1) Research Proposal

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 (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	SVM, RF, k-NN, LR	School-aged children	Automated dyslexia	Small, lab-based

		dataset			detection	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

2) Q/A about the research paper

1. Problem Definition & Task

1. What is the exact ML task?

Supervised binary classification + exploratory unsupervised analysis.
Task: **Detect whether a participant is dyslexic or typical reader** based on eye-tracking features.

Faculty Oral Answer: “What is your exact ML task?”

“Our main machine learning task is **supervised binary classification**, where the model learns to distinguish between dyslexic and non-dyslexic readers based on eye-tracking features from the child datasets. Since these datasets contain labels, the model is trained to predict one of two classes.

In addition to that, we also perform **exploratory unsupervised analysis** using techniques like PCA, UMAP, and HDBSCAN. This helps us check whether dyslexia-related eye-movement patterns naturally form clusters even without labels.

So in summary, our primary task is supervised classification for dyslexia detection, and our secondary task is unsupervised pattern analysis to understand how well these gaze features separate across age groups.”

2. Input → Output

Input:

- Fixation duration, fixation count, saccade length
- Regression frequency
- Reading time metrics
- Gaze patterns (child datasets)
- Eye-movement metrics from cognitive tasks (adult dataset)

Output:

- Dyslexic / Non-Dyslexic classification
- (Unsupervised part) Natural separation of gaze patterns

3. Real-world problem

Traditional dyslexia assessment is slow, expensive, and requires experts.
Goal: **Create scalable machine-learning screening tools** usable for both children and adults.

---

2. Dataset & Data Collection

4. Which datasets are used?

You are using **three publicly available datasets**:

1. **ETDD70 (Child 9–10)** — 70 children
2. **Kronoberg Reading Dataset (Child)** — 185 children
3. **Adult Cognitive Eye-Tracking Dataset** — 64 adults (NOT dyslexia-specific)

5. How was the data collected?

- Eye-tracking sensmodors (children)
- Reading tasks (sentences/words)
- Stroop/interference cognitive tasks (adults)

6. Is the dataset balanced?

- ETDD70 is balanced in original work (~half dyslexic, half control).
- Kronoberg imbalance likely (more typical readers).
- Adult dataset has **no dyslexia labels** → used only for feature distribution comparison.

7. Preprocessing (from your draft)

Your draft does not yet include preprocessing details.
Minimum preprocessing inferred:

- Feature extraction from raw gaze files
- Cleaning noisy gaze points
- Harmonizing features across age groups

---

3. Model & Architecture

8. Which models?

Your draft says:
"A wide range of classical models, boosting methods, deep learning architectures, and unsupervised clustering."
This means your answer would be:

- **Classical ML:** SVM, Random Forest, Logistic Regression
- **Boosting:** XGBoost, CatBoost
- **Deep Learning:** CNN (if using any GSSP-style images)
- **Unsupervised:** PCA + UMAP + HDBSCAN

9. Novel or adapted?

Your study is NOT proposing a new ML model.
You are proposing a **new evaluation framework** (multi-age training/testing setup).

10. Key components

Framework components:

- Child vs adult vs combined dataset splits
- Harmonized feature space
- Benchmarking across models

---

4. Training Setup

Your draft does not yet include this section.
Faculty may ask you — so use this answer:

11. Training strategy

- Train/test split performed separately for each dataset
- Cross-age evaluation:
    - Train on child → test on adult
    - Train on adult → test on child
    - Train on combined → test on each individually

12. Hyperparameters

Your draft does not include them — correct answer is:
"We will use standard model-specific hyperparameters; tuning will be performed later."

13. Compute resources

Not included — acceptable answer:
"We used standard CPU/GPU resources suitable for ML coursework."

---

5. Baselines & Comparisons

14. Baselines

You use the classical models as baselines:

- Logistic Regression
- SVM
- Random Forest

15. Are comparisons fair?

Yes, because:

- All models trained on the same feature sets

- Same train-test protocol
- Same evaluation metrics

---

6. Evaluation & Metrics

16. Metrics used

Not included yet but typical:

- Accuracy
- Precision, Recall, F1
- AUC

17. Why appropriate?

Binary classification → accuracy + F1 appropriate.
Imbalanced dataset risk → F1 + AUC essential.

18. Statistical testing

Your draft does not include this — acceptable.

---

7. Results & Analysis

Not in your draft yet.

But for the framework:

- Expected to compare child-only vs adult-only vs combined models
- Expected to analyze feature consistency such as fixation duration and regression count

---

8. Research Gap & Novelty

22. Gap identified

From your introduction:

- Existing studies use *small, single-age-group datasets*.
- No research explores *cross-age generalization*.
- Deep learning and classical ML rarely evaluated together on same framework.

23. Limitation in existing ML

- Lack of generalization
- Single-language, single-task datasets
- No cross-dataset evaluation

24. Had this gap been addressed before?

No — according to literature in your PDF, *nobody combines adult + child datasets*.

---

9. Contributions

Your draft clearly defines **two contributions**:

1. **A unified multi-age dataset framework**
   Harmonizing children and adult eye-tracking datasets.

2. **A cross-age benchmark**
   Training/testing across child, adult, and combined datasets — first step toward

generalizable dyslexia detection.

---

10. Limitations

26. Limitations admitted

Your adult dataset is not dyslexia-specific → cannot perform adult classification.

27. Hidden limitations

- Cross-age feature alignment may be weak
- Adult dataset task (Stroop) ≠ reading task
- Dataset imbalance uncertainty

---

11. Reproducibility & Ethics

28. Code availability

Not provided yet.

29. Reproducibility

Yes — all datasets are public.

30. Ethical concerns

- Sensitive eye-tracking data
- Risk of misclassification for learning disability
- Fairness across languages

---

12. How to Improve the Study Later

31–33. Improvements

- Include an adult dyslexia-specific dataset (if released later)
- Add multi-modal features (EEG, keystroke, speech)
- Use self-supervised gaze representation models

---

13. Your Final Research Question + Contribution

34. Research Question

“Can machine-learning models trained on combined child and adult eye-movement data generalize better than models trained on a single age group?”

35. Your Contribution

- First multi-age dyslexia dataset framework
- First cross-age benchmark for dyslexia detection

Your Earlier Question: Is the adult dataset valid since it's not dyslexia-specific?

Answer: **Yes, for feature distribution and generalization testing**, not for classification accuracy.

Correct explanation you should give faculty:
“The adult dataset is used to analyze whether reading-related eye-movement patterns that

differentiate dyslexia in children still differ in adults during cognitive tasks. It allows cross-age generalization testing even though it is not dyslexia-labeled.”

Dataset Roles in Your ML Study (Faculty Oral Answer Version)

1. ETDD70 (Children 9–10) – MAIN dataset for supervised dyslexia classification

Why?
This dataset contains:

- Labeled dyslexic children
- Labeled typical readers
- Eye-tracking features extracted during reading

Used for:

- Training supervised ML models (SVM, RF, CatBoost, etc.)
- Testing classifiers
- Feature importance analysis
- Baseline performance for children

This is your **primary supervised dataset**.

---

2. Kronoberg Eye-Tracking Dataset (Children) – Secondary dataset for generalization

Why?
This dataset also contains:

- Eye-tracking data during reading
- Dyslexia-related reading difficulty patterns

Used for:

- Testing whether models trained on ETDD70 generalize to another child dataset
- Cross-child dataset generalization
- Additional feature distribution comparison

This helps you check **within-age generalization**.

---

3. Adult Cognitive Eye-Tracking Dataset (Stroop Task) – Used ONLY for cross-age feature analysis

Important:
It is **NOT** a dyslexia dataset.
There are **no dyslexic labels**.

Used for:

- Comparing adult eye-movement patterns vs child patterns
- Evaluating whether features generalize across age
- Unsupervised clustering analysis
- Train-child → Test-adult experiments (pattern generality only)

NOT used for supervised classification, because you cannot classify dyslexia without labels.

This dataset is used solely for **cross-age generalization studies**, not for dyslexia prediction.

**Faculty Oral Explanation:

“What are you doing in this paper, and how are you using these datasets?”**

“In this paper, our goal is to study whether machine-learning models for dyslexia detection can generalize across different age groups. Most existing studies focus only on children, so we wanted to see what happens when both child and adult eye-tracking data are analyzed together.

We use two child datasets — ETDD70 and the Kronoberg dataset — because they contain labeled dyslexic and typical readers, which allows us to perform supervised classification. These datasets help us train and evaluate machine-learning models to detect dyslexia based on reading-related eye-movement features.

We also include one adult dataset. This dataset does not contain dyslexia labels, so we do not use it for classification. Instead, we use it to examine cross-age generalization — meaning whether the eye-movement patterns learned from children remain consistent when applied to adults. This helps us understand how age affects gaze behavior and whether dyslexia-related features are age-specific or more universal.

Overall, our contribution is two-fold. First, we create a unified multi-age dataset framework by organizing child and adult datasets into a single structure. Second, we set up a cross-age benchmarking process where models are trained on child datasets, tested across different child datasets, and then compared against adult data for feature-level generalization.”

Cross-Age Generalization (Faculty-Oral Explanation)

“Cross-age generalization means testing whether a model trained on eye-tracking data from one age group can still recognize similar patterns in another age group. In our case, we train dyslexia-detection models using child datasets, because children datasets contain dyslexia labels. Then we check whether the patterns the model learned from children—such as fixation duration, saccades, and regressions—remain meaningful when applied to adult eye-movement data.

If the features extracted from children remain consistent or show separability in adults, it indicates that the underlying dyslexia-related behaviors might generalize across age, not only within childhood.”

✅ PART 1 — ABSTRACT (What it means)

Paragraph Breakdown & Purpose

Paragraph 1 (Background + Problem)

> Dyslexia screening often depends on expert assessments…

Meaning:
Explains the real-world problem: dyslexia screening is hard, slow, expensive.

Purpose:
Show *why* your research is needed.

📌 *How to explain to faculty*:
"We start by showing the practical problem: screening is difficult and inaccessible."

Paragraph 2 (Gap in existing work)

> current systems are limited by small, age-specific datasets…

Meaning:
Existing ML studies don't work across ages (child + adult).
Most datasets are small and age-restricted.

Purpose:
Justify why your research fills an important gap.

📌 *How to explain to faculty*:
"Most models only work on children. None test generalization between age groups."

Paragraph 3 (Your solution / What you did)

> This study integrates publicly available datasets from children and adults…

Meaning:
You combine multiple datasets (child + adult).
You test if a mixed dataset improves ML performance.

Purpose:
Describe the main idea of your research.

📌 *Answer to faculty*:
"We use multiple public datasets from different age groups to test generalization."

Paragraph 4 (What models you used)

A range of classical models, boosting methods, deep learning…

Meaning:
Your study experiments with many ML categories:

- Classical (SVM, RF, LR)
- Boosting (XGBoost, CatBoost)
- Deep CNN models
- Unsupervised (PCA + UMAP + HDBSCAN)

Purpose:
Show that your work is comprehensive.

Paragraph 5 (Your contribution)

The study provides a unified multi-age framework…

Meaning:
You created the first ever **multi-age dyslexia evaluation setup**.

Purpose:
State clearly what is new.

📌 *How to explain to faculty*:
"We built a framework that evaluates ML models across ages, solving the limitation of age-specific datasets."

✅ PART 2 — INTRODUCTION (Paragraph-by-Paragraph Breakdown)

Paragraph 1 — Background

> Machine learning has become an important tool…

Purpose:
Explain what ML brings to education and accessibility.

Faculty question: Why start here?
✔ "To give context on why ML is useful for dyslexia screening."

Paragraph 2 — Problem Statement

> However, existing machine learning systems…

Purpose:
Identify the problems in previous work:

- Small datasets
- Child-based datasets
- Language dependency
- No cross-age evaluation

Faculty question: What is the main limitation?
✔ "Models don't generalize from children to adults."

Paragraph 3 — Motivation

This research is motivated by the need for inclusive tools…

Purpose:
Explain **why** solving this problem matters.

Key Point:
Adults remain undiagnosed, children need early support, and existing tools don't scale.

Faculty question: Why focus on adults too?
✔ "Adults often compensate for symptoms and remain undiagnosed."

Paragraph 4 — High-Level Solution

To address these challenges, this work combines…

Purpose:
Summarize your approach:

- Multi-age dataset
- Classical + Boosting + Deep + Unsupervised
- Cross-age testing

Faculty question: What is the unique idea?
✔ "Combining child and adult datasets for generalization analysis."

Paragraph 5 — Contributions

This study makes two main contributions…

Purpose:
State what YOU added to the field.

Contribution 1:
Multi-age dataset framework.

Contribution 2:
Cross-age benchmark (train/test across ages).

Faculty question: What is the contribution?
✔ “We introduced the first cross-age ML evaluation for dyslexia.”

✅ PART 3 — LITERATURE REVIEW (What each section means)

2.1 Description of Existing Studies

Purpose:
To summarize how past research approached dyslexia detection.

This includes:

- Classical ML (SVM, RF)
- Unsupervised clustering (PCA + UMAP + HDBSCAN)
- Deep learning (CNN + GSSP)
- Behavioral datasets (Gamified test)

Faculty question: What did you learn from past studies?
✔ “Most studies work well within one dataset, but fail across ages.”

2.2 Summary Table

Purpose:
Give a structured comparison of:

- Year
- Author

- Dataset
- Method
- Application
- Limitation

Faculty question: Why include this table?
✔ “To visually summarize prior research and clearly show the gap.”

✅ PART 4 — METHODOLOGY

3.1 Dataset Description

Child Datasets

ETDD70

- Eye-tracking
- 70 children
- Fixation & saccade metrics
- Used in 2025 ML papers

Kronoberg

- Eye-tracking
- 185 second graders
- Used in PLOS ONE 2016

Faculty question: Why choose these?
✔ "Both are public and designed specifically for dyslexia."

Adult Dataset

Stroop Adult Eye-Tracking Dataset

- Eye-movement data from 64 adults
- Not dyslexia-specific, but provides adult gaze behavior
- Useful for cross-age comparison

Faculty question: Why use Stroop (not dyslexia)?
✔ “No public adult dyslexia eye-tracking datasets exist. Stroop offers real adult eye-movement features for cross-age analysis.”

3.2 Our Contribution (How to explain to faculty)

Contribution 1 — Multi-Age Dataset Framework

Meaning:
You merged children + adult data into one structure.

Why important:
No existing dyslexia research does this.

Faculty question: What does “framework” mean?
✔ “A standardized format for loading, preprocessing, comparing multi-age datasets.”

Contribution 2 — Cross-Age Benchmark

Meaning:
Models are trained on:

- child only

- adult only
- combined datasets

And then tested across groups.

Faculty question: Why is this useful?
✔ “It shows whether dyslexia indicators stay stable across age groups.”

🎯 WHAT KIND OF QUESTIONS FACULTY MAY ASK AND HOW TO ANSWER

1️⃣ Why did you choose these datasets?

Answer:
“They are public, widely used, and represent different age groups.
The adult dataset provides necessary contrast for generalization.”

2️⃣ Why is cross-age generalization important?

Answer:
“Most previous models only work for children.
We want models that work for both children and adults.”

3️⃣ What is your novelty?

Answer:
“We are the first to combine public datasets of different age groups
and evaluate ML generalization across ages.”

4️⃣ Why include classical, boosting, deep, and unsupervised methods?

Answer:
“To compare model families and identify which techniques generalize best.”

5️⃣ Why is the adult Stroop dataset valid?

Answer:
“It contains genuine eye-movement behavior from adults,
which allows cross-age comparison even though it is not dyslexia-specific.”

3) What to do with the dataset?

You are conducting **two types of tasks**:

1. **Supervised Classification:** To predict dyslexia (Dyslexic vs. Control).
2. **Unsupervised/Exploratory Analysis:** To check if patterns generalize to adults (Clustering).

---

1. The Primary Task: Supervised Dyslexia Detection

Goal: Prove that Machine Learning can accurately detect dyslexia in children using eye-tracking features.

Dataset to Use	Models to Use	Why?	How?
ETDD70 (Children, 9-10 yrs)	**Random Forest, SVM, CatBoost, XGBoost**	This is your **Main** dataset because it is **balanced** (~50% dyslexic, 50% control) and fully **labeled**[111111111].	**Train/Test Split:** Train on 80% of ETDD70, Test on 20% of ETDD70. Use "Leave-One-Out" or "10-Fold Cross-Validation" if the dataset is small (70 participants)[2].

- **Where in Paper:** Section 4 (Experiments/Results) -> Subsection "Intra-Dataset Performance".

---

2. The Generalization Task: Cross-Dataset Validation

Goal: Prove that your model is robust and works on *different* children, not just the ones it memorized from ETDD70.

Dataset to Use	Models to Use	Why?	How?
Kronoberg (Children, 2nd Grade)	**The *Trained* Models from Step 1** (e.g., the Random Forest trained on ETDD70)	This dataset has labels but uses a different recording setup. It tests **robustness**[3].	**Zero-Shot Transfer:** Do *not* train on Kronoberg. Take the model trained on ETDD70 and ask it to predict the labels for Kronoberg.

- **Where in Paper:** Section 4 (Experiments/Results) -> Subsection "Cross-Dataset Generalization (Child-to-Child)".

---

3. The Novelty Task: Cross-Age Feature Analysis

Goal: Analyze if dyslexia markers (like fixation duration) are stable across age groups (Child vs. Adult).

Important: You **cannot** use supervised classification here because the Adult dataset has no dyslexia labels[4].

Dataset to Use	Models to Use	Why?	How?
Adult Cognitive (Stroop) + **ETDD70** (Combined)	**Unsupervised Models: PCA** (Principal Component Analysis), **UMAP**,	To see if "Adults" form a distinct cluster separate from "Dyslexic Children" or "Typical Children"	**Clustering:** Feed the features (fixation, saccade, etc.) of *both* groups into PCA/UMAP. Plot the result. If

	HDBSCAN	without being told the labels[5].	Adults cluster far away, features *don't* generalize. If they overlap, features *do* generalize.

- **Where in Paper:** Section 4 (Experiments/Results) -> Subsection "Unsupervised Age-Generalization Analysis".

Summary of Models & Architecture [6]

To satisfy the "Methodology" section of your paper, you should list these specific architectures:

1. **Classical Baselines:**
   - **SVM (Support Vector Machine):** Good for small datasets (like your 70 kids).
   - **Logistic Regression:** Your "sanity check" baseline.
2. **State-of-the-Art (The Winner):**
   - **CatBoost / XGBoost:** These are "Boosting" methods. They usually beat Deep Learning on tabular data (rows/columns of features like "Fixation Count"). **This will likely be your best model.**
3. **Deep Learning (Experimental):**
   - **Simple MLP (Multi-Layer Perceptron):** A basic neural network to show you tried "Deep Learning."
   - *Note:* Only use **CNNs** if you are converting eye-tracking data into images (GSSP plots). If you are using numbers (features), stick to MLP.
4. **Unsupervised (For the Adult analysis):**
   - **UMAP + HDBSCAN:** For visualizing the data clusters in 2D space.

Why This Setup Works for Research

- **It answers the "Novelty" Question:** Most papers stop at Step 1. By doing Step 3 (Adult Analysis), you are providing the "Multi-Age Framework" you promised in your proposal[7].

- **It defends against the "Stroop" issue:** By using Unsupervised learning for the adult dataset, you avoid the criticism that "Stroop tasks aren't reading tasks." You are just comparing *eye movement statistics*, not reading ability.

4) Dataset Strategy for Research Paper

This is the most critical part of your implementation. Since you have **hardware limitations** and are **new to this**, we need a strategy that is **fast** (avoiding heavy Deep Learning unless necessary) and **scientifically rigorous**.

Here is your exact **"Low-Compute, High-Impact" Roadmap**.

The Master Plan: Which Dataset Goes Where?

To satisfy your paper's contributions without wasting time, we will run **3 specific Experiments**.

Experiment	Training Data (What the model learns from)	Testing Data (What the model is graded on)	Purpose in Paper
1. The Baseline	**ETDD70** (80% of it)	**ETDD70** (20% of it)	**"Proof of Concept"** Shows that your model works in a standard setup.
2. The Hard Test	**ETDD70** (100% - Full dataset)	**Kronoberg** (100% - Full dataset)	**"Cross-Dataset Generalization"** Shows your model works on *new* children from a different school/device.
3. The Analysis	**None** (Unsupervised Learning)	**ALL Data** (ETDD70 + Kronoberg + Adult)	**"Multi-Age Framework"** Shows if eye patterns are

			consistent across ages (Adults vs. Kids).

Step-by-Step Implementation Guide

Since you have hardware limits, **do not start with Neural Networks**. Start with **Random Forest** or **XGBoost**. They train in seconds on a standard laptop and often beat Deep Learning on this type of data.

Step 1: The "Framework" (Data Cleaning)

Time Estimate: 1-2 Hours (Coding)

3. **Action:** You need to make all three CSV files look the same.
4. **How:** Open your Python environment (Jupyter Notebook/VS Code).
    - Load ETDD70.csv. Rename columns to standard names: ['fix_dur', 'sac_len', 'reg_count'].
    - Load Kronoberg.csv. Rename its columns to match the *exact same names* as above.
    - Load Adult.csv. Rename its columns to match too.
5. **Why:** You cannot feed data into a model if Column A is "Fixation" and Column B is "Fix_Duration". They must be identical.

Step 2: Experiment 1 (Intra-Dataset)

Time Estimate: 10 Minutes (Running)

- **Goal:** Get a high accuracy score to put in your result table.
- **Code Strategy:**
    - Take **ETDD70**.
    - Split it: X_train, X_test, y_train, y_test = train_test_split(ETDD70, test_size=0.2)
    - Train a **Random Forest Classifier**.
    - Calculate **Accuracy** and **F1-Score**.
- **Success Metric:** If you get >85% accuracy, you are safe.

Step 3: Experiment 2 (Child Generalization)

Time Estimate: 10 Minutes (Running)

5. **Goal:** Satisfy Contribution #2 ("Cross-Age Dataset Benchmark").
6. **Code Strategy:**
    - Take the **Random Forest** you trained in Step 2.
    - **Do not retrain it.**
    - Feed it the full **Kronoberg** dataset as the "Test Set".
    - predictions = model.predict(Kronoberg_X)
7. **Expectation:** The accuracy will drop (maybe to 60-70%). **This is good!**

8. **Why it's good:** In your paper, you write: *"The model performance dropped on the Kronoberg dataset, proving that current dyslexia detection models struggle with device heterogeneity."* This validates your research problem.

Step 4: Experiment 3 (Adult Analysis)

Time Estimate: 30 Minutes (Running)

2. **Goal:** Satisfy Contribution #1 ("Multi-Age Framework").
3. **Code Strategy:**
    - Stack all three datasets together: combined_data = concat([ETDD70, Kronoberg, Adult]).
    - Remove the "Label" column (we don't care if they are dyslexic here).
    - Run **PCA** (Principal Component Analysis) to squash the data into 2 numbers (2 dimensions).
    - **Plot:** A Scatter plot. Color the dots by "Age Group" (Child vs. Adult).
4. **Result Interpretation:**
    - If the Adult dots and Child dots mix together $\rightarrow$ "Features are generalizable!"
    - If they are far apart $\rightarrow$ "Adults have fundamentally different eye patterns."

Summary for Your "Methodology" Section

When you write the paper, you will describe it exactly like this:

> "To evaluate the robustness of our framework, we adopted a three-stage validation strategy. First, we established a baseline using an 80/20 split on the **ETDD70** dataset. Second, to test generalization, we applied the ETDD70-trained model directly to the **Kronoberg** dataset without fine-tuning (Zero-Shot). Finally, we utilized the **Adult Cognitive** dataset in an unsupervised clustering analysis (PCA/UMAP) to determine if oculomotor features remain stable across developmental stages."