An Explainable Ensemble Machine Learning for Early Detection of Dyslexia

Author: Patrick Nwogu
Affiliation: National Open University of Nigeria
Corresponding Author: Prof.Chigozirim Ajaegu , Dr. Faruk Umar Ambursa , Dr. Femi Adeluyi
Correspondence : punwogu@gmail.com

Abstract

Dyslexia is a specific learning disorder associated with persistent difficulties in accurate and fluent word recognition, decoding and spelling. Early identification can support timely educational intervention, but conventional assessment may be resource-intensive and computational prediction models may introduce an additional challenge: highly accurate machine-learning models can be difficult for educators and other stakeholders to interpret. This study develops an explainable ensemble machine-learning framework for early dyslexia detection using behavioural learning data. The study uses the supplied Dyt-desktop dataset containing 3,644 observations, 196 predictors and 392 dyslexia-positive observations. Random Forest (RF), XGBoost and Extra Trees (ET) were evaluated using stratified five-fold cross-validation, followed by Random-Forest Recursive Feature Elimination (RF-RFE) and probability-level logistic-regression stacking. The empirical results show that XGBoost was the strongest base learner, obtaining 90.64% accuracy, 29.08% recall, 40.07% F1-score, 0.8845 ROC-AUC and 0.3912 Matthews Correlation Coefficient (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 proposed heterogeneous stacking ensemble produced the highest recall (71.17%), F1-score (51.71%) and MCC (0.4644), although its accuracy decreased to 85.70%. The findings demonstrate that ensemble learning can substantially improve sensitivity-oriented dyslexia screening. However, because interpretability is essential when predictive outputs may influence educational decisions, the study proposes SHAP-based global and local explanations as an explainability layer for the validated ensemble architecture. The work therefore contributes an empirically grounded framework combining predictive performance, feature reduction and explainability, while recognizing that machine-learning predictions should support rather than replace professional dyslexia assessment.

Keywords: Dyslexia; Explainable Artificial Intelligence; Ensemble Machine Learning; Random Forest; XGBoost; Extra Trees; RF-RFE; SHAP; Educational Data Mining; Early Detection.

1. Introduction

Dyslexia is a neurodevelopmental learning disorder characterized by persistent difficulties involving word recognition, decoding, spelling and reading fluency [1]. Because difficulties may become more pronounced as academic demands increase, early identification is important for providing appropriate educational support. Conventional identification generally combines educational, behavioural and cognitive assessment. Although such assessment remains essential, the increasing availability of digital learning and behavioural data creates opportunities for machine-learning systems to function as complementary early-warning tools.

Machine learning has been investigated for dyslexia-risk prediction using behavioural and digital assessment information. Rello et al. demonstrated the potential of online gamified behavioural assessment for predicting dyslexia risk [2]. More recent reviews have emphasized that artificial-intelligence approaches to dyslexia detection involve a range of behavioural, cognitive and neuroimaging modalities, while also identifying dimensionality reduction, limited datasets and interpretability as continuing challenges [7].

The use of ensemble machine learning is attractive because different algorithms can learn different aspects of a complex feature space. Random Forest combines multiple randomized decision trees [3], XGBoost uses gradient boosting [4], and Extra Trees introduces additional randomization during tree construction [5]. Combining these learners can therefore provide complementary predictive information.

However, predictive accuracy alone is insufficient for an educational screening system. A teacher or specialist may reasonably ask: **Why was this learner classified as potentially dyslexic? Which behavioural characteristics contributed to the prediction? How strongly did each characteristic influence the result?** These questions motivate Explainable Artificial Intelligence (XAI).

Lundberg and Lee introduced SHAP (SHapley Additive exPlanations) as a unified framework for attributing prediction contributions to individual features [6]. SHAP is particularly relevant to complex models because it can provide both global explanations of feature importance and local explanations of individual predictions.

The present study therefore extends the author's earlier ensemble-machine-learning research by developing an **explainable ensemble framework for early dyslexia detection**. Importantly, the empirical analysis reported here is based on the supplied dataset and previously documented model experiments. SHAP is introduced as the explainability methodology; actual SHAP values are not represented as already computed in this manuscript.

1.1 Research Problem

A major challenge in dyslexia prediction is the coexistence of high-dimensional behavioural data and an imbalanced target variable. The supplied dataset contains 3,644

observations but only 392 positive dyslexia cases. Consequently, a model can achieve apparently high accuracy while identifying very few dyslexia cases. For example, the existing Random Forest result achieves 89.65% accuracy but only 5.87% recall.

A second challenge is model transparency. Complex ensemble models may provide better predictive performance but are more difficult to interpret than simple statistical models. The 2024 review of AI-based dyslexia detection specifically identifies the black-box nature of ML models and the need for robust and interpretable approaches as continuing concerns [7].

This study addresses these challenges through a combination of RF-RFE feature selection, heterogeneous ensemble learning and an explainability framework.

1.2 Objectives

The study seeks to:

1. evaluate RF, XGBoost and Extra Trees for dyslexia prediction;
2. investigate the effect of RF-RFE feature selection;
3. compare individual models with heterogeneous stacking;
4. identify the strongest predictive configuration using multiple evaluation metrics;
5. establish an explainability framework capable of identifying the contribution of behavioural features to predictions; and
6. formulate an interpretable early-warning architecture for dyslexia screening.

2. Related Literature

2.1 Machine Learning for Dyslexia Detection

The application of machine learning to dyslexia detection has expanded as researchers have gained access to behavioural, linguistic, cognitive and neuroimaging data. Rello et al. showed that digital behavioural assessment can contain useful information for identifying dyslexia risk [2]. Such findings support the broader concept of using machine learning as an early-warning mechanism.

However, dyslexia is heterogeneous, and no single behavioural indicator necessarily provides sufficient predictive information. Consequently, models capable of learning nonlinear relationships among multiple variables are attractive.

2.2 Ensemble Learning

Random Forest is an ensemble approach based on multiple decision trees trained using randomized sampling and feature selection [3]. It is particularly useful for structured data because it can represent nonlinear relationships and interactions without requiring a strictly linear relationship between predictors and outcome.

XGBoost is a scalable gradient-boosting framework that constructs an additive sequence of decision trees [4]. Its ability to model nonlinear interactions and complex feature relationships makes it suitable for high-dimensional structured datasets.

Extra Trees, proposed by Geurts et al., introduces additional randomization into tree construction [5]. Although individual tree structures differ from those of conventional Random Forest models, the resulting diversity can make Extra Trees useful as a complementary ensemble component.

2.3 Feature Selection

High-dimensionality is a recurring challenge in dyslexia prediction. The supplied dataset contains 196 predictors, including demographic variables and repeated performance measures from linguistic tasks. Feature selection can reduce redundancy, improve computational efficiency and potentially improve model generalization.

The RF-RFE approach used in the present study employs Random Forest importance information to recursively eliminate less informative variables. This approach is particularly relevant because the objective is not merely dimensionality reduction but identification of predictors that contribute to classification.

Recent literature reviews have emphasized the importance of dimensionality-reduction and feature-selection methods in AI-based dyslexia detection [7].

2.4 Explainable Artificial Intelligence

Explainability is increasingly important when machine-learning outputs affect people. Complex models may produce strong predictive performance while offering limited insight into the reasoning behind individual predictions. XAI attempts to bridge this gap by producing interpretable representations of model behaviour.

SHAP is based on Shapley-value concepts and assigns feature contribution values to predictions [6]. A SHAP explanation can therefore answer two complementary questions:

- **Global explanation:** Which variables are generally most influential across the dataset?
- **Local explanation:** Which variables contributed to a particular learner being classified as high-risk or low-risk?

Lundberg and Lee describe SHAP as a unified framework for additive feature attribution and demonstrate its applicability to complex predictive models.

For dyslexia screening, this distinction is important. A global feature-importance plot can identify influential behavioural variables, whereas a local explanation can help an educator understand why a particular learner received a risk prediction.

3. Materials and Methods

3.1 Dataset

The empirical analysis uses the supplied **Dyt-desktop.csv** dataset. The dataset contains 3,644 observations and 196 predictors. The variables include demographic characteristics and performance measures associated with 32 linguistic tasks. The task-level variables include measures such as clicks, hits, misses, score, accuracy and miss rate.

The target variable is binary, representing dyslexia and non-dyslexia observations.

Table 1. Class Distribution

Class	Frequency	Percentage
Non-dyslexia	3,252	89.24%
Dyslexia	392	10.76%
Total	**3,644**	**100.00%**

The dataset therefore exhibits substantial class imbalance. An all-negative classifier would achieve approximately 89.24% accuracy without detecting any dyslexia-positive cases. Accordingly, accuracy cannot be used as the sole measure of model quality.

3.2 Data Preprocessing

The documented analysis used preprocessing procedures including conversion of numeric-looking fields, treatment of missing values, categorical encoding and class weighting. Numerical missing values were median-imputed, while categorical values were treated using the most-frequent category. The analysis employed stratified five-fold cross-validation to preserve the class distribution between folds.

The use of class weighting was important because dyslexia-positive cases represented only 10.76% of observations.

3.3 Base Learners

Three tree-based algorithms were evaluated:

Random Forest: An ensemble of randomized decision trees [3].

XGBoost: A gradient-boosting algorithm that sequentially improves prediction through tree-based learners [4].

Extra Trees: An extremely randomized tree ensemble designed to introduce additional randomness into tree construction [5].

The documented baseline configuration used 250 trees for RF and ET and 250 estimators for XGBoost, with the latter using a learning rate of 0.05, maximum depth of five, subsampling of 0.90 and feature subsampling of 0.90.

3.4 RF-RFE Feature Selection

RF-RFE was applied to reduce the transformed predictor space. Feature selection was conducted within the training folds rather than on the complete dataset, reducing the possibility of information leakage. The retained feature subset was limited to 60 transformed predictors or approximately one-half of the transformed feature space, subject to the documented minimum rule.

3.5 Stacking Ensemble

The proposed ensemble uses the probability predictions of RF, XGBoost and ET as meta-features:

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

A logistic-regression meta-classifier then estimates:

[P(Y=1)= {1+}]

The default classification threshold was 0.50. The meta-classifier used out-of-fold probabilities to reduce leakage between the base learners and meta-learner.

3.6 Explainability Framework

The explainability layer is designed around SHAP. For each final prediction, SHAP values can be used to decompose the model output into feature-level contributions. The conceptual form is:

[f(x)=*0*+{i=1}^{p}_i]

where (f(x)) is the model output, (_0) is the baseline prediction and (_i) represents the contribution of feature (i).

The proposed analysis has two levels:

1. **Global SHAP analysis** – rank predictors according to mean absolute SHAP values.
2. **Local SHAP analysis** – explain individual dyslexia-risk predictions by identifying features pushing predictions toward or away from the positive class.

This approach follows the SHAP framework developed by Lundberg and Lee [6].

Because SHAP computation was not part of the supplied empirical results, this manuscript does **not** fabricate feature rankings or SHAP numerical values. Instead, the validated predictive results establish the model to which the explainability analysis should subsequently be applied.

3.7 Evaluation Metrics

The models were evaluated using:

- Accuracy
- Precision
- Recall/Sensitivity
- Specificity
- F1-score
- ROC-AUC
- Matthews Correlation Coefficient

For an imbalanced screening problem, recall, F1 and MCC are particularly important because they reveal whether the classifier is actually identifying minority-class cases.

4. Results

Table 2. Empirical Performance of the Evaluated 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
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
Stacked Ensemble	85.70%	40.61%	**71.17%**	87.45%	**51.71%**	0.8839	**0.4644**

Source: empirical analysis of the supplied Dyt-desktop dataset.

4.1 Base-Model Performance

XGBoost achieved the highest accuracy among the three base models at 90.64%. It also obtained the highest recall, F1-score, ROC-AUC and MCC. Its ROC-AUC of 0.8845 indicates strong ranking performance compared with RF and ET.

Random Forest obtained 89.65% accuracy and very high specificity of 99.75%, but its recall was only 5.87%. Extra Trees demonstrated the highest precision at 78.95% and specificity at 99.88%, but its recall was only 3.83%.

These findings demonstrate why accuracy and specificity should not be interpreted independently.

4.2 Effect of RF-RFE

RF-RFE improved all three tree-based models on the principal performance measures.

For XGBoost, RF-RFE increased:

- Accuracy: **90.64% → 90.81%**
- Recall: **29.08% → 31.89%**
- F1: **40.07% → 42.74%**
- ROC-AUC: **0.8845 → 0.8858**
- MCC: **0.3912 → 0.4122**

The improvement in MCC is particularly important because MCC evaluates the complete confusion matrix and is less misleading under class imbalance.

4.3 Ensemble Performance

The stacking ensemble produced a substantially different operating profile.

Compared with XGBoost, recall increased from 29.08% to 71.17%:

[71.17-29.08=42.09]

Thus, the ensemble improved recall by **42.09 percentage points**.

MCC also increased:

[0.4644-0.3912=0.0732]

However, accuracy declined from 90.64% to 85.70%, while specificity decreased to 87.45%.

This is not evidence that the ensemble is universally superior. Instead, it demonstrates that the ensemble is more sensitive to the minority class at the tested threshold.

5. Explainability Analysis and Proposed SHAP Interpretation

The major contribution of the proposed framework is the integration of explainability with ensemble prediction.

5.1 Why Explainability Is Necessary

An educator receiving an ML prediction should not be required to accept the prediction as an unexplained numerical output. An explainable model can provide contextual information about the variables contributing to the risk estimate.

This is particularly important for dyslexia because behavioural measures may be correlated with multiple educational and linguistic factors. A feature that is statistically predictive should not automatically be interpreted as a causal factor.

5.2 Global Explanation

The proposed global SHAP analysis should calculate:

[Importance_j= *{i=1}^{N}*|{ij}|]

where (Importance_j) represents the mean absolute SHAP contribution of feature (j).

The resulting ranking can identify which behavioural and demographic variables contribute most strongly to ensemble predictions.

Importantly, such a ranking would answer **what the model uses**, rather than necessarily answering **what causes dyslexia**.

5.3 Local Explanation

For an individual learner, the SHAP explanation can identify variables that push the prediction toward the dyslexia-positive class and variables that push it toward the non-dyslexia class.

A hypothetical explanation could therefore be represented conceptually as:

[Baseline +_1+_2+_3++_p Risk Score]

The proposed system would enable an educator to inspect the principal contributing variables rather than simply receiving a binary prediction.

5.4 Relationship Between RF-RFE and SHAP

RF-RFE and SHAP serve different purposes.

RF-RFE answers:

Which predictors should remain in the predictive model?

SHAP answers:

> **How do the retained predictors contribute to individual and aggregate predictions?**

Combining the two therefore creates a useful methodological sequence:

[Data Preprocessing RF-RFE Ensemble Prediction SHAP Explanation]

This architecture combines predictive modelling with interpretability while maintaining a clear separation between feature selection, prediction and explanation.

6. Discussion

The empirical findings demonstrate that ensemble learning can provide meaningful improvements for dyslexia screening, but model performance depends strongly on the evaluation objective.

The strongest individual configuration was RF-RFE + XGBoost, which achieved 90.81% accuracy and 0.8858 ROC-AUC. This suggests that feature reduction can improve XGBoost's use of the high-dimensional behavioural feature space.

The stacking ensemble produced the highest recall, F1 and MCC. Its 71.17% recall is especially relevant to early screening because the primary objective of an early-warning system may be to avoid overlooking potentially affected learners.

Nevertheless, the reduction in accuracy from 90.81% for RF-RFE + XGBoost to 85.70% for the stacking model demonstrates that sensitivity comes with a specificity trade-off. Therefore, the appropriate model depends on the operational objective.

Objective	Best configuration
Highest accuracy	**RF-RFE + XGBoost**
Highest ROC-AUC	**RF-RFE + XGBoost**
Strongest base learner	**XGBoost**
Highest precision	**Extra Trees**
Highest specificity	**Extra Trees**
Highest recall	**Stacked Ensemble**
Highest F1	**Stacked Ensemble**
Highest MCC	**Stacked Ensemble**

These results are consistent with the previous empirical analysis of the supplied dataset.

The explainability component strengthens the proposed architecture because predictive performance alone does not establish educational usefulness. Recent XAI research

emphasizes transparency, accountability and fairness as important considerations when complex models are used in consequential domains.

For dyslexia prediction, explanations should therefore be presented as **model reasoning**, not as proof of causation. If a particular task-performance variable has a high SHAP value, this means that the model relied strongly on that variable; it does not demonstrate that the variable caused dyslexia.

7. Implications for Educational AI

The proposed framework can be used as an early-warning mechanism within a broader educational assessment process.

A potential operational workflow is:

[Learner Behavioural Data RF-RFE RF/XGBoost/ET Stacking Risk Prediction
SHAP Explanation Professional Assessment]

Such a system could potentially assist educators in prioritizing learners for additional assessment.

However, the system should **not automatically diagnose dyslexia**. Dyslexia assessment requires appropriate professional and educational evaluation. The ML system should therefore function as a decision-support or screening mechanism.

Explainability can also help identify potential model weaknesses. If SHAP analysis reveals that the model relies disproportionately on demographic variables rather than task-performance measures, this may indicate a need for additional fairness or feature-engineering analysis.

8. Limitations

Several limitations should be recognized.

First, the study uses one supplied dataset. External validation on an independent population is therefore required before generalization.

Second, the dataset is substantially imbalanced, with only 10.76% positive cases. Although class weighting and stratified validation were used, further precision-recall analysis is required.

Third, the reported experiments used a default classification threshold of 0.50. Threshold optimization should be performed inside the validation procedure rather than selected after observing test results.

Fourth, the current manuscript does not claim completed SHAP computations. The explainability framework is methodologically specified, but actual global and local SHAP plots should be generated in the next empirical phase.

Fifth, hyperparameter optimization was not exhaustive. Future work should use nested cross-validation and systematic optimization.

Finally, predictive association should not be confused with causation. Features identified by SHAP as influential should be interpreted as model predictors rather than causal determinants of dyslexia.

9. Future Research

Future research should extend the present study in six directions.

First, the RF-RFE procedure should be subjected to feature-stability analysis to determine whether the same predictors are selected across folds.

Second, SHAP should be computed for the final RF-RFE + XGBoost model and the stacking ensemble. Global SHAP rankings, dependence plots and individual force/waterfall explanations should be reported.

Third, precision-recall AUC should be added because dyslexia is the minority class.

Fourth, bootstrap confidence intervals and paired statistical comparisons should be used to quantify uncertainty around model differences.

Fifth, subgroup analysis should investigate whether model performance varies across relevant demographic and language groups.

Sixth, independent external validation should be performed before deployment.

These recommendations are consistent with the limitations and future-analysis requirements identified in the previous empirical work.

10. Conclusion

This study developed an explainable ensemble-machine-learning framework for early dyslexia detection using the supplied Dyt-desktop behavioural dataset. The empirical results demonstrate that XGBoost is the strongest base learner, while RF-RFE + XGBoost provides the strongest individual configuration, achieving 90.81% accuracy and 0.8858 ROC-AUC. The heterogeneous stacking ensemble provides a different advantage, achieving 71.17% recall, 51.71% F1-score and 0.4644 MCC.

The results demonstrate that the best model depends on the screening objective. RF-RFE + XGBoost is preferable when overall accuracy and ranking performance are emphasized, whereas the stacking ensemble is preferable when sensitivity to dyslexia-positive cases is prioritized.

The principal methodological contribution is the integration of **RF-RFE feature selection, heterogeneous ensemble learning and SHAP-based explainability**. Such integration can move dyslexia prediction beyond a black-box classification task toward a more transparent decision-support framework.

However, the present study deliberately distinguishes **validated empirical findings** from **proposed explainability analysis**. SHAP values should be computed and externally validated before claims are made about the most influential individual predictors. The resulting explainable model should be used as an early-warning mechanism to support professional assessment rather than as an autonomous diagnostic system.

Overall, the study provides an empirical foundation for developing transparent, sensitivity-aware and externally validated AI-assisted dyslexia screening systems.

References

[1] Lyon, G. R., Shaywitz, S. E., & Shaywitz, B. A. (2003). A definition of dyslexia. *Annals of Dyslexia*, 53, 1–14.

[2] Rello, L., Baeza-Yates, R., Ali, A., Bigham, J. P., & Serra, M. (2020). Predicting risk of dyslexia with an online gamified test. *PLOS ONE*, 15(12), e0241687. https://doi.org/10.1371/journal.pone.0241687.

[3] Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.

[4] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

[5] Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63, 3–42.

[6] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.

[7] Alkhurayyif, Y., & Sait, A. R. W. (2024). A review of artificial intelligence-based dyslexia detection techniques. *Diagnostics*, 14(21), 2362.

[8] Paglialunga, A., & Melogno, S. (2025). The effectiveness of artificial intelligence-based interventions for students with learning disabilities: A systematic review. *Brain Sciences*, 15(8), 806.