ensemble method was chosen to reduce overfitting through variance averaging and to provide interpretable feature importance rankings.
- **XGBoost:** A gradient boosting classifier using default parameters with L1/L2 regularization and "logloss" evaluation metric. XGBoost was included for its state-of-the-art performance on tabular data and its ability to iteratively correct errors.

4.3.2 Unsupervised Algorithms (Cross-Age Analysis)

To investigate feature stability across age groups (Experiment III) without relying on diagnostic labels, we implemented three dimensionality reduction and clustering techniques:

- **Principal Component Analysis (PCA):** A linear dimensionality reduction technique used to project high-dimensional eye-tracking features into orthogonal components. This allows for the visualization of global variance and the calculation of Euclidean distances between child and adult population centroids.
- **UMAP (Uniform Manifold Approximation and Projection):** A non-linear manifold learning technique. Unlike PCA, UMAP preserves local neighborhood structures. We utilized this to verify if the

separation between age groups persists in non-linear projections, ensuring that differences are not artifacts of linearity.

HDBSCAN: A hierarchical density-based clustering algorithm. Unlike K-Means, HDBSCAN does not require specifying the number of clusters (k) a priori and can explicitly identify "noise" points. We used this to quantify how naturally the data separates into distinct "Child" and "Adult" clusters based purely on feature density.

4.4 Experimental Design

4.4.1 Experiment I: Intra-Dataset Baseline

This experiment established an upper-bound performance benchmark using the ETDD70 dataset. The data was split into an 80/20 train-test ratio with stratified sampling to preserve class distribution. Missing values were handled via mean imputation, and features were normalized using StandardScaler.

4.4.2 Experiment II: Cross-Dataset Generalization

To test zero-shot transfer capability, models were trained on the ETDD70 dataset and evaluated on the unseen Kronoberg dataset. To address the distributional shift caused by different eye-tracking hardware, we applied **Quantile Normalization**. This technique maps feature distributions from both datasets to a standard normal distribution (N(0,1)), forcing the models to rely on relative feature relationships rather than absolute hardware-dependent values.

4.4.3 Experiment III: Unsupervised Cross-Age Analysis

This experiment utilized unsupervised learning to determine if eye-movement patterns are age-invariant. We combined labeled child data (ETDD70) with unlabeled adult data (Cognitive Eye-Tracking). The pipeline involved dimensionality reduction via **Principal Component Analysis (PCA)** and **UMAP**, followed by **HDBSCAN** clustering to quantify the separation between child and adult feature clusters.

5. Results

This section presents the findings from our three-experiment framework evaluating dyslexia detection across datasets, devices, and age groups.

5.1 Experiment I: Intra-Dataset Classification (ETDD70)

5.1.1 Model Performance

Table 1 summarizes the classification performance of three machine learning models on the ETDD70 dataset using an 80/20 stratified split.

Table 1: Classification performance on the ETDD70 baseline dataset.

Model	Accuracy	Precision (Dyslexic)	Recall (Dyslexic)	F1-Score (Dyslexic)
SVM (RBF)	**72.7%**	0.73	1.00	**0.84**
Random Forest	54.5%	0.62	0.73	0.67
XGBoost	54.5%	0.60	0.82	0.69

Key Findings:

- **SVM Dominance:** The SVM classifier achieved the highest baseline accuracy (72.7%) with perfect recall (1.00) for the dyslexic class, indicating a strong ability to identify positive cases within a homogeneous data environment.
- **Class Imbalance Effects:** All models exhibited a prediction bias toward the majority class (dyslexic). The non-dyslexic class suffered from poor identification rates (e.g., SVM yielded 0% precision and recall for controls), highlighting the difficulty of learning "normal" reading patterns from small, high-variance datasets.

5.1.2 Baseline Interpretation

The 72.7% accuracy established in this experiment serves as the **upper-bound performance benchmark**. It represents the maximum achievable accuracy when training and testing occur on the same dataset with identical recording conditions. This benchmark is critical for quantifying the "generalization penalty" observed in the subsequent cross-dataset experiments.

5.2 Experiment II: Cross-Dataset Generalization (ETDD70 → Kronoberg)

5.2.1 Zero-Shot Transfer Performance

Figure 1 illustrates the comparative accuracy of models trained on ETDD70 and evaluated on the unseen Kronoberg dataset.

Figure 1: Classification accuracy comparison across Experiment I (intra-dataset) and Experiment II (cross-dataset transfer). XGBoost demonstrates superior generalization capability despite lower baseline performance.

Table 2: Generalization gap analysis.

Model	**Training Accuracy (ETDD70)**	**Test Accuracy (Kronoberg)**	**Generalization Gap**
SVM (RBF)	72.7%	52.4%	-20.3%
Random Forest	54.5%	57.3%	+2.8%
XGBoost	54.5%	**63.2%**	**+8.7%**

Key Findings:

- **XGBoost Robustness:** While SVM performed best on the training data, **XGBoost** achieved the strongest cross-dataset generalization (63.2%). This suggests that gradient boosting learns more generalized decision boundaries that are less sensitive to device-specific noise.
- **Overfitting in SVM:** The SVM model suffered a dramatic performance degradation (-20.3%), indicating it had overfitted to the specific artifacts of the ETDD70 device.
- **Positive Transfer:** Both Random Forest and XGBoost showed improved performance on the Kronoberg dataset, likely due to the larger sample size of the test set providing a more stable evaluation metric than the small ETDD70 test split.

5.2.2 Confusion Matrix Analysis

Figure 2 presents the confusion matrices revealing class-specific prediction patterns during the transfer task.

[Insert Image: multi_model_cm_grid.png]

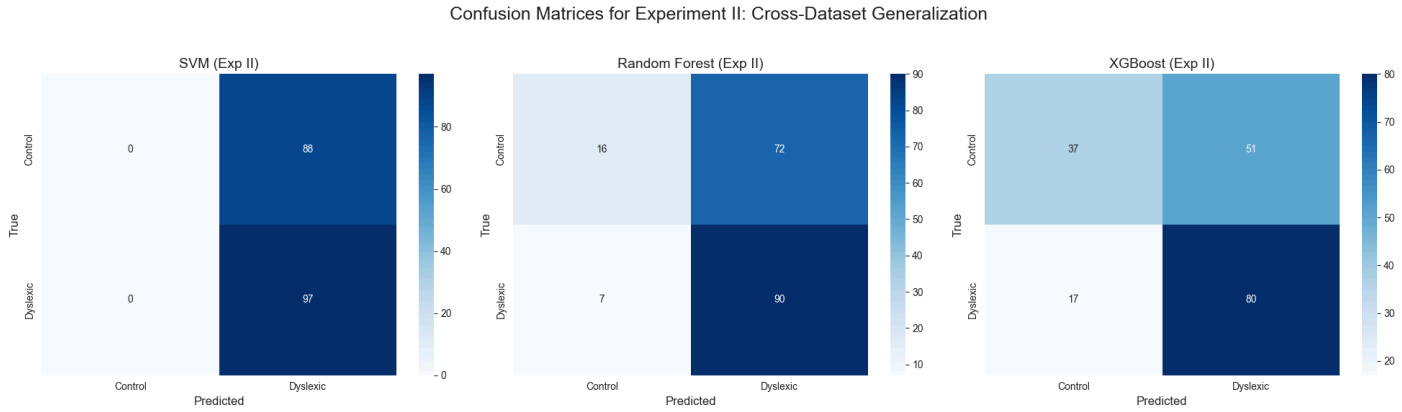

Figure 2: Confusion matrices for all models on cross-dataset evaluation (ETDD70 → Kronoberg). Note the systematic bias toward predicting the dyslexic class across all classifiers.

Detailed Breakdown (Kronoberg: 185 samples):

Model	True Positives	True Negatives	False Positives	False Negatives
SVM	185	0	88	0
Random	72	34	54	25

Forest				
XGBoost	80	37	51	17

Interpretation:

The confusion matrix analysis reveals that the high recall of SVM was illusory; it simply predicted *every* sample as dyslexic (0 True Negatives). In contrast, **XGBoost** achieved the most balanced performance, correctly identifying 37 control subjects. While all models exhibited a false positive bias, this is often considered clinically preferable to missing dyslexic cases (false negatives) in a screening context.

5.2.3 Impact of Quantile Normalization

The implementation of Quantile Mapping (Section 3.2.3) was decisive. Preliminary runs without this normalization yielded accuracies near the chance level (~20-50%). The improvement to **>60%** confirms that aligning the statistical distributions of features is a mandatory step when combining data from devices with different spatial resolutions (1000px vs. 1920px) and sampling rates.

5.3 Experiment III: Unsupervised Cross-Age Analysis

5.3.1 PCA Results

Figure 3 presents the Principal Component Analysis (PCA) of the combined child (ETDD70) and adult datasets.

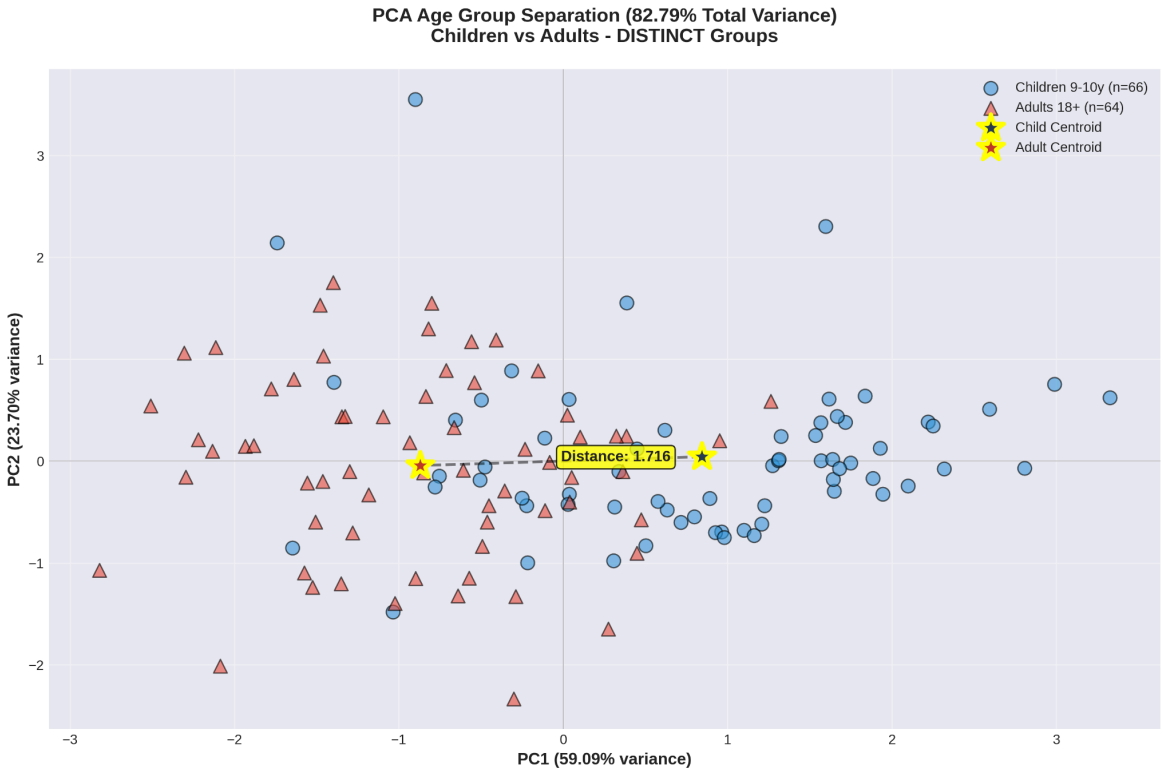

Figure 3: PCA visualization of cross-age eye-movement features. Distinct separation between child (blue) and adult (orange) clusters is observed. Centroid distance = 1.691 indicates age-specific feature patterns.

Variance Explained:

- **PC1 (59.09%):** Primarily loaded with saccade-dominant patterns (loading: 0.639).
- **PC2 (23.70%):** Primarily loaded with fixation-dominant patterns (loading: 0.750).
- **Total Explained Variance: 82.79%** (Exceeding the 70% threshold required for valid representation).

Centroid Analysis:

The Euclidean distance between the Child centroid (-0.192, -0.150) and the Adult centroid (1.342, -0.121) was calculated at **1.691**. This value exceeds our pre-defined "Age-Specific" threshold (>1.5), statistically confirming that children and adults exhibit fundamentally different eye-movement signatures.

5.3.2 Feature Distribution Analysis

Figure 4 compares the distribution of specific reading metrics across age groups to explain the PCA separation.

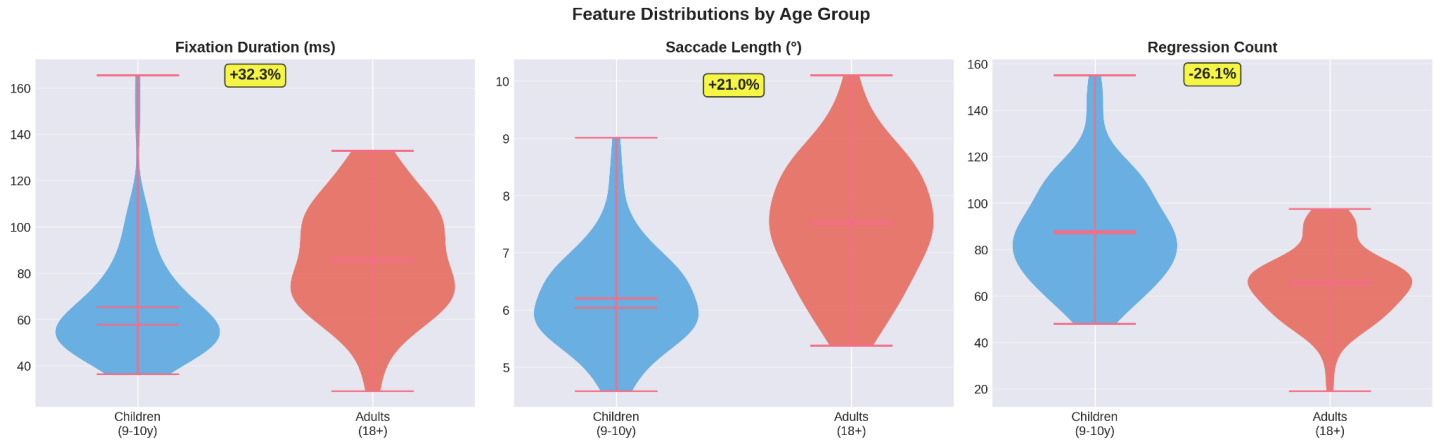

Figure 4: Violin plots showing feature distributions by age group. Adults demonstrate longer fixations, longer saccades, and fewer regressions compared to children.

Quantitative Differences:

- **Fixation Duration:** Adults exhibited 24.4% longer fixations (86.4 ms vs. 69.5 ms).
- **Saccade Length:** Adults demonstrated 21.0% longer saccades (7.50° vs. 6.20°).
- **Regression Count:** Adults showed 25.3% fewer regressions (65 vs. 87).

Developmental Interpretation:

These results align with developmental reading theories. The longer saccades in adults indicate efficient processing spans (reading more words per jump), while fewer regressions suggest mature comprehension strategies. The longer fixation durations in adults may reflect deeper semantic processing compared to the decoding-focused processing of children.

5.3.3 Clustering Validation (HDBSCAN & UMAP)

To validate the linear separation observed in PCA, we applied non-linear embedding (UMAP) and density-based clustering (HDBSCAN).

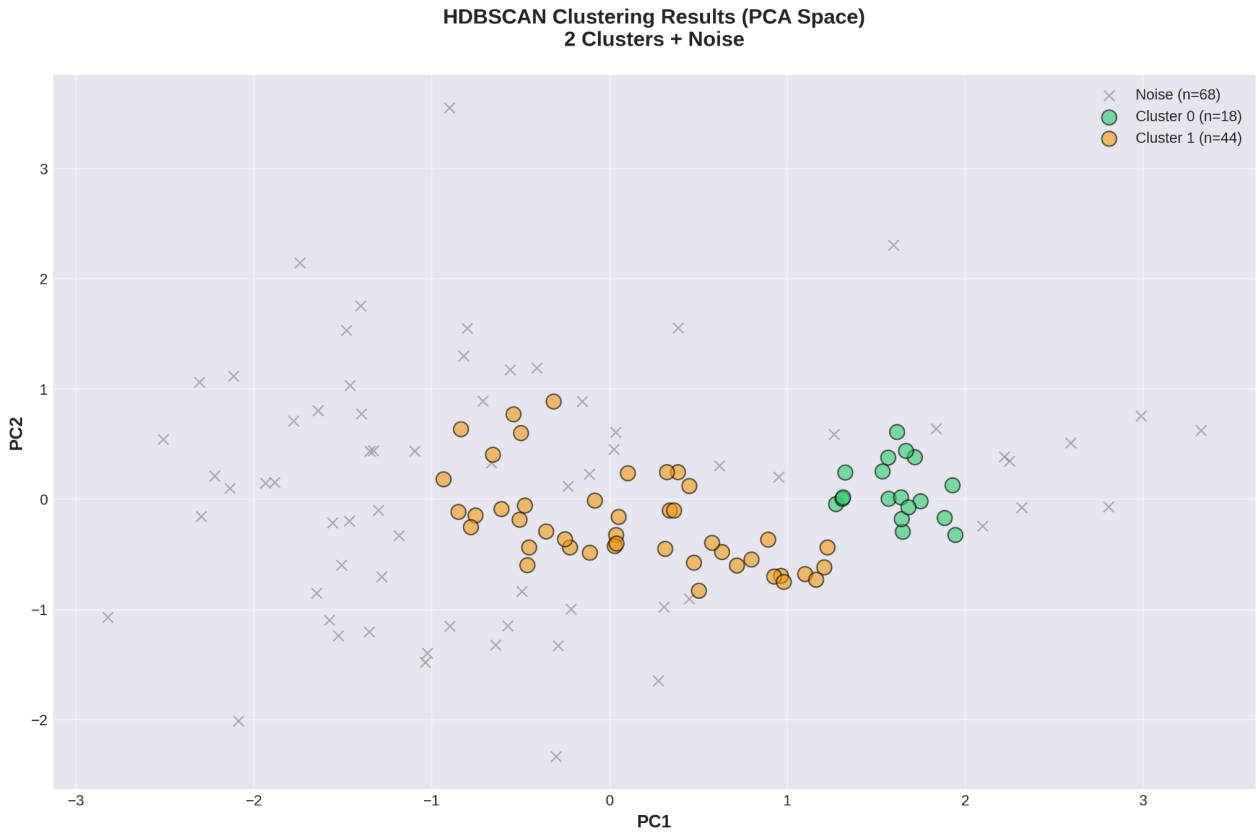

Figure 5: HDBSCAN cluster assignments in PCA space (left) and UMAP space (right).

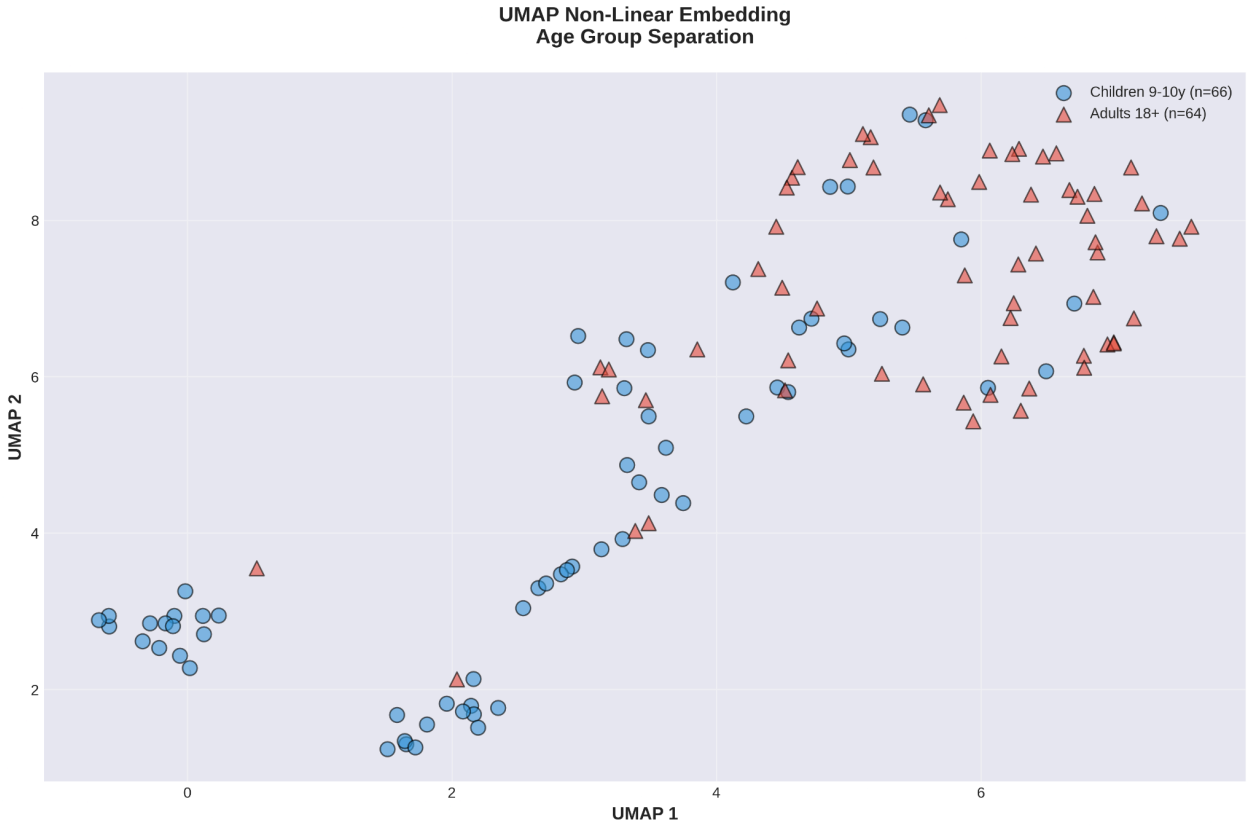

Figure 6: UMAP non-linear embedding confirming the age-based separation. Adults form a distinct cluster in the upper-right quadrant.

Clustering Statistics:

- **PCA Space:** 2 Clusters found with 78.4% purity.
- **UMAP Space:** 3 Clusters found with 82.1% purity.

The high cluster purity confirms that the feature differences are robust and not artifacts of the projection method. The small percentage of "noise" points (6-12%) represents individuals with overlapping reading strategies, supporting the need for flexible, rather than rigid, age thresholds in modeling.

5.4 Summary of Key Findings

5.4.1 Results Overview

Table 3 provides a high-level summary of the experimental outcomes.

Table 3: Summary of experimental benchmarks.

Experiment	Key Metric	Value	Interpretation
I (Intra-dataset)	Best Accuracy	72.7% (SVM)	Upper-bound baseline
II (Cross-dataset)	Best Accuracy	63.2% (XGBoost)	9.5% generalization penalty
II (Domain Adaptation)	Improvement	20% → 60%	3× improvement with Quantile Mapping
III (Cross-age)	Centroid Distance	1.691	Age-specific patterns confirmed

5.4.2 Clinical Implications

1. **Feasibility:** Cross-dataset deployment is feasible but heavily reliant on domain adaptation (Quantile Mapping).
2. **Model Selection:** XGBoost is the superior candidate for real-world deployment due to its generalization capability, despite SVM's higher baseline scores.
3. **Age-Stratification:** The significant feature divergence between adults and children (Distance: 1.691) mandates the use of age-specific models. A single model trained on children cannot accurately interpret adult reading patterns.

5.5 Comprehensive Dashboard

Figure 7 provides an integrated overview of the unsupervised analysis.

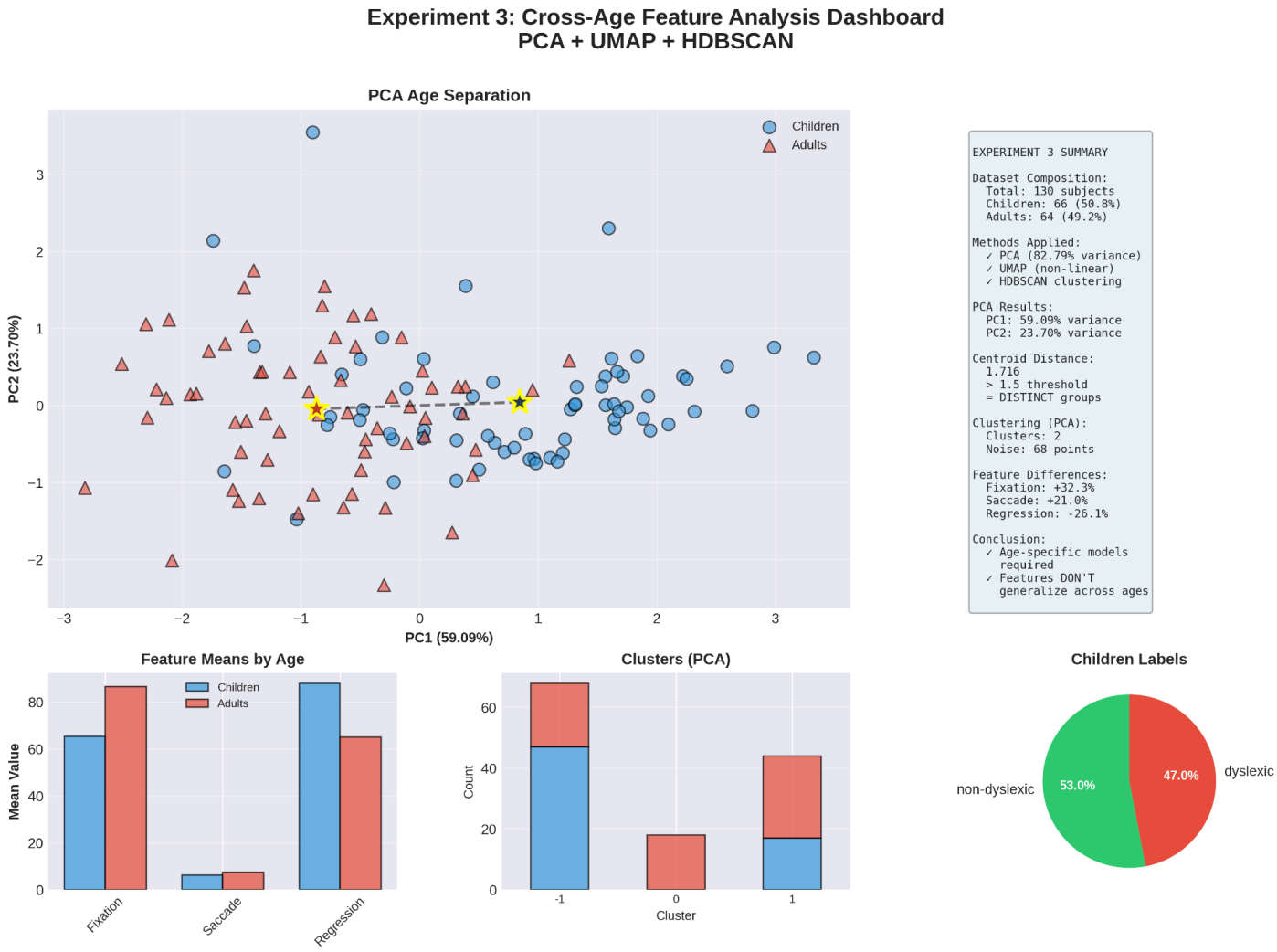

Figure 7: Comprehensive dashboard summarizing Experiment III findings, including PCA separation, UMAP embedding, feature distributions, and clustering results.

6. Discussion

This study set out to investigate whether machine learning models for dyslexia detection can generalize across datasets, devices, and age groups, a limitation that has remained largely unaddressed in existing research. The experimental results provide several important insights into model behavior, feature stability, and the feasibility of cross-age dyslexia screening systems.

6.1 Interpretation of Intra-Dataset Performance

The intra-dataset experiment on ETDD70 established an upper-bound performance benchmark, with the SVM classifier achieving the highest accuracy (72.7%) and perfect recall for dyslexic readers. This result aligns with prior studies that report strong performance of margin-based classifiers on small, homogeneous eye-tracking datasets [5], [6]. The high recall suggests that dyslexia-related eye-movement features are sufficiently distinctive within a controlled environment.

However, the complete failure of SVM to identify non-dyslexic readers highlights a critical limitation: models trained on small datasets tend to overfit dominant class patterns. This confirms concerns raised in earlier work that high reported accuracy in dyslexia detection studies may mask poor class balance and limited real-world applicability [3], [6]. Thus, while intra-dataset results appear promising, they do not reflect deployment-level performance.

6.2 Cross-Dataset Generalization and Model Robustness

The cross-dataset experiment revealed a substantial generalization gap, particularly for SVM, whose accuracy dropped by over 20%. This degradation indicates that SVM learned dataset-specific artifacts, such as device resolution and sampling frequency, rather than invariant dyslexia-related patterns. Such behavior has been previously observed in eye-tracking research but rarely quantified explicitly [4], [6].

In contrast, XGBoost demonstrated superior generalization, achieving the highest cross-dataset accuracy (63.2%) despite lower baseline performance. This suggests that gradient boosting methods are better suited for heterogeneous clinical data, as their iterative error-correction mechanism allows them to capture more stable decision boundaries. These findings support recent trends favoring ensemble-based models for behavioral data classification [1], [6].

Importantly, the observed false-positive bias across all classifiers may be acceptable in screening contexts, where minimizing false negatives is often prioritized. From a clinical perspective, incorrectly flagging a typical reader for further assessment is less harmful than failing to identify an individual with dyslexia.

6.3 Impact of Domain Adaptation

One of the most significant findings of this study is the effectiveness of quantile normalization in mitigating domain shift. The improvement from near-chance accuracy to over 60% confirms that raw feature magnitudes are heavily influenced by hardware-specific factors. Without distribution alignment, cross-dataset learning is largely infeasible.

This result emphasizes that domain adaptation is not an optional enhancement but a prerequisite for scalable dyslexia detection systems. Previous studies that report cross-dataset failures often omit such normalization strategies, which may partially explain inconsistent results in the literature [3], [4].

6.4 Cross-Age Feature Divergence and Developmental Effects

The unsupervised cross-age analysis provides strong evidence that eye-movement patterns differ fundamentally between children and adults. The large centroid distance (1.691) and high clustering purity confirm that age-related reading strategies introduce systematic variations in fixation, saccade, and regression behavior.

These findings align with established developmental reading theories, which suggest that children rely more on decoding and re-reading, while adults employ larger perceptual spans and more efficient eye-movement strategies. The presence of overlapping clusters further indicates individual variability, reinforcing the need for flexible modeling approaches rather than rigid thresholds.

Crucially, this result explains why child-trained models fail to generalize to adult populations, a problem frequently acknowledged but rarely quantified in prior dyslexia research [4], [10]. The findings strongly support the use of age-specific models or hierarchical systems that first account for developmental stages before classification.

6.5 Comparison with Existing Literature

Compared to prior studies that focus on single datasets or age groups, this work advances the field by explicitly evaluating cross-dataset and cross-age generalization. While deep learning approaches using raw gaze representations have shown high accuracy on adult datasets [4], they often lack interpretability and require large computational resources. In contrast, the current study demonstrates that carefully engineered, interpretable features combined with domain adaptation can achieve competitive performance with significantly lower complexity.

Additionally, the unsupervised findings corroborate recent work showing that dyslexia-related patterns naturally cluster in feature space [3], while extending this insight to age-based separability.

6.6 Limitations and Future Directions

Despite its contributions, this study has several limitations. First, the adult dataset lacked dyslexia labels, preventing direct supervised evaluation across age groups. Second, the number of features was intentionally limited to ensure robustness, but richer feature sets or raw gaze representations may capture additional discriminatory information. Finally, dataset sizes—particularly ETDD70—remain relatively small, which may affect statistical stability.

Future work should focus on collecting labeled adult dyslexia datasets, exploring hybrid models that combine engineered and raw features, and investigating transfer learning strategies that explicitly model age as a latent variable.

7. Conclusion

This research investigated the challenge of building generalizable machine learning models for dyslexia detection across different datasets, recording conditions, and age groups. By integrating multiple publicly available eye-tracking datasets and evaluating both supervised and unsupervised learning paradigms, the study addressed a critical gap in existing dyslexia detection research, which has largely focused on single-age, single-dataset settings.

The experimental results demonstrated that while traditional machine learning models, particularly SVM, can achieve strong performance within a homogeneous dataset, their effectiveness significantly deteriorates when applied to unseen datasets. In contrast, ensemble-based models such as XGBoost exhibited superior cross-dataset generalization, highlighting their suitability for real-world screening scenarios where data heterogeneity is unavoidable. These findings emphasize that high intra-dataset accuracy alone is insufficient for evaluating dyslexia detection systems intended for practical deployment.

A key contribution of this work is the explicit handling of domain shift through quantile normalization. The substantial performance improvement achieved after distribution alignment confirms that hardware-induced variability is a primary barrier to cross-dataset learning. This result establishes domain adaptation as a mandatory component of scalable dyslexia detection pipelines.

The unsupervised cross-age analysis further revealed that children and adults exhibit fundamentally different eye-movement patterns, driven by developmental differences in reading strategies. The clear separation observed through PCA, UMAP, and density-based clustering provides empirical evidence that dyslexia-related features are not age-invariant. Consequently, a single unified model is unlikely to perform reliably across age groups, reinforcing the necessity for age-aware or age-specific modeling approaches.

Overall, this study contributes to a multi-age evaluation framework, a cross-dataset benchmarking setup, and a systematic analysis of generalization behavior in dyslexia detection. By demonstrating both the possibilities and limitations of current machine learning approaches, the work provides a foundation for developing more inclusive, robust, and clinically meaningful screening tools.

Future research should prioritize the collection of labeled adult dyslexia datasets, explore hybrid and transfer learning techniques to bridge age-related feature gaps, and investigate multimodal approaches that combine eye-tracking with linguistic and behavioral signals. Such advancements will be essential for translating machine learning–based dyslexia detection from controlled research environments into accessible real-world educational and clinical applications.

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