

Explainable and Uncertainty-Aware Eye-Tracking for Dyslexia-Related Reading Support

Asesh Kumar Tripathy¹, Rupesh Ravi M R², Syeda Husna Mehanoor³, Soorya P⁴, Nancy V⁵, Reji R⁶, Gourav Kalra⁷, Javed Ahmad⁸, Dheeraj Tiger⁹, Divya Saleela¹⁰

¹Associate Professor, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India

²Assistant Professor, Thangal Kunju Musaliar Institute of Technology, Kerala, India

³Associate Professor, Malla Reddy Technical Campus, Malla Reddy Vishwavidyapeeth (Deemed to be University), Telangana, India.

⁴Assistant Professor, Thangal Kunju Musaliar Institute of Technology, Kerala, India

⁵Assistant Professor, Easwari Engineering College, Ramapuram, Chennai, India

⁶Professor, Thangal Kunju Musaliar Institute of Technology, Kerala, India

⁷Associate Professor, School of Engineering and Technology, CGC University, Punjab, India

⁸Associate Professor, Sharda University, Greater Noida, India

⁹Research Scholar, The NorthCap University, Gurugram, Haryana, India

¹⁰Assistant Professor, College of Engineering Aranmula, Kerala, India

asesh.tripathy@gmail.com¹, rupeshravi@tkmit.ac.in², drsyedahusna.mehanoor@mrsv.edu.in³, sooryamurali1987@gmail.com⁴, nancy.victorsagayee@gmail.com⁵, rejir@tkmit.ac.in⁶, gkalra89@gmail.com⁷, javed.jmi08@gmail.com⁸, dheerajtiger@gmail.com⁹, divyasal009@cearanmula.ac.in¹⁰

Corresponding Author: divyasal009@cearanmula.ac.in

¹Assistant Professor, College of Engineering Aranmula, Kerala, India
divyasal009@cearanmula.ac.in¹

HOW TO CITE:

Asesh Kumar Tripathy, Rupesh Ravi M R, Syeda Husna Mehanoor, Soorya P, Nancy V, Reji R, Gourav Kalra, Javed Ahmad, Dheeraj Tiger, Divya Saleela (2026). Explainable and Uncertainty-Aware Eye-Tracking for Dyslexia-Related Reading Support. International Journal of Special Education, 41(13s), 460-476

ABSTRACT:

Early identification of dyslexia-related reading difficulties is essential for timely educational support, but automated approaches must remain interpretable, uncertainty-aware and subordinate to professional assessment. This study presents an explainable eye-tracking framework for analysing dyslexia-related reading patterns across syllable, meaningful-text and pseudo-text tasks. A publicly available dataset comprising 70 children, including 35 dyslexic and 35 non-dyslexic readers, was evaluated using participant-independent five-fold cross-validation. Fixation, saccade, pupil and reading-duration features were extracted separately from each task and eye-movement event family. Logistic Regression, Support Vector Machine, Random Forest and Extra Trees classifiers were compared with EduGaze-Select, a task-aware ensemble that combined task-specific predictions using validation-derived fusion weights. Split conformal prediction was incorporated to identify uncertain cases requiring further educational assessment. Logistic Regression achieved the strongest overall performance, with 90.00% accuracy, 90.00% balanced accuracy, an ROC-AUC of 0.9347 and a PR-AUC of 0.9534. EduGaze-Select achieved 82.86% balanced accuracy and an ROC-AUC of 0.9061 while providing task-level interpretability and calibrated referral support. The conformal component attained 94.29% empirical coverage against a nominal target of 90%, produced unique decisions for 74.29% of participants and referred 25.71% for further assessment. These findings demonstrate that explainable and uncertainty-aware eye-tracking analysis can support dyslexia-related reading assessment

COPYRIGHT STATEMENT:

Copyright: © 2026 Authors.
Open access publication under the terms and conditions

of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).	while preserving human oversight and avoiding automated diagnostic claims. Keywords: Dyslexia; eye tracking; reading difficulties; special education; explainable artificial intelligence; uncertainty quantification; educational decision support.
--	--

1. Introduction

Dyslexia is a persistent learning difficulty that affects accurate and fluent word recognition, decoding and spelling. Delayed identification can restrict learners' access to the curriculum, reduce academic confidence and increase the need for intensive educational support [1] [2]. Although formal assessment remains essential, it can be time-consuming and may not always be readily available in school settings. This has encouraged interest in complementary technologies that can support early educational screening while preserving the role of teachers, psychologists and specialist assessors.

Eye tracking provides an objective means of examining reading behaviour through fixations, saccades, regressions, pupil measurements and gaze progression. Previous studies have reported that dyslexic readers may display more frequent or prolonged fixations, shorter forward saccades and increased regressions during reading [3]. These patterns do not indicate that dyslexia is caused by visual impairment, but they can provide useful evidence about decoding effort, reading fluency and text-processing difficulty [4]. Eye-movement information may therefore complement conventional assessment by revealing how a learner processes written material rather than relying only on final reading scores.

Machine-learning research has increasingly used eye-tracking data to distinguish dyslexic and non-dyslexic readers. Early studies mainly combined handcrafted fixation and saccade features with conventional classifiers such as Support Vector Machines, Logistic Regression and Random Forests [5] [6] [7], while later work examined complete gaze signals, spatiotemporal representations, fixation images and deep neural networks [8]. Classification performance is also influenced by the reading task and validation design. Syllable, meaningful-text and pseudo-text tasks place different demands on decoding and

semantic processing, and studies have shown that task choice can materially affect discrimination between reading groups [9]. In addition, datasets commonly contain several recordings per participant; if observations from the same child appear in both training and testing sets, the model may learn participant-specific characteristics rather than generalisable reading patterns.

Despite continued progress, existing studies retain several limitations. Many prioritise accuracy without adequately explaining which reading behaviours influence the prediction. Multiple tasks are often merged without quantifying the contribution of each condition, while small participant samples are sometimes analysed using high-dimensional representations that increase the risk of overfitting. Most importantly, conventional classifiers force every learner into a binary category even when the evidence is uncertain. This is unsuitable for special education, where an ambiguous output should prompt further professional assessment rather than an unsupported automated judgement.

This study addresses these limitations using the ETDD70 dataset, which contains eye-tracking recordings from 70 children, equally divided between dyslexic and non-dyslexic readers, across syllable, meaningful-text and pseudo-text tasks [10]. Fixation, saccade, pupil and reading-duration features are engineered separately for each task and event family. Logistic Regression, Support Vector Machine, Random Forest and Extra Trees classifiers are compared with EduGaze-Select, a task-aware ensemble that combines task-specific predictions using validation-derived fusion weights. All preprocessing and feature selection are conducted within participant-independent validation folds. Split conformal prediction is additionally used to generate non-unique prediction sets for uncertain cases, thereby

supporting specialist referral rather than forced classification [11].

The principal contributions of this study are fourfold. First, it develops a structured representation of eye-tracking behaviour that preserves the distinction between reading tasks and event families. Second, it applies participant-independent five-fold cross-validation with inner validation for fusion-weight estimation and fold-specific feature selection to reduce identity leakage and high-dimensional overfitting. Third, it preserves the contribution of syllable, meaningful-text, pseudo-text and combined-task models through a transparent task-aware fusion mechanism and educationally interpretable feature domains. Fourth, it incorporates calibrated uncertainty as a selective referral mechanism, extending eye-tracking classification from accuracy-centred prediction towards responsible educational decision support. The framework is intended to supplement professional assessment and should not be interpreted as an automated diagnostic system.

The remainder of the paper is organised as follows. Section 2 describes the dataset, preprocessing, feature engineering, classification models, participant-level validation and conformal referral procedure. Section 3 presents the classification, calibration, explainability and uncertainty results. Section 4 discusses the educational implications and study limitations. Section 5 concludes the study and outlines

directions for future validation and broader educational application.

2. Methodology

This section describes the development and evaluation of EduGaze-Select. The methodology comprised dataset acquisition, task- and event-specific preprocessing, participant-level feature engineering, comparative classification, task-aware probability fusion, participant-independent validation, uncertainty estimation and explainability analysis. The complete framework is illustrated in Figure 1.

2.1 Framework Architecture

EduGaze-Select was designed as an explainable and uncertainty-aware framework for analysing dyslexia-related reading behaviour from eye-tracking data. The architecture contains six stages: dataset acquisition, data-quality control, feature extraction, participant-level modelling, task-aware probability fusion and uncertainty-based educational referral.

Recordings from syllable, meaningful-text and pseudo-text tasks were processed through separate fixation, saccade and metric branches. The extracted features were combined at participant level and supplied to conventional classifiers and task-specific ensemble components. Validation-derived weights were then used to combine task-level probabilities, while split conformal prediction converted the final output into either a unique prediction or a recommendation for further professional assessment.

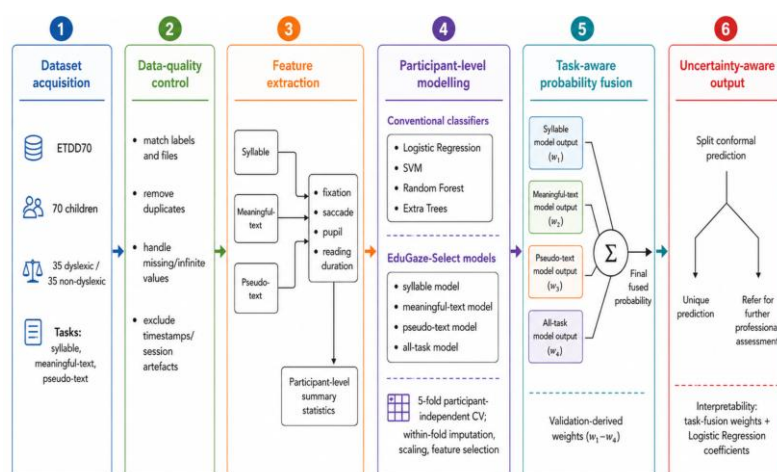


Figure 1. Architecture of the EduGaze-Select framework for explainable and uncertainty-aware dyslexia-related reading support.

2.2 Dataset

The study used the publicly available ETDD70 Eye-Tracking Dyslexia Dataset. It contains recordings from 70 Czech children aged 9–10 years, including 35 dyslexic and 35 non-dyslexic readers. Each participant completed three reading tasks: syllable reading, meaningful-text reading and pseudo-text reading.

The repository provides raw gaze, fixation, saccade and summary-metric files for each participant and task. Only fixation, saccade and metric files were used in this study. Raw sampling-level files were excluded because they contained substantial redundancy and absolute recording timestamps that could introduce session-specific artefacts. Duplicate archive entries were removed using participant, task and file-family identifiers.

2.3 Preprocessing and Feature Engineering

Preprocessing was conducted separately for each reading task and event family. Participant identifiers, task labels and file types were recovered from the file names and matched with the supplied class-label file. Numerical values were converted to consistent formats, infinite values were replaced with missing values, and variables with excessive missingness or no variation were removed.

Absolute timestamps and session start and end times were excluded from the predictor set. Recording duration was retained only as a within-session difference. Fixation files were used to derive fixation count, duration, gaze position, dispersion and pupil features. Saccade files provided event count, duration, amplitude, velocity, direction and start–end position features. Metric files supplied task-level reading duration, fixation frequency, saccade frequency and regression-related measures.

For each valid variable, the count, mean, standard deviation, median, first quartile, third quartile, interquartile range, minimum and maximum were calculated. Every feature retained its task and event-family origin. The final modelling matrix contained one row per participant.

2.4 Classification and Task-Aware Fusion

Four conventional classifiers were evaluated: Logistic Regression, Support Vector Machine, Random Forest and Extra Trees. These models

represented linear, kernel-based and ensemble-learning approaches.

EduGaze-Select contained four components: a syllable model, a meaningful-text model, a pseudo-text model and an all-task model. Each component produced a probability estimate using its corresponding feature set. Inner-validation ROC-AUC values were converted into non-negative fusion weights.

The final probability for participant i was calculated as:

$$\hat{p}_i = \sum_{t=1}^T w_t \hat{p}_{it} \quad (1)$$

where $\hat{p}_{it}$ is the probability produced for participant i by component t , w_t is the validation-derived weight assigned to that component, and T is the total number of model components. The fusion weights were normalised so that their sum equalled one.

2.5 Participant-Independent Validation

Five-fold stratified cross-validation was used to evaluate generalisation to unseen participants. The participant-level feature matrix contained one row per child, and each participant was assigned exclusively to a single cross-validation fold. Within each fold, median imputation, robust scaling, mutual-information feature selection and model fitting were performed using only the training participants.

A maximum of 40 features was retained for the conventional classifiers, while smaller task-specific subsets were selected for the EduGaze-Select components. This fold-specific procedure prevented information from the test participants from influencing preprocessing, feature selection or model fitting.

Performance was evaluated using accuracy, balanced accuracy, precision, sensitivity, specificity, F1-score, Matthews correlation coefficient, ROC-AUC and PR-AUC. Calibration was assessed using the Brier score, log loss and expected calibration error.

2.6 Uncertainty and Explainability

Split conformal prediction was used to avoid forcing a binary decision for every participant. Within each outer cross-validation fold, the development data were divided into a proper-training subset and a calibration subset. The task-aware ensemble was fitted only on the proper-

training participants, and nonconformity scores were then calculated using the held-out calibration participants. A separate conformal threshold was estimated within each outer fold and applied only to the corresponding unseen test participants.

Prediction sets were generated at a nominal coverage level of 90%. A singleton prediction set represented a unique model decision, whereas a two-class or empty prediction set indicated insufficient certainty and triggered referral for further professional assessment. The numbers of singleton, two-class and empty prediction sets were recorded separately.

Model interpretation was supported through task-fusion weights and standardised Logistic Regression coefficients. The coefficients retained their direction, indicating whether each feature increased or decreased the predicted probability of dyslexia-related reading difficulty. The selected features were mapped to educationally meaningful domains such as reading fluency, fixation behaviour, visual progression, re-reading and task-related effort. The explainability analysis was intended to support interpretation of model outputs and not to provide an independent diagnosis of dyslexia.

3. Results

This section presents the participant-level classification, calibration, uncertainty and explainability results obtained from the five-fold out-of-fold predictions.

3.1 Classification Performance

Logistic Regression achieved the strongest overall performance, with 90.00% accuracy, 90.00% balanced accuracy, 88.57% sensitivity, 91.43% specificity, an F1-score of 0.8986, an ROC-AUC of 0.9347 and a PR-AUC of 0.9534. The Support Vector Machine achieved 85.71% balanced accuracy, followed by Random Forest at 84.29%. Extra Trees and EduGaze-Select each achieved 82.86% balanced accuracy. EduGaze-Select produced an ROC-AUC of 0.9061 and a PR-AUC of 0.9369.

Figure 2 compares the participant-level discrimination performance of the evaluated models. The receiver operating characteristic curves in Figure 2(a) show that Logistic Regression achieved the highest ROC-AUC, while the Support Vector Machine and EduGaze-Select also produced values above 0.90. The precision–recall curves in Figure 2(b) show that Logistic Regression achieved the highest average precision of 0.953, followed by the Support Vector Machine at 0.945, EduGaze-Select at 0.937, Random Forest at 0.932 and Extra Trees at 0.931.

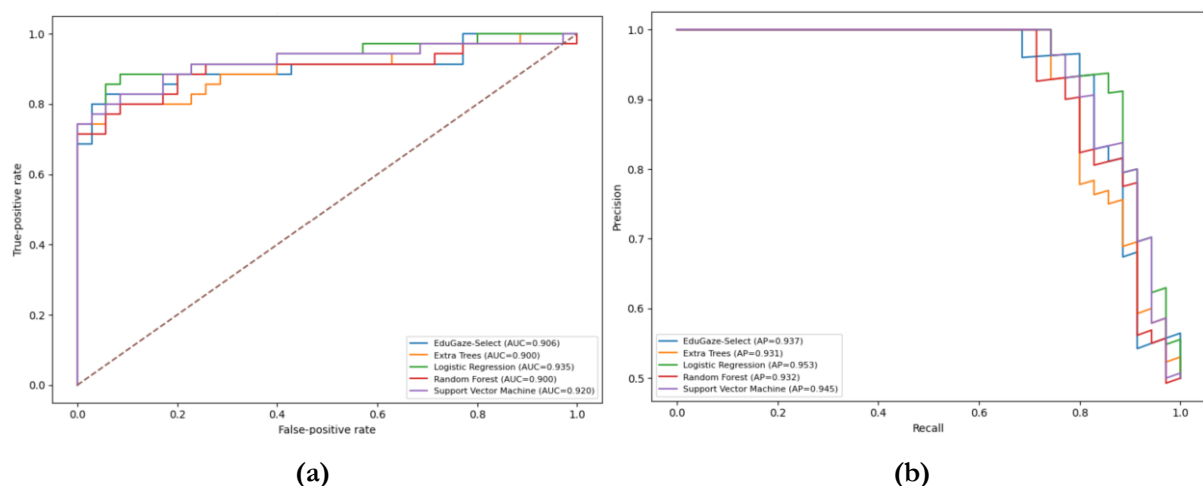


Figure 2. Participant-level discrimination performance of the evaluated models: (a) receiver operating characteristic curves and (b) precision–recall curves.

3.2 Probability Calibration

Logistic Regression produced the lowest Brier score of 0.0932 and the lowest log loss of 0.3067,

while the Support Vector Machine achieved the lowest expected calibration error of 0.0691. EduGaze-Select obtained a Brier score of 0.1167,

a log loss of 0.3799 and an expected calibration error of 0.1334. The calibration patterns of all evaluated models are shown in Figure 3.

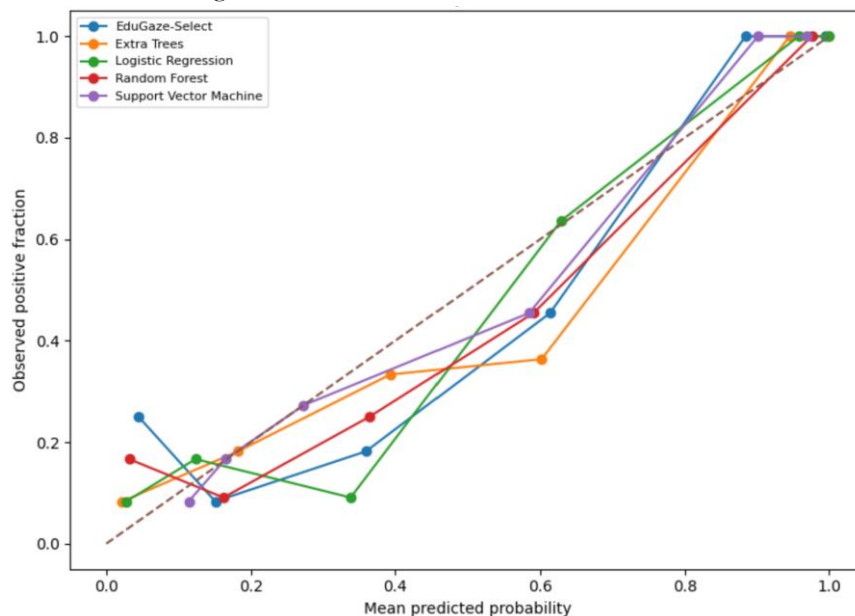


Figure 3. Calibration curves for participant-level dyslexia-related reading-pattern predictions.

3.3 EduGaze-Select Predictions and Referral

EduGaze-Select achieved equal sensitivity and specificity of 82.86%, indicating balanced

identification of dyslexic and non-dyslexic readers. Its out-of-fold classification pattern is shown in Figure 4.

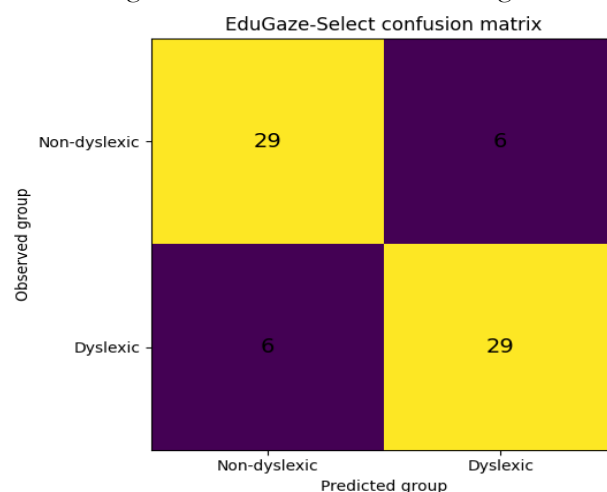


Figure 4. Confusion matrix for the out-of-fold EduGaze-Select predictions.

At a nominal conformal coverage target of 90%, the framework achieved 94.29% empirical coverage, corresponding to 66 of 70 participants. Singleton prediction sets were produced for 52 participants (74.29%), while 18 participants (25.71%) received two-class prediction sets and were referred for further professional assessment. No empty prediction sets were

produced. The mean prediction-set size was 1.257.

3.4 Explainability

The strongest predictors were meaningful-text reading dwell time and fixation count, followed by pseudo-text first-visit fixation duration and dwell-time measures. These task- and event-specific features retained their original measurement context and were interpreted in

relation to reading fluency, re-reading behaviour and task-related effort, as illustrated in Figure 5.

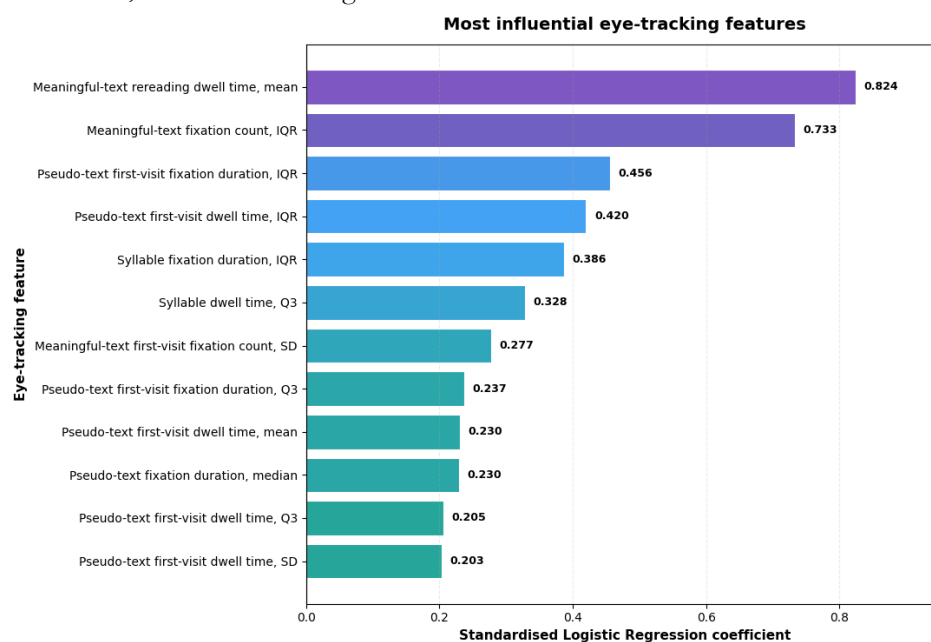


Figure 5. Standardised Logistic Regression coefficients for the most influential task- and event-specific eye-tracking features.

4. Discussion

The findings demonstrate that participant-level eye-tracking features can distinguish dyslexic and non-dyslexic reading patterns with strong performance. Logistic Regression achieved the highest accuracy of 90.00% and an ROC-AUC of 0.9347, indicating that carefully engineered task-specific features can provide effective discrimination without requiring a highly complex model.

The results also show the value of preserving the origin of features across syllable, meaningful-text and pseudo-text tasks. This task-specific structure supports clearer interpretation of the reading behaviours associated with model decisions and avoids treating all eye-movement measurements as a single undifferentiated representation.

EduGaze-Select did not exceed Logistic Regression in classification accuracy, but it provided additional educational value through task-aware fusion and uncertainty-based referral. Feature-level interpretation was provided separately using standardised Logistic Regression coefficients. Split conformal prediction achieved 94.29% empirical coverage, produced unique decisions for 74.29% of participants and referred 25.71% for further professional assessment. This

selective approach is appropriate for special education because it avoids forcing a categorical decision when the model evidence is insufficient. The findings should nevertheless be interpreted cautiously because the analysis was based on 70 participants and a high-dimensional engineered feature space. Although participant-independent validation and fold-specific feature selection reduced the risk of leakage and overfitting, the results remain specific to the available dataset and should be viewed as evidence of potential educational decision support rather than independent diagnostic validity.

5. Conclusion and Future Work

This study developed an explainable and uncertainty-aware eye-tracking framework for dyslexia-related reading support. Participant-level analysis across syllable, meaningful-text and pseudo-text tasks showed that engineered fixation, saccade, pupil and reading-duration features could distinguish dyslexic and non-dyslexic reading patterns with strong performance. Logistic Regression achieved the best overall classification result and supported feature-level interpretation through standardised coefficients, while EduGaze-Select provided task-aware probability fusion and calibrated

referral for uncertain cases. The conformal component enabled the framework to withhold a unique decision when uncertainty was high, thereby supporting professional oversight and reducing the risk of unsupported automated classification.

Future work should validate the framework using larger, multilingual and independently collected datasets. Further studies should examine

performance across different age groups, eye-tracking devices and reading materials, and investigate the integration of gaze features with phonological awareness, reading attainment and teacher observations. These extensions would help determine whether the framework can support broader educational assessment while remaining transparent, reliable and appropriately subordinate to specialist judgement.

References

- M. J. Snowling, C. Hulme, and K. Nation, 'Defining and understanding dyslexia: past, present and future', *Oxf. Rev. Educ.*, vol. 46, no. 4, pp. 501–513, Jul. 2020, doi: 10.1080/03054985.2020.1765756.
- J. Miciak and J. M. Fletcher, 'The Critical Role of Instructional Response for Identifying Dyslexia and Other Learning Disabilities', *J. Learn. Disabil.*, vol. 53, no. 5, pp. 343–353, Sep. 2020, doi: 10.1177/0022219420906801.
- E. I. Toki, 'Using Eye-Tracking to Assess Dyslexia: A Systematic Review of Emerging Evidence', *Educ. Sci.*, vol. 14, no. 11, p. 1256, Nov. 2024, doi: 10.3390/educsci14111256.
- T. K. C. Nguyen, D. D. Le, T. H. Le, T. C. H. Nguyen, and T. D. Ngo, 'The use of eye tracking in supporting individuals with dyslexia: a review', *Disabil. Rehabil. Assist. Technol.*, vol. 20, no. 5, pp. 1183–1198, Jul. 2025, doi: 10.1080/17483107.2024.2437697.
- 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, Oct. 2020, doi: 10.1016/j.cmpb.2020.105538.
- B. Nerušil, J. Polec, J. Škunda, and J. Kačur, 'Eye tracking based dyslexia detection using a holistic approach', *Sci. Rep.*, vol. 11, no. 1, p. 15687, Aug. 2021, doi: 10.1038/s41598-021-95275-1.
- P. Haller, A. Säuberli, S. Kiener, J. Pan, M. Yan, and L. Jäger, 'Eye-tracking based classification of Mandarin Chinese readers with and without dyslexia using neural sequence models', in *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, Abu Dhabi, United Arab Emirates (Virtual): Association for Computational Linguistics, 2022, pp. 111–118. doi: 10.18653/v1/2022.tsar-1.10.
- R. Svaricek, N. Dostalova, J. Sedmidubsky, and A. Cernek, 'INSIGHT: Combining Fixation Visualisations and Residual Neural Networks for Dyslexia Classification From Eye-Tracking Data', *Dyslexia*, vol. 31, no. 1, p. e1801, Feb. 2025, doi: 10.1002/dys.1801.
- H. W. Catts, C. W. Haynes, and R. M. Joshi, 'Defining dyslexia: 2025 revision', *Ann. Dyslexia*, Apr. 2026, doi: 10.1007/s11881-026-00363-4.
- J. Sedmidubsky, N. Dostalova, R. Svaricek, and W. Culemann, 'ETDD70: Eye-Tracking Dataset for Classification of Dyslexia Using AI-Based Methods', in *Similarity Search and Applications*, vol. 15268, E. Chávez, B. Kimia, J. Lokoč, M. Patella, and J. Sedmidubsky, Eds, in *Lecture Notes in Computer Science*, vol. 15268, Cham: Springer Nature Switzerland, 2025, pp. 34–48. doi: 10.1007/978-3-031-75823-2_3.
- A. N. Angelopoulos and S. Bates, 'Conformal Prediction: A Gentle Introduction', *Found. Trends® Mach. Learn.*, vol. 16, no. 4, pp. 494–591, Mar. 2023, doi: 10.1561/22000000101.

Author Declaration

The authors confirm that this manuscript presents original work that has not been published previously and is not under consideration elsewhere. All sources have been appropriately acknowledged. The authors declare that they have no conflicts of interest.