Eye-Tracking Driven Dyslexia Detection: A Data-Efficient Approach Using Synthetic Augmentation and XGBoost Classifier

Rachana Ramchandar
CSE, PES University
Bengaluru, India
rachramchandar10@gmail.com

Samyuktha S
CSE, PES University
Bengaluru, India
samyukthasundarrajan@gmail.com

Shreya Mittal
CSE, PES University
Bengaluru, India
shreya5.mittal3@gmail.com

S Adarsh Nayak
CSE-AIML, PES University
Bengaluru, India
nayakadu004@gmail.com

Harshitha Pakati V
CSE, PES University
Bengaluru, India
harshitha05pakati@gmail.com

Mamatha H R
CSE, PES University
Bengaluru, India
mamathahr@pes.edu

***Abstract*—Early and accurate detection of Developmental Learning Disorders (DLD), such as dyslexia, remains a key challenge due to limited availability of publicly accessible data and diagnostic complexity. In this study, we propose a dyslexia detection model that promises efficient usage of data, using eye-tracking features extracted from reading behaviour. To overcome the challenge of sample scarcity, we augmented the publicly available Zenodo dataset (70 subjects) with 140 synthetically generated samples, forming a balanced dataset of 210 participants (105 dyslexic, 105 non-dyslexic). We extracted nine detailed spatiotemporal and behavioral features, such as fixation duration, saccade length, dispersion metrics, number of lines fixated, and trained an XGBoost classifier. The model was evaluated through 5-fold cross-validation, achieving a mean accuracy of 85.1% ($\pm$2.7). The final model achieved a training AUC of 0.9928 with an accuracy of 0.9464 and a test AUC of 0.9819, with an accuracy of 0.9268. No over-fitting was observed in the model (Train-Test gap: 0.0178). Our results show that, by combining synthetic augmentation with meticulously engineered features, advanced ensemble methods such as XGBoost can successfully support early screening of dyslexia through eye-movement analysis.**

***Index Terms*—Developmental learning disorders (DLD), Dyslexia detection, Eye tracking biomarkers, Neuro-cognitive engineering, Biomedical data engineering, Computational Neuroscience.**

I. Introduction

Developmental Learning Disorder (DLD) is a neurodevelopmental disorder based on biological factors and leads to abnormalities at the cognitive level, which are related to the behavioral signs of the disorder. It is defined by consistent difficulties in the learning and use of academic skills, notably in reading, writing, and/or mathematics [1].

Neuroimaging studies consistently show that fluent word recognition depends on the functional integrity of two-well established posterior systems in the left hemisphere (LH), systems that are characteristically disrupted in individuals with developmental dyslexia. Studies such as [2], [3] show evidences of abnormal neural activity observed in readers diagnosed with dyslexia in the left posterior temporo-parietal (TP) cortex-middle temporal gyrus, superior temporal gyrus, supramarginal gyrus and angular gyrus, the left occipito-temporal (OT) cortex-inferior temporal gyrus and fusiform gyrus, and the left frontal cortex-inferior frontal gryus and precentral gryus. These findings align with meta-analytic evidence of dyslexic underactivation relative to typical readers in the left TP and OT cortex. [3].

Eye-tracking research studies such as [4], [5] confirm these neurobiological observations by recognising distinct oculomotor patterns in dyslexic readers. Compared to non-dyslexic readers, those with dyslexia show longer fixation durations, shorter saccades, more regressions, and overall slower reading speeds. These measurable features have enabled the development of rapid eye-tracking based screening tools; for example, RADAR, [5] exhibited over 94% accuracy in distinguishing dyslexic readers using fixation and saccade features.

Studies like [6] have shown that conventional diagnostic assessments are time consuming and susceptible to examiner bias, necessitating the need for more scalable and data-driven screening methods. In this context, recent machine learning (ML) efforts, such as DysLexML [7] have shown promise by applying linear models to eye-tracking features, achieving classification accuracies above 95%. However, the drawback of these systems is that they require larger, more balanced datasets than the ones that are typically available.

To address the challenges of data scarcity and model explainability, this study proposes an XGBoost [8]–based dyslexia detection framework trained on an augmented eye-tracking dataset. From nine spatio-temporal and behavioral features, we synthesize additional samples to create a balanced dataset for reliable model training. Our approach resides within the domains of **computational neuroscience**, **biomedical data engineering**, and **neuro-cognitive engineering**, aiming to support early and accessible dyslexia screening.

Contribution of this work: This study makes three primary contributions: (1) we propose a perturbation-based data

augmentation strategy tailored for small dyslexia datasets; (2) we design a compact but informative set of nine gaze-based features, balancing interpretability with predictive performance; and (3) we demonstrate the superiority of XGBoost over alternative classifiers while benchmarking its performance against state-of-the-art models.

II. Literature Survey

The integration of eye-tracking technology and machine learning (ML) has opened new opportunities for reliable, noninvasive screening of developmental dyslexia. Traditional diagnostic methods primarily rely on psycholinguistic and neuropsychological evaluations, which are often time-consuming, require specialized expertise, and are prone to variability between contexts [6]. In contrast, eye-tracking methods provide objective spatio-temporal metrics such as fixation durations, saccade lengths, regressions, and scanpath complexity, parameters that are measurably altered in dyslexic readers [4], [5].

Several computational frameworks have emerged for dyslexia detection using gaze data. The RADAR system [5] achieved over **94%** accuracy in distinguishing dyslexic from non-dyslexic readers using fixation and saccade metrics. Asvestopoulou et al. [7] developed DysLexML, an ML-based screening tool for dyslexia using eye-tracking features during silent reading activities. Their best performing model was a linear SVM model trained on a reduced set of three gaze metrics - achieved a classification accuracy of **97%**, demonstrating both high predictive power and robustness to noise.

Hybrid approaches such as SVM with Particle Swarm Optimization (SVM-PSO) have demonstrated an accuracy of **95.6%** when trained on fixation and saccadic features [9]. Deep learning models resulted in further improvement of performance: CNNs trained on raw gaze sequences achieved up to **96.6%** accuracy on balanced and available datasets [4].

Benfatto et al. [10] used eye-tracking data collected during reading tasks and extracted 48 gaze features. Using an SVM classifier, they achieved **95.6% accuracy**, showing the potential of gaze-based ML models for early screening of dyslexia. Meena et al. [11] developed a virtual keyboard integrated with gaze-based typing assessments, using Linear Discriminant Analysis (LDA) to classify children with an AUC of **0.91**.

Moreover, a range of machine learning (ML) methods and models, such as KNN, SVM, Decision Trees, and CNNs, have been explored across multiple datasets. Studies such as [6] show that when trained on well-curated eye-tracking data, these models often achieve accuracies exceeding **90%**. Vajs et al. [12] evaluated traditional machine learning models on eye-tracking-based VAS event detection features. The best overall accuracy of **88.9%**, was achieved by a Support Vector Machine (SVM) model, while Logistic Regression yielded the highest specificity of **92.7%**, highlighting its effectiveness in distinguishing between dyslexic and non-dyslexic readers.

Despite progress, existing models still face challenges such as data imbalance, overfitting, and lack of explainability. To address these gaps, our work proposes an XGBoost-based classifier trained on a synthesized, balanced dataset derived from eye-tracking features. Rather than relying on a fixed feature dataset, the model dynamically identifies and leverages the most informative features during training.

III. Methodology

A. Dataset: ETDD70 Eye-Tracking Dyslexia Dataset

This study makes use of the ETDD70 dataset [14], which was developed as part of a dyslexia diagnosis project. The dataset contains eye-tracking recordings of 70 Czech children (aged 9 to 10 years), evenly divided between diagnosed dyslexic and non-dyslexic participants. Each participant's eye movements were recorded during three text-reading tasks-syllable reading (Task 1), meaningful text reading (Task 4), and pseudo-text reading (Task 5), each task designed to elicit different cognitive and visual processing behaviors. Eye movements were recorded using a standardized setup and fixations were extracted with the help of i2mc algorithm, which is optimized for children's eye-tracking data.

To overcome the challenge of data scarcity, we focused exclusively on Task 4 (meaningful text reading) and employed synthetic data augmentation to this subset to improve generalization and balance. A total of 140 synthetic subjects (2 per original participant) were generated using a perturbation-based method that retains core visual and temporal patterns. Specifically, horizontal fixation coordinates (`fix_x`) were perturbed with 3–5% noise to simulate natural variability, while vertical alignment (`fix_y`) was preserved to maintain reading structure. Fixation durations were also slightly varied ($\pm 5\%$) while maintaining the original sequence and rhythm. Statistical validation, including Kolmogorov-Smirnov (KS) tests, Mean Absolute Error (MAE) analysis, and Pearson correlation checks, confirmed that the synthetic data retained key distributional characteristics without duplicating any real data. This augmentation expanded our dataset to 210 participants, evenly balanced across dyslexic and non-dyslexic groups, and based exclusively on the meaningful text reading task.

Dataset Justification: We selected the ETDD70 dataset because it is one of the few publicly available and well-annotated eye-tracking resources designed explicitly for dyslexia classification. Its controlled experimental design—featuring standardized text stimuli and uniform recording conditions—ensures consistency across participants, reducing confounding variables common in reading studies. The dataset's focus on Czech children aged 9–10 provides an ideal developmental window for observing reading-related oculomotor patterns before compensatory reading strategies typically emerge in older students. Furthermore, the reading tasks were performed in the native language of participants, which minimizes linguistic interference and isolates visual–attentional markers of dyslexia. Although region-specific, these controlled characteristics make ETDD70 a strong benchmark for evaluating model generalization and data-efficient augmentation techniques.

B. Feature Extraction

For each subject, five metrics were extracted. These were used to engineer nine meaningful features covering a range of spatial, temporal, and behavioral aspects of eye movement. These included the average and variability of fixation durations, mean fixation coorinates, total fixation count, dispersion along both axes, average saccade length, and the number of distinct lines or areas of interest focused on. These features capture various dimensions of reading efficiency and oculomotor control that have been previously associated with dyslexia in literature. Feature values were standardized using z-score normalization to facilitate robust model training and improve convergence.

C. Model Architecture

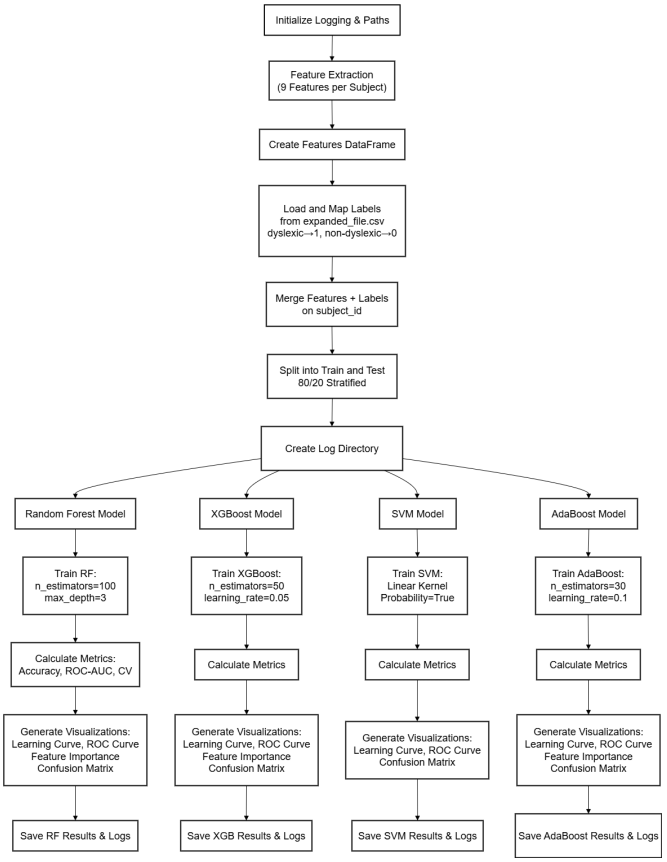

Fig. 1. Overall workflow of the proposed dyslexia detection pipeline, illustrating the complete process from raw fixation logs to XGBoost-based classification output.

Class-wise metrics summary: For the dyslexic class — Precision = **1.00**, Recall = **0.86**, F1 = **0.92**; For the non-dyslexic class — Precision = **0.88**, Recall = **1.00**, F1 = **0.93**.

We implemented an *XGBoost classifier* to distinguish between dyslexic and non-dyslexic participants using nine extracted eye-tracking features. XGBoost builds an ensemble of decision trees trained via gradient boosting, optimizing the following regularized objective function:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} \ell(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{1}$$

where $\ell(\hat{y}_i, y_i)$ is the logistic loss, and $\Omega(f_k)$ is the regularization term for the k^{th} tree:

$$\Omega(f_k) = \gamma T_k + \frac{1}{2}\lambda \sum_{j=1}^{T_k} w_j^2 \tag{2}$$

Here, T_k is the number of leaves in tree k, w_j is the weight of leaf j, and γ and λ are regularization parameters controlling model complexity.

The predicted probability for the dyslexic class is given by the sigmoid function:

$$P(y = 1 \mid \mathbf{x}) = \sigma\left(\sum_{k=1}^{K} f_k(\mathbf{x})\right), \quad \sigma(z) = \frac{1}{1 + e^{-z}} \tag{3}$$

All features were standardized to zero mean and unit variance:

$$x_i' = \frac{x_i - \mu_i}{\sigma_i} \tag{4}$$

Feature selection was guided by Random Forest-based importance scores to retain the most discriminative metrics. The final selected features included:

- Mean and standard deviation of fixation durations
- Total number of fixations
- Mean saccade length
- Fixation-per-line ratio
- Mean and dispersion of fixation coordinates

The final XGBoost model was trained using the following hyper-parameters:

- Learning rate: 0.05
- Max depth: 2
- Number of estimators: 50
- λ (L2 regularization): 1.0
- α (L1 regularization): 0.1
- Min child weight: 3
- `scale_pos_weight`: class imbalance ratio

After training, feature importance analysis revealed that fixation count, average fixation duration, and saccade-related features were the most influential in predicting dyslexia. All modeling and evaluation were conducted using `xgboost` and `scikit-learn` libraries in Python.

IV. Results and Discussion

To evaluate the effectiveness of our dyslexia detection pipeline, we compared many machine learning models such as Support Vector Machine (SVM), Random Forest, AdaBoost, and XGBoost, using nine carefully engineered eye-tracking features that capture both spatial and temporal aspects of reading behavior. As shown in Table I, **XGBoost consistently demonstrated superior performance**, achieving a training

TABLE I
COMPREHENSIVE PERFORMANCE COMPARISON OF ML TECHNIQUES

ML Technique	Train Accuracy	Test Accuracy	AUC (Train)	AUC (Test)	Precision (Dys)	Recall (Dys)	F1 (Dys)	Precision (Non)	Recall (Non)
AdaBoost	87.5%	86%	0.96	0.97	0.80	0.82	0.81	0.87	0.85
Random Forest	93%	86%	0.99	0.96	0.84	0.83	0.83	0.88	0.87
SVM	90%	86%	0.97	0.96	0.83	0.81	0.82	0.86	0.88
XGBoost	**95%**	**93%**	**0.99**	**0.98**	**1.00**	**0.86**	**0.92**	**0.88**	**1.00**

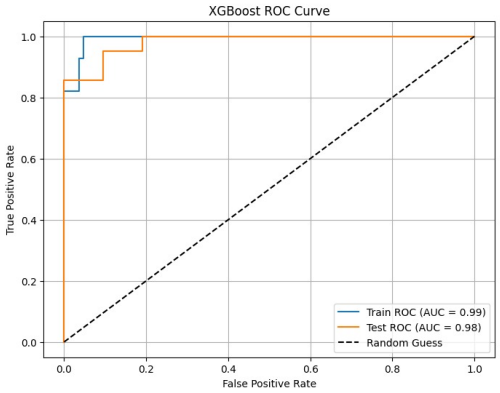

Fig. 2. XGBoost-ROC Curve

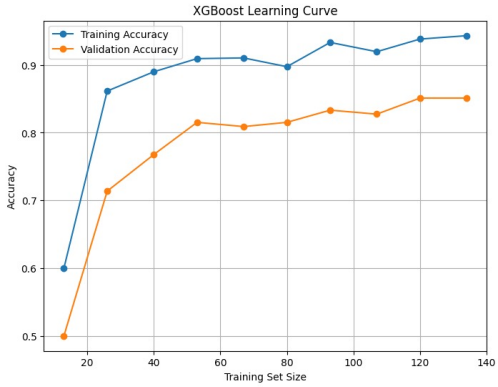

Fig. 3. XGBoost-Learning Curve

accuracy of **94.6%** and a test accuracy of **92.9%**. It also recorded the highest ROC-AUC scores: **0.9928** as shown in, (Figure 2), on the training set and **0.9819** on the test set, showing exceptional class separability. The 5-fold cross-validation method confirmed the model's reliability, giving a mean accuracy of **85.1%** with a standard deviation of **±2.7**. Importantly, the model was trained on a synthetically augmented dataset comprising 105 dyslexic and 105 non-dyslexic samples. Class-wise evaluation exhibited a precision of **1.00** and recall of **0.86** for the dyslexic class, and a precision of **0.88** and recall of **1.00** for the non-dyslexic class, with F1-scores of **0.92** and **0.93**, respectively. The train-test accuracy gap of only **0.0178** indicated minimal overfitting. Learning curve analysis, shown in (Figure 3) revealed progressive improvements in validation accuracy with more data, and training and validation scores remained closely aligned. Feature importance analysis highlighted fixation count, average fixation duration, and saccade length as the most predictive variables, consistent with existing literature. Collectively, these results affirm that our XGBoost-based model, when paired with meaningful gaze-based features, provides a reliable and accurate approach for early dyslexia detection through eye-tracking.

A. Comparison with State-of-the-Art

Our XGBoost model achieves competitive performance while maintaining efficiency and interpretability. DysLexML [7] reported 97% accuracy with three gaze metrics using SVM, RADAR [5] achieved 94% with fixation/saccade metrics, and CNN-based deep models [4] achieved 96.6% using raw gaze data. While our accuracy (93%) is slightly lower, our approach relies only on nine interpretable features and an augmented dataset, providing a strong trade-off between simplicity, efficiency, and predictive strength.

V. CONCLUSION

In summary, our contributions lie in: (i) developing a perturbation-based augmentation method to address small dataset issues, (ii) creating a compact and interpretable nine-feature set capturing critical oculomotor patterns, and (iii) showing that XGBoost not only outperforms other traditional classifiers but also approaches state-of-the-art models in accuracy with fewer features and lower data requirements.

VI. FUTURE WORK

While the proposed XGBoost-based model demonstrates strong performance in classifying dyslexic and non-dyslexic readers, there remain several avenues for further exploration. First, expanding the dataset to include more diverse age groups, linguistic backgrounds, and reading environments would help validate the model's generalizability across populations. Additionally, integrating sequential features such as scanpath patterns and fixation sequences using temporal models like LSTM or Transformer architectures could capture dynamic reading behavior more effectively. Incorporating multimodal data—including EEG or pupil dilation—may further enhance diagnostic precision. Lastly, translating this model into a real-time, accessible screening tool with an intuitive user interface could support early dyslexia detection in educational and clinical settings, especially in low-resource environments.

ACKNOWLEDGMENT

We would like to thank the Center for Healthcare Engineering and Learning (cHEAL) under the Department of Computer

Science and Engineering, PES University, for providing us with the necessary resources, research environment, and institutional support that helped with the successful completion of this project. Their guidance played a crucial role in shaping the direction and outcome of our research.

REFERENCES

[1] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*, 5th ed. Arlington, VA: American Psychiatric Publishing, 2013.

[2] K. R. Pugh *et al.*, "Functional neuroimaging studies of reading and reading disability (developmental dyslexia)," *Ment. Retard. Dev. Disabil. Res. Rev.*, vol. 6, no. 3, pp. 207–213, 2000.

[3] F. Richlan, M. Kronbichler, and H. Wimmer, "Functional abnormalities in the dyslexic brain: A quantitative meta-analysis of neuroimaging studies," *Hum. Brain Mapp.*, vol. 30, pp. 3299–3308, 2009.

[4] Nerušil, B., Polec, J., Škunda, J. et al. Eye tracking based dyslexia detection using a holistic approach. Sci Rep 11, 15687 (2021)

[5] I. Smyrnakis *et al.*, "RADAR: A novel fast-screening method for reading difficulties with special focus on dyslexia," *arXiv preprint arXiv:1611.05669*, 2016.

[6] F. Richlan, "The functional neuroanatomy of developmental dyslexia across languages and writing systems," *Front. Psychol.*, vol. 11, p. 155, 2020.

[7] T. Asvestopoulou *et al.*, "DysLexML: Screening tool for dyslexia using machine learning," *arXiv preprint arXiv:1903.06274*, 2019.

[8] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785-794, doi: 10.1145/2939672.2939785.

[9] A. Jothi Prabha and R. Bhargavi, "Predictive model for dyslexia from fixations and saccadic eye movement events," *Comput. Methods Programs Biomed.*, vol. 195, p. 105538, 2020.

[10] M. N. Benfatto et al., "Screening for dyslexia using eye tracking during reading," *PLoS ONE*, vol. 11, no. 12, p. e0165508, 2016.

[11] Y. K. Meena, H. Cecotti, B. Bhushan, A. Dutta, and G. Prasad, "Detection of dyslexic children using machine learning and multimodal Hindi language eye-gaze-assisted learning system," *IEEE Trans. Hum.-Mach. Syst.*, vol. 53, no. 1, pp. 122–131, 2022.

[12] I. Vajs, T. Papic, V. Kovic, A. M. Savic, and M. M. Jankovic, "Accessible dyslexia detection with real-time reading feedback through robust interpretable eye-tracking features," *Brain Sci.*, vol. 13, no. 3, p. 405, 2023.

[13] Sedmidubsky, J., Dostalova, N., Svaricek, R., & Culemann, W. (2024). ETDD70: Eye-tracking dataset for classification of dyslexia using AI-based methods. In Proceedings of the 17th International Conference on Similarity Search and Applications (SISAP) (pp. 1-14). Springer.

[14] Dostalova, N., Svaricek, R., Sedmidubsky, J., Culemann, W., Sasinka, C., Zezula, P., & Cenek, J. (2024). ETDD70: Eye-tracking Dyslexia Dataset [Data set]. Zenodo. https://doi.org/10.5281/zenodo.13332134