ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

Approaching Early Dyslexia Prediction with Ensemble Machine Learning: A Secondary Empirical Analysis of the Rello Et Al. Public Behavioural Dataset

Patrick Ufomba Nwogu, Prof. Chigozirim Ajaegbu; Dr. Faruk Umar Ambursa; Dr. Femi Adeluyi

(PhD Researcher on Management Information Systems) National Open University of Nigeria

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150800116

Received: 03 September 2026; Accepted: 08 September 2026; Published: 19 September 2026

ABSTRACT

Dyslexia is a specific learning disorder associated with persistent difficulties in reading accuracy, decoding, spelling and fluent word recognition. Early identification is important because timely educational support can reduce the educational consequences of undetected reading difficulties. The increasing availability of digital behavioural data provides an opportunity to complement conventional assessment with machine-learning-based screening approaches. This study presents a secondary empirical investigation of ensemble machine learning for early dyslexia prediction using the publicly available behavioural dataset originally reported by Rello et al. The study uses the Dyt-desktop dataset containing 3,644 observations and 196 predictive variables derived from demographic characteristics and performance across 32 linguistic tasks. The target distribution consists of 392 dyslexia cases and 3,252 non-dyslexia cases, producing a substantial class imbalance. Three tree-based ensemble learners—Random Forest (RF), Extreme Gradient Boosting (XGBoost) and Extra Trees (ET)—were evaluated. Random-Forest Recursive Feature Elimination (RF-RFE) was additionally investigated as a feature-selection strategy, while probability-level stacking combined predictions from RF, XGBoost and ET. Models were evaluated using accuracy, precision, recall, specificity, F1-score, ROC-AUC and Matthews Correlation Coefficient (MCC). XGBoost produced the strongest individual baseline performance, achieving 90.64% accuracy, 64.41% precision, 29.08% recall, 98.06% specificity, 40.07% F1-score, 0.8845 ROC-AUC and 0.3912 MCC. RF-RFE + XGBoost improved performance to 90.81% accuracy, 31.89% recall, 42.74% F1-score, 0.8858 ROC-AUC and 0.4122 MCC. The heterogeneous stacking model achieved substantially higher recall at 71.17%, together with an F1-score of 51.71% and MCC of 0.4644, although accuracy and specificity declined to 85.70% and 87.45%, respectively. The findings demonstrate that ensemble architecture materially affects the operating characteristics of dyslexia prediction. In particular, accuracy-oriented models and sensitivity-oriented screening models represent different deployment objectives. The study contributes a secondary empirical evaluation of the Rello dataset and demonstrates that heterogeneous ensemble learning can provide a potentially useful sensitivity-oriented screening strategy. Nevertheless, the findings should be interpreted as predictive screening evidence rather than clinical diagnosis, and independent external validation, threshold optimization, calibration and explainable artificial intelligence are required before practical deployment.

Keywords: Dyslexia; Ensemble Machine Learning; Secondary Empirical Study; Random Forest; XGBoost; Extra Trees; RF-RFE; Stacking; Behavioural Data; Early Prediction; Educational Artificial Intelligence.

INTRODUCTION

Dyslexia is a specific learning disorder characterized by persistent difficulties with accurate and fluent word recognition, decoding and spelling [1]. It is one of the most widely investigated learning difficulties and can have significant consequences when identification and appropriate support are delayed. Importantly, dyslexia should not be interpreted as a measure of intelligence or motivation; rather, it concerns specific difficulties associated with reading and language processing.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

Early identification is particularly important in educational environments. Learners who experience persistent reading difficulties may require additional assessment and targeted educational support. Conventional dyslexia assessment remains essential and normally involves standardized educational, linguistic, cognitive and behavioural measures. However, professional assessment may require substantial time and specialist expertise. Consequently, computational methods capable of supporting early screening have received increasing research attention.

The rapid growth of digital educational environments has created new opportunities for machine learning (ML). Learners interact with digital systems through measurable actions such as clicks, response accuracy, task completion, errors and response patterns. These behavioural signals may contain information associated with underlying learning characteristics. Rello et al. [2] demonstrated this possibility through an online gamified dyslexia assessment. Their study included more than 3,600 participants and reported that the predictive model correctly detected more than 80% of participants with dyslexia. They subsequently evaluated the approach using a separate dataset containing more than 1,300 participants in a different tablet-based environment.

The Rello et al. dataset is therefore particularly valuable for secondary research. The original study established the feasibility of machine-learning-based dyslexia screening, while subsequent researchers can investigate alternative algorithms, feature-selection strategies, ensemble architectures and evaluation criteria using the publicly available data. The original publication makes the data openly accessible through the associated Kaggle dataset.

The present research builds upon the author's earlier work on ensemble machine learning for early dyslexia prediction. Rather than treating a single algorithm as sufficient, the study examines complementary tree-based ensemble learners. Random Forest (RF) represents bagging-based learning [3], XGBoost represents gradient boosting [4], and Extra Trees (ET) provides an additional randomized tree ensemble [5]. Their differences in learning strategy create the possibility that they will make different prediction errors.

A major methodological challenge is the substantial class imbalance in the dataset. The uploaded Dyt-desktop dataset contains 3,644 observations, of which 392 are dyslexia cases and 3,252 are non-dyslexia cases. Consequently, dyslexia cases constitute only 10.76% of the observations. A classifier that predicts every observation as non-dyslexic would achieve approximately 89.24% accuracy while failing to identify any dyslexia case.

This characteristic creates an important methodological problem. Accuracy alone may provide a misleading representation of screening effectiveness. A model with 90% accuracy may still have poor clinical or educational screening utility if it fails to identify learners belonging to the minority dyslexia class.

The present study consequently evaluates accuracy together with precision, recall, specificity, F1-score, ROC-AUC and MCC.

The research objectives are to:

1. conduct a secondary empirical analysis of the Rello et al. public dyslexia dataset;
2. evaluate RF, XGBoost and ET for dyslexia prediction;
3. investigate the contribution of RF-RFE feature selection;
4. examine probability-level stacking of heterogeneous learners;
5. compare accuracy-oriented and sensitivity-oriented performance;
6. determine whether ensemble learning provides useful predictive diversity; and

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

7. identify methodological requirements for future AI-assisted dyslexia screening.

The principal research question is:

How effectively can heterogeneous ensemble machine-learning models predict dyslexia from behavioural learning data, and which ensemble configuration provides the most appropriate balance between overall predictive performance and sensitivity to dyslexia cases?

BACKGROUND AND RELATED LITERATURE

Dyslexia and Early Identification

Dyslexia is commonly characterized by difficulties in accurate and/or fluent word recognition together with poor spelling and decoding ability [1]. Its effects may become increasingly visible as educational demands increase. Early identification can allow appropriate instructional strategies to be introduced before reading difficulties become entrenched.

The use of computational approaches does not replace professional assessment. Instead, predictive models may serve as preliminary screening tools capable of identifying learners who could benefit from more detailed assessment.

The Rello et al. study is particularly important in this context. Their online gamified assessment demonstrated that behavioural interaction data can support dyslexia prediction [2]. The study was conducted at considerable scale, with more than 3,600 participants, and subsequently tested the methodology in another environment.

This provides a strong basis for secondary empirical analysis.

Behavioural Data and Digital Assessment

Traditional dyslexia assessment may involve reading tests, spelling measures, phonological assessments and professional interpretation. Digital assessments introduce additional behavioural variables.

The Dyt-desktop dataset contains demographic variables and repeated performance measures across linguistic tasks. Variables include Gender, Nativelang, Otherlang, Age, Clicks, Hits, Misses, Score, Accuracy and Missrate.

Such variables can capture different dimensions of task behaviour.

For example:

- Clicks can reflect interaction behaviour;
- Hits indicate successful responses;
- Misses indicate unsuccessful responses;
- Score summarizes task performance;
- Accuracy represents correctness; and
- Missrate provides an error-oriented performance measure.

Because these variables are measured repeatedly across linguistic tasks, the resulting feature space contains potentially useful information but also substantial redundancy.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

Machine Learning for Dyslexia Detection

Machine learning has been investigated using behavioural, handwriting, eye-tracking, EEG, neuroimaging and other modalities [6].Behavioural data have practical advantages for educational applications because they can potentially be collected through ordinary digital interactions without requiring specialized clinical equipment.

Rello et al. [2] demonstrated the feasibility of this approach through an online gamified test. Their research reported that machine learning could detect a substantial proportion of participants with dyslexia and subsequently showed robustness across a different device and testing environment.

The present study extends this line of research by applying multiple ensemble architectures to the same general dataset.

Ensemble Machine Learning

Random Forest

Random Forest is a bagging-based ensemble introduced by Breiman [3]. It constructs multiple decision trees using randomized samples and feature subsets and combines their predictions.

For a classification problem:

[_{RF}=mode(T_1(X), T_2(X),…T_B(X))]

where (T_b) represents the (b)-th decision tree.

Random Forest is suitable for structured data because it can capture nonlinear relationships and interactions without requiring explicit specification of the functional form.

XGBoost

XGBoost is a scalable implementation of gradient boosting introduced by Chen and Guestrin [4].

The model constructs an additive prediction:

[*i*={k=1}^{K}f_k(x_i)]

where each (f_k) represents a decision-tree learner.

Unlike bagging, boosting constructs trees sequentially, allowing later trees to concentrate on errors made by earlier learners.

The original analysis used XGBoost with 250 estimators, a learning rate of 0.05 and maximum depth of five.

Extra Trees

Extra Trees, introduced by Geurts et al. [5], increases randomization during tree construction. Greater randomization can reduce correlation among trees and potentially improve generalization.

For the present study, ET provides an important comparison with RF because both are tree ensembles but employ different randomization mechanisms.

Stacking

Stacking combines heterogeneous learners through a second-level meta-model [7].

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

For this study:

[M=[P_{RF} ,P_{XGB},P_{ET}]]

where (P_{RF}), (P_{XGB}) and (P_{ET}) represent predicted probabilities.

A logistic-regression meta-classifier can then be expressed as:

[P(Y=1)= {1+e^{-(_0+*1P*{RF}+ *2P*{XGB}+*3P*{ET})}}]

The principal advantage is that the meta-classifier can learn when predictions from one base learner should be weighted differently from those of another.

Dataset and Secondary Research Design

Source Dataset

The empirical study uses the uploaded **Dyt-desktop.csv** dataset associated with the Rello et al. dyslexia research.

The original publication describes the dataset as part of research into prediction of dyslexia risk using an online gamified assessment. The published data are openly available, supporting independent analysis and reproducibility.

The present work is explicitly a **secondary empirical study**. It does not claim to have collected a new participant cohort.

Dataset Characteristics

The uploaded dataset contains:

Characteristic	Value
Observations	3,644
Predictors	196
Dyslexia cases	392
Non-dyslexia cases	3,252
Dyslexia prevalence	10.76%
Non-dyslexia prevalence	89.24%
Linguistic tasks	32
Target	Binary

These characteristics are consistent with descriptions of the public dataset in published secondary research. A separate published study also describes the dataset as comprising 3,644 subjects and 196 attributes representing dyslexic and non-dyslexic classes.

The uploaded research material likewise identifies the 196 predictors and 32 linguistic tasks.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

Class Imbalance

The class distribution is:

[P(Dyslexia)==0.1076]

and:

[P(NonDyslexia)==0.8924.]

Thus, the dataset contains approximately 8.3 non-dyslexia observations for every dyslexia observation.

This imbalance is a central issue in the study.

If all participants were classified as non-dyslexic:

[Accuracy==89.24%.]

Therefore, an accuracy close to 89% is not necessarily evidence of a useful dyslexia screening system.

METHODOLOGY

Overall Experimental Framework

The research workflow is:

[Dyt Dataset Data Preprocessing Stratified Validation Feature Processing Base Ensemble Models RF-RFE Performance Evaluation.]

The earlier research pipeline employed preprocessing, categorical encoding, class weighting and stratified five-fold cross-validation.

Data Preprocessing

The preprocessing procedure consists of:

1. identification of the binary target;
2. conversion of numeric-looking variables;
3. treatment of missing numerical observations;
4. treatment of missing categorical observations;
5. categorical encoding;
6. class-weighted training;
7. stratified validation; and
8. fold-specific feature selection.

The use of fold-specific preprocessing is important because information from validation observations should not influence model fitting.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

Feature Representation

The original feature space contains 196 predictors.

These predictors combine demographic information and repeated task-performance variables.

Because the same performance indicators are repeated across linguistic tasks, the feature space may contain correlated or redundant predictors.

This motivates the use of RF-RFE.

Random-Forest Recursive Feature Elimination

Recursive Feature Elimination ranks predictors according to model-derived importance and progressively removes less important predictors.

The procedure is:

[X {196} RF Ranking Feature Elimination X_{selected}.]

The earlier work used RF-RFE to obtain a reduced feature representation. The research documentation identifies approximately 60 transformed predictors after feature selection.

Feature selection may improve:

- computational efficiency;
- model simplicity;
- resistance to irrelevant variables;
- generalization; and
- interpretability.

However, feature selection should not automatically be assumed to improve predictive performance. Its contribution must be evaluated empirically.

Performance Measures

Seven principal performance measures were used.

Accuracy

[Accuracy]

Accuracy measures the proportion of correctly classified observations.

Precision

[Precision]

Precision indicates the proportion of predicted dyslexia cases that are actually dyslexia cases.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

Recall

[Recall]

Recall is particularly important for screening because it measures the proportion of actual dyslexia cases detected.

Specificity

[Specificity]

Specificity measures the ability to correctly identify non-dyslexia cases.

F1-score

[F1= {Precision+Recall}]

F1 provides a balance between precision and recall.

ROC-AUC

ROC-AUC summarizes discrimination across classification thresholds.

A value closer to one indicates stronger separation between the two classes.

Matthews Correlation Coefficient

MCC incorporates TP, TN, FP and FN and is useful under class imbalance.

Because the dyslexia class represents only 10.76% of the data, MCC is an important complementary metric.

EMPIRICAL RESULTS

Base-Model Results

The previously established empirical analysis of the uploaded dataset produced the following results.

Table 1. Performance of Base and Ensemble Models

Model	Accuracy	Precision	Recall	Specificity	F1	ROC-AUC	MCC
Random Forest	89.65%	74.19%	5.87%	99.75%	10.87%	0.8773	0.1896
XGBoost	90.64%	64.41%	29.08%	98.06%	40.07%	0.8845	0.3912
Extra Trees	89.54%	78.95%	3.83%	99.88%	7.30%	0.8709	0.1593
RF-RFE + RF	90.07%	73.44%	11.99%	99.48%	20.61%	0.8813	0.2705

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

RF-RFE + XGBoost	90.81%	64.77%	31.89%	97.91%	42.74%	0.8858	0.4122
RF-RFE + ET	89.96%	69.12%	11.99%	99.35%	20.43%	0.8771	0.2597
RF-XGB-ET Stacking	85.70%	40.61%	71.17%	87.45%	51.71%	0.8839	0.4644

Source: secondary analysis of the uploaded Dyt-desktop dataset.

Random Forest

Random Forest achieved 89.65% accuracy and 99.75% specificity.

However, recall was only 5.87%.

This means that despite correctly classifying most non-dyslexia observations, the model identified only a small fraction of dyslexia-positive observations at the evaluated classification threshold.

Its ROC-AUC of 0.8773 indicates that the model nevertheless possesses substantial ranking discrimination.

This distinction is important.

A model can have good discrimination but poor recall at a particular classification threshold.

XGBoost

XGBoost achieved the strongest individual base-model performance.

Its results were:

- Accuracy = 90.64%;
- Precision = 64.41%;
- Recall = 29.08%;
- Specificity = 98.06%;
- F1 = 40.07%;
- ROC-AUC = 0.8845; and
- MCC = 0.3912.

Compared with Random Forest, recall increased from 5.87% to 29.08%.

The increase is:

[29.08-5.87=23.21]

percentage points.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

XGBoost therefore provides substantially better minority-class identification than RF at the evaluated threshold.

Extra Trees

Extra Trees achieved:

- Accuracy = 89.54%;
- Precision = 78.95%;
- Recall = 3.83%;
- Specificity = 99.88%;
- F1 = 7.30%;
- ROC-AUC = 0.8709; and
- MCC = 0.1593.

Its precision and specificity are strong, but recall is extremely low.

Thus, ET is conservative in its positive predictions.

For a high-specificity application, this characteristic may be useful. For broad dyslexia screening, however, the low recall is problematic.

Effect of RF-RFE

RF-RFE was applied to investigate whether reducing the predictor space improves performance.

Table 2. XGBoost Before and After RF-RFE

Metric	XGBoost	RF-RFE + XGBoost	Change
Accuracy	90.64%	90.81%	+0.17 pp
Precision	64.41%	64.77%	+0.36 pp
Recall	29.08%	31.89%	+2.81 pp
Specificity	98.06%	97.91%	−0.15 pp
F1	40.07%	42.74%	+2.67 pp
ROC-AUC	0.8845	0.8858	+0.0013
MCC	0.3912	0.4122	+0.0210

RF-RFE produced improvements across accuracy, precision, recall, F1 and MCC.

The most notable improvement is recall, which increased by 2.81 percentage points.

Although the ROC-AUC increase is small, the improvement in MCC indicates better overall binary classification quality.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

This suggests that feature reduction may remove some redundancy without eliminating the information required for dyslexia prediction.

Stacking Ensemble

The heterogeneous stacking model combines RF, XGBoost and ET.

The result is substantially different from the individual models.

Table 3. Stacking Performance

Metric	Stacking
Accuracy	85.70%
Precision	40.61%
Recall	**71.17%**
Specificity	87.45%
F1-score	**51.71%**
ROC-AUC	0.8839
MCC	**0.4644**

The stacking model achieves 71.17% recall.

Compared with XGBoost:

[71.17-29.08=42.09]

percentage points.

This is the largest improvement observed in the experiment.

The stacking model also produces the highest F1-score and MCC.

However, the improvement in sensitivity comes at a cost.

Accuracy decreases from 90.64% for XGBoost to 85.70%.

Specificity decreases from 98.06% to 87.45%.

Precision decreases from 64.41% to 40.61%.

The stacking model therefore changes the operating point of the prediction system toward greater sensitivity.

DISCUSSION

Principal Finding

The most important finding is that ensemble architecture influences the balance between sensitivity and specificity.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

XGBoost is the strongest individual model, while stacking provides the most sensitivity-oriented prediction.

This distinction is important for dyslexia screening.

If the objective is to minimize false positives, a high-specificity model may be preferred.

If the objective is to minimize false negatives, a high-recall model may be preferred.

For early screening, the second objective can be especially important.

A learner incorrectly classified as non-dyslexic may not receive additional assessment.

By contrast, a false positive screening result can trigger further professional assessment without constituting a diagnosis.

Therefore, the stacking model's lower specificity should not automatically be interpreted as a failure.

Comparison with the Original Rello Study

The original Rello et al. study demonstrated the feasibility of machine-learning-based dyslexia screening using

behavioural data. The researchers reported that their model correctly detected over 80% of participants with dyslexia and subsequently evaluated robustness in another dataset and device environment.

The present research differs in purpose.

The original research established the feasibility of the behavioural approach.

The present secondary study evaluates alternative ensemble configurations using the same public-data research foundation.

Consequently, the results should not be interpreted as a direct replication of the original experiment.

Rather, they represent a secondary computational investigation.

Why Accuracy Alone Is Inadequate

The dataset provides a particularly clear demonstration of why accuracy can be misleading.

The majority-class baseline is approximately 89.24%.

Random Forest achieves 89.65%.

At first glance, this may appear satisfactory.

However, RF recall is only 5.87%.

Thus, its apparent accuracy advantage over the majority classifier provides little evidence of useful dyslexia screening.

XGBoost increases accuracy to 90.64% while also increasing recall to 29.08%.

The stacking ensemble decreases accuracy to 85.70% but increases recall to 71.17%.

If the models were ranked only by accuracy:

[RF-RFE+XGB > XGB > RF > ET > Stacking.]

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

If they were ranked by recall:

[Stacking > XGB > RF-RFE+XGB > RF-RFE+RF/ET > RF/ET.]

If ranked by MCC:

[Stacking > RF-RFE+XGB > XGB > RF > RF-RFE+RF > RF-RFE+ET > ET.]

Therefore, the preferred model depends on the evaluation criterion.

Practical Screening Interpretation

The results support a two-stage screening concept.

Stage 1: Sensitive computational screening

A stacking model can identify learners who have behavioural patterns associated with dyslexia.

Stage 2: Professional assessment

Learners identified as potentially at risk can undergo conventional assessment.

The workflow can be represented as:

[Digital Behaviour ML Screening Risk Classification Professional Assessment Educational Intervention.]

The model should therefore not be presented as a diagnostic replacement.

Implications for Educational AI

The study has several implications for educational technology.

First, behavioural learning data can support automated risk screening.

Second, ensemble learning can provide alternative operating characteristics.

Third, class imbalance must be explicitly addressed.

Fourth, evaluation should incorporate recall and MCC rather than accuracy alone.

Fifth, models intended for educational use should provide interpretable outputs.

A practical system could return:

- predicted dyslexia risk;
- confidence/probability;
- major contributing behavioural features;
- screening threshold;
- recommendation for additional assessment.

This would be more useful than providing a binary prediction alone.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

Explainability Considerations

Although the current empirical analysis focuses primarily on predictive performance, explainability should form part of future development.

SHAP (SHapley Additive exPlanations) provides a framework for explaining individual predictions and global model behaviour [8].

The SHAP representation is:

[f(x)=*0*+{j=1}^ {p}_j]

where (_j) represents the contribution of feature (j).

For the dyslexia problem, SHAP could identify whether variables such as task accuracy, miss rate, hits, misses or other behavioural measures contribute strongly to a model's prediction.

However, predictive importance should not be interpreted as causality.

If a particular behavioural feature receives a high SHAP value, this means that the model uses that feature strongly for prediction; it does not prove that the feature causes dyslexia.

This distinction is especially important in educational and health-adjacent applications.

METHODOLOGICAL RELIABILITY

Several safeguards are necessary to make the findings reproducible.

Stratified Validation

Because the positive class is relatively small, stratification should be maintained across validation folds.

The earlier research used stratified five-fold evaluation.

Leakage Prevention

Feature selection must be conducted using only training observations.

If RF-RFE is performed on the entire dataset before cross-validation, information from validation observations can indirectly influence feature selection.

The correct procedure is:

[Training Fold RF-RFE Model Validation Fold.]

Threshold Optimization

The reported results use the evaluated classification threshold rather than assuming that a single threshold is universally appropriate.

Future experiments should optimize the threshold according to the screening objective.

For example:

[Threshold^*= _t F1(t)]

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

or:

[Threshold^*= _t [Recall(t)]]

subject to an acceptable specificity constraint.

LIMITATIONS

This study has several limitations.

Secondary Dataset

The research uses an existing public dataset and therefore depends on the original study's participant characteristics, measurement procedures and data-collection protocol.

Class Imbalance

Only 10.76% of the observations are dyslexia-positive.

Single Dataset

The models require validation using independent populations before generalization.

Threshold Dependence

Recall, precision and specificity depend on the classification threshold.

Dataset-Specific Behaviour

The models may learn patterns specific to the Dyt assessment environment.

No Clinical Diagnosis

The target variable represents the dataset's dyslexia classification and should not be interpreted as a substitute for comprehensive professional diagnosis.

Explainability

Although ensemble models can achieve strong predictive performance, additional explainability analysis is needed before educational deployment.

Future Research

Future research should extend the study in several directions.

Hyperparameter Optimization

RF, XGBoost and ET should be systematically optimized using randomized or Bayesian search.

Threshold Optimization

Different operating points should be evaluated to identify thresholds appropriate for screening.

Precision-Recall Analysis

Because of class imbalance, precision-recall curves and PR-AUC should complement ROC-AUC.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

External Validation

The optimized model should be evaluated on an independent dyslexia dataset.

Shap Explainability

Global and local SHAP analysis should be performed to identify influential behavioural predictors.

Calibration

Predicted probabilities should be calibrated so that risk scores correspond more closely to observed probabilities.

Fairness Analysis

Model performance should be investigated across relevant demographic groups.

Model Stability

Feature rankings should be evaluated across folds to determine whether the same behavioural variables consistently contribute to prediction.

Recommended Ensemble Architecture

Based on the empirical findings, the proposed architecture is:

This architecture provides a logical extension of the earlier empirical work.

Research Contributions

The study makes five principal contributions.

Contribution 1: Secondary empirical validation

It provides an independent computational investigation using the public behavioural dataset originally reported by Rello et al.

Contribution 2: Heterogeneous ensemble comparison

It compares three different tree-based ensemble mechanisms.

Contribution 3: Feature-selection analysis

It evaluates RF-RFE as a method for reducing the 196-dimensional predictor space.

Contribution 4: Sensitivity-oriented ensemble prediction

It demonstrates that stacking can substantially increase dyslexia recall relative to individual learners.

Contribution 5: Multi-metric evaluation

It demonstrates empirically why accuracy alone is insufficient under substantial class imbalance.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

CONCLUSION

This study presented a secondary empirical investigation of ensemble machine learning for early dyslexia prediction using the publicly available behavioural dataset associated with Rello et al.

The dataset contains 3,644 observations, 196 predictors and 392 dyslexia-positive cases. The 10.76% prevalence of dyslexia creates substantial class imbalance and makes conventional accuracy-based model selection inadequate.

Among the individual base learners, XGBoost provided the strongest overall performance, achieving 90.64% accuracy, 29.08% recall, 40.07% F1-score, 0.8845 ROC-AUC and 0.3912 MCC.

RF-RFE + XGBoost improved performance to 90.81% accuracy, 31.89% recall, 42.74% F1-score, 0.8858 ROC-AUC and 0.4122 MCC.

The most significant result was obtained from the heterogeneous RF-XGBoost-ET stacking model. Stacking achieved 71.17% recall, 51.71% F1-score and 0.4644 MCC, representing substantial improvements in sensitivity-oriented performance.

However, stacking reduced accuracy and specificity to 85.70% and 87.45%, respectively.

The findings consequently demonstrate that ensemble model selection should be driven by the intended application.

For an accuracy-oriented predictive system, RF-RFE + XGBoost provides the strongest configuration among the evaluated models.

For a screening-oriented system in which minimizing missed dyslexia cases is prioritized, the stacking architecture provides a substantially more sensitive operating point.

The study therefore supports the proposition that ensemble machine learning can provide a useful computational foundation for early dyslexia risk screening using behavioural learning data.

Nevertheless, the models should be viewed as **screening-support systems rather than autonomous diagnostic instruments**. Future research should integrate systematic hyperparameter optimization, threshold optimization, PR-AUC, probability calibration, external validation and SHAP-based explainability.

Ultimately, the value of an AI-based dyslexia screening system should not be judged solely by how accurately it predicts a label. Its practical value depends on whether it can reliably identify learners who require further assessment while minimizing unnecessary referrals, remaining interpretable, generalizable and ethically appropriate.

REFERENCES

1. G. R. Lyon, S. E. Shaywitz, and B. A. Shaywitz, “A definition of dyslexia,” Annals of Dyslexia, vol. 53, pp. 1–14, 2003.
2. L. Rello, R. Baeza-Yates, A. Ali, J. P. Bigham, and M. Serra, “Predicting risk of dyslexia with an online gamified test,” PLOS ONE, vol. 15, no. 12, e0241687, 2020.
3. L. Breiman, “Random forests,” Machine Learning, vol. 45, pp. 5–32, 2001.
4. 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, pp. 785–794, 2016.
5. P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” Machine Learning, vol. 63, pp. 3–42, 2006.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

6. Y. Alkhurayyif and A. R. W. Sait, “A review of artificial intelligence-based dyslexia detection techniques,” Diagnostics, vol. 14, no. 21, 2362, 2024.
7. D. H. Wolpert, “Stacked generalization,” Neural Networks, vol. 5, no. 2, pp. 241–259, 1992.
8. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Advances in Neural Information Processing Systems, vol. 30, pp. 4765–4774, 2017.
9. J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization,” Journal of Machine Learning Research, vol. 13, pp. 281–305, 2012.
10. J. Snoek, H. Larochelle, and R. P. Adams, “Practical Bayesian optimization of machine learning algorithms,” in Advances in Neural Information Processing Systems, vol. 25, 2012.
11. F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.