ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep376

Feature Selection and Ensemble Learning for Robust Dyslexia Prediction: An Empirical Study

Patrick Ufomba Nwogu[1]; Chigozirim Ajaegbu[2]; Dr. Faruk Umar Ambursa[3]; Dr. Femi Adeluyi[4]

[1,2,3,4] (Phd Researcher; Management Information Systems)
National Open University of Nigeria

Publication Date: 2026/09/19

Abstract: Dyslexia is a neurodevelopmental learning difficulty that can affect reading, spelling, decoding and academic achievement. Early identification is important because timely intervention can reduce the consequences of delayed recognition. This study investigates feature selection and ensemble machine learning for robust dyslexia prediction using the publicly available behavioural dataset associated with Rello et al. The uploaded Dyt-desktop dataset contains 3,644 observations, 196 predictor attributes and a binary dyslexia outcome. The proposed analytical framework combines Random Forest Recursive Feature Elimination (RF-RFE) with Random Forest (RF), XGBoost (XGB) and Extremely Randomized Trees (ET), followed by heterogeneous stacking. Models are evaluated using accuracy, precision, recall, specificity, F1-score, ROC-AUC and Matthews correlation coefficient (MCC). The results show that RF-RFE improves individual predictive performance, with RF-RFE-XGBoost obtaining 90.81% accuracy, 31.89% recall, 42.74% F1-score, 0.8858 ROC-AUC and 0.4122 MCC. The stacking ensemble provides substantially stronger dyslexia-class detection, achieving 71.17% recall, 51.71% F1-score and 0.4644 MCC, although specificity decreases to 87.45%. Tuned XGBoost is the strongest individual optimized learner, achieving 91.22% accuracy and 0.8921 ROC-AUC. The findings demonstrate that ensemble learning can provide a more screening-oriented decision boundary than accuracy-focused individual classifiers in an imbalanced dyslexia dataset.

Keywords: *Dyslexia Prediction; Machine Learning; Feature Selection; RF-RFE; Random Forest; XGBoost; Extra Trees; Ensemble Learning; Stacking; Learning Disabilities.*

How to Cite: Patrick Ufomba Nwogu (2026) Feature Selection and Ensemble Learning for Robust Dyslexia Prediction: An Empirical Study. *International Journal of Innovative Science and Research Technology*,11(9), 512-517. https://doi.org/10.38124/ijisrt/26sep376

I. INTRODUCTION

Dyslexia is a specific learning difficulty associated with persistent difficulties in accurate and fluent word recognition, decoding and spelling. Early recognition is important because intervention can be introduced before difficulties become entrenched. Conventional assessment typically involves standardized educational and psychological procedures and specialist interpretation. Although such assessment remains essential, digital behavioural data create opportunities for scalable computational screening.

Rello et al. demonstrated that behavioural interaction data collected during an online gamified assessment could support machine-learning prediction of dyslexia risk. Their study involved more than 3,600 participants and subsequently evaluated the approach on a separate dataset. This work established an important basis for investigating machine learning from digital behavioural indicators.

Recent reviews show growing use of artificial intelligence for dyslexia detection but also identify challenges involving limited datasets, feature dimensionality, interpretability and generalization. These issues motivate approaches that combine feature selection with heterogeneous machine-learning models.

This study therefore investigates whether feature selection and ensemble learning can improve robust dyslexia prediction from behavioural data. The research questions are: (RQ1) How effectively does feature selection improve prediction? (RQ2) Which individual learner performs best? (RQ3) Does stacking improve dyslexia-class detection? and (RQ4) Which metrics

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep376

most appropriately describe performance under class imbalance?

II. RELATED WORK

➢ *Machine Learning for Dyslexia Detection*

Machine learning has been applied to dyslexia detection using behavioural, linguistic, eye-tracking and neurophysiological information. Behavioural approaches are particularly attractive for scalable screening because they can potentially be integrated into digital learning environments. Rello et al. showed that gamified behavioural assessment could support dyslexia-risk prediction, while later reviews have emphasized the need for stronger validation and transparent reporting.

➢ *Feature Selection*

High-dimensional predictors can contain redundant and weakly informative variables. Feature selection reduces dimensionality and may improve generalization. Recursive Feature Elimination repeatedly removes lower-importance predictors according to an underlying estimator. RF-RFE is appropriate for this study because Random Forest can estimate nonlinear feature importance without requiring strong distributional assumptions.

➢ *Ensemble Learning*

Ensemble learning combines multiple models to exploit complementary decision structures. Random Forest uses bagging and randomized feature selection; XGBoost uses sequential gradient boosting; Extra Trees introduces additional split randomization. Their methodological diversity provides a rationale for stacking. Wolpert's stacked generalization framework motivates learning a second-level classifier from the outputs of first-level learners.

III. MATERIALS AND METHODS

➢ *Dataset*

The uploaded Dyt-desktop.csv file contains 3,644 observations and 197 columns: 196 predictors and one binary target variable, Dyslexia. The dataset contains 3,252 observations labelled No and 392 labelled Yes, giving a dyslexia prevalence of 10.76%. Predictors include demographic variables and task-level behavioural measures such as clicks, hits, misses, scores, accuracy and miss-rate variables across multiple tasks.

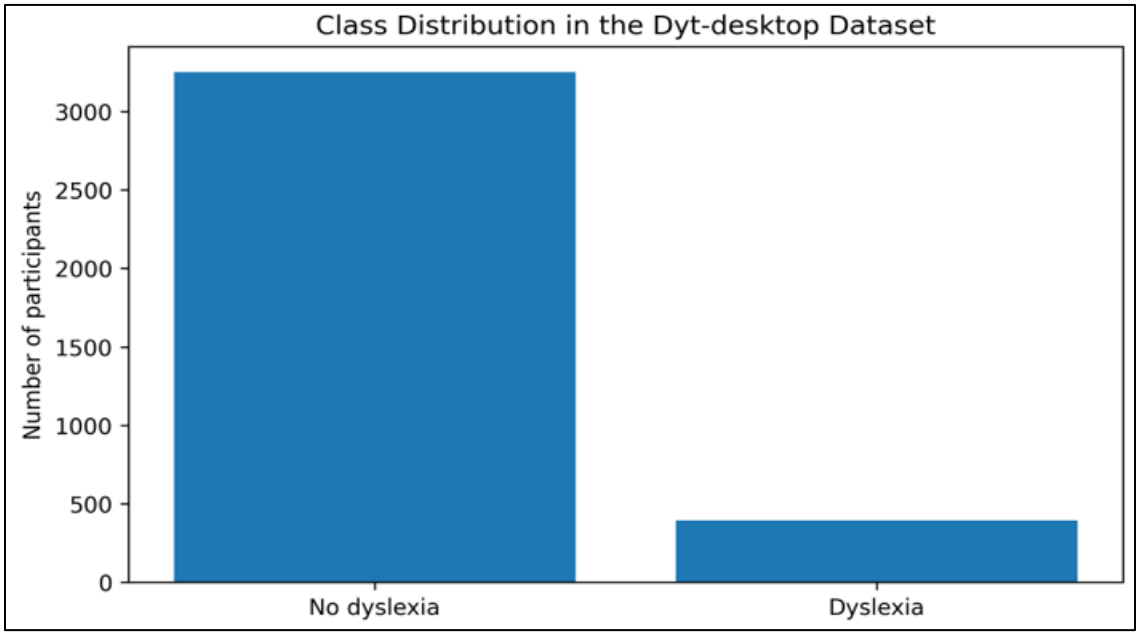

Fig 1. Class Distribution in the Uploaded Dyt-Desktop Dataset.

Table 1 Class Distribution in the Uploaded Dyt-Desktop Dataset.

Characteristic	Value
Observations	3,644
Predictors	196
Target	Dyslexia
No dyslexia	3,252 (89.24%)
Dyslexia	392 (10.76%)
Columns	197

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep376

➢ *Preprocessing*

Categorical predictors included Gender, Nativelang and Otherlang. Numerical predictors were processed as quantitative variables, while categorical variables were encoded for model training. The target variable was converted to a binary representation. An 80:20 stratified train-test split was used to preserve the class ratio.

➢ *RF-RFE Feature Selection*

Let F0={f1,f2,...,fp} represent the original predictor set. RF-RFE trains a Random Forest, ranks predictors by importance and recursively eliminates the least informative variables. The process yields a reduced feature subset F*. The selected variables are then supplied to the downstream learners. The objective is to remove noise and redundancy while retaining predictive information.

➢ *Base Learners*

Random Forest combines randomized decision trees through aggregation. XGBoost constructs an additive sequence of boosted trees. Extra Trees uses additional randomization in tree construction. These models were selected because their different learning mechanisms can generate complementary prediction errors.

➢ *Stacking Ensemble*

The base predictions are represented as M=[P_RF,P_XGB,P_ET]. A logistic meta-classifier estimates the final probability: $P(Y=1)=1/[1+\exp(-(\beta0+\beta1P\_RF+\beta2P\_XGB+\beta3P\_ET))]$. A default threshold of 0.5 is used for classification. This architecture follows the principle of stacked generalization.

IV. EVALUATION METRICS

Because dyslexia-positive observations represent only 10.76% of the dataset, accuracy alone is inadequate. Accuracy, precision, recall, specificity, F1-score, ROC-AUC and MCC were therefore reported. Recall is especially important because false negatives represent potentially missed cases requiring further assessment. MCC provides an additional balanced measure for imbalanced binary classification.

V. RESULTS

Table 2 Base Models

Model	Accuracy	Precision	Recall	Specificity	F1	ROC-AUC	MCC
RF	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

XGBoost was the strongest initial individual model, producing the highest accuracy and ROC-AUC and considerably higher recall than RF and Extra Trees.

Table 3 RF-RFE Models

Model	Accuracy	Precision	Recall	Specificity	F1	ROC-AUC	MCC
RF-RFE + RF	90.07%	73.44%	11.99%	99.48%	20.61%	0.8813	0.2705
RF-RFE + XGB	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-RFE-XGBoost produced the strongest feature-selected individual performance. Relative to the original XGBoost model, recall increased from 29.08% to 31.89%, F1 from 40.07% to 42.74%, and MCC from 0.3912 to 0.4122.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep376

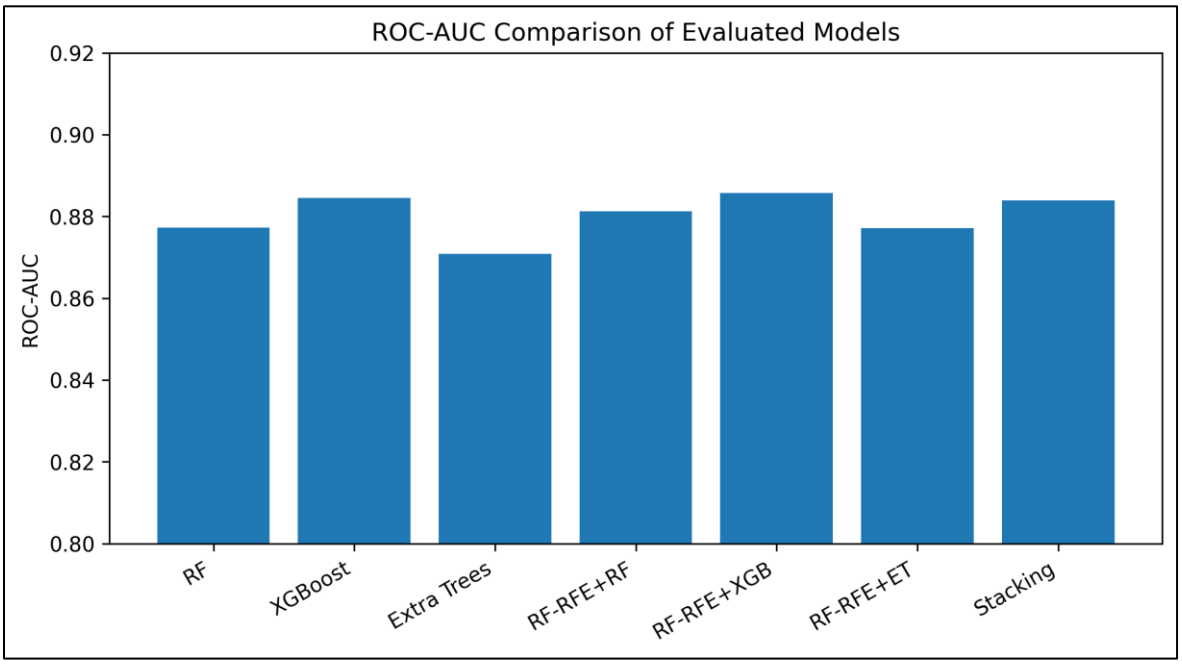

Fig 2. ROC-AUC comparison across evaluated models.

Table 4 Stacking Ensemble

Model	Accuracy	Precision	Recall	Specificity	F1	ROC-AUC	MCC
RF-XGB-ET Stacking	85.70%	40.61%	71.17%	87.45%	51.71%	0.8839	0.4644

The stacking ensemble produced the highest dyslexia-class recall, F1-score and MCC. Its recall of 71.17% substantially exceeded the best individual recall, but this gain was accompanied by lower specificity and accuracy.

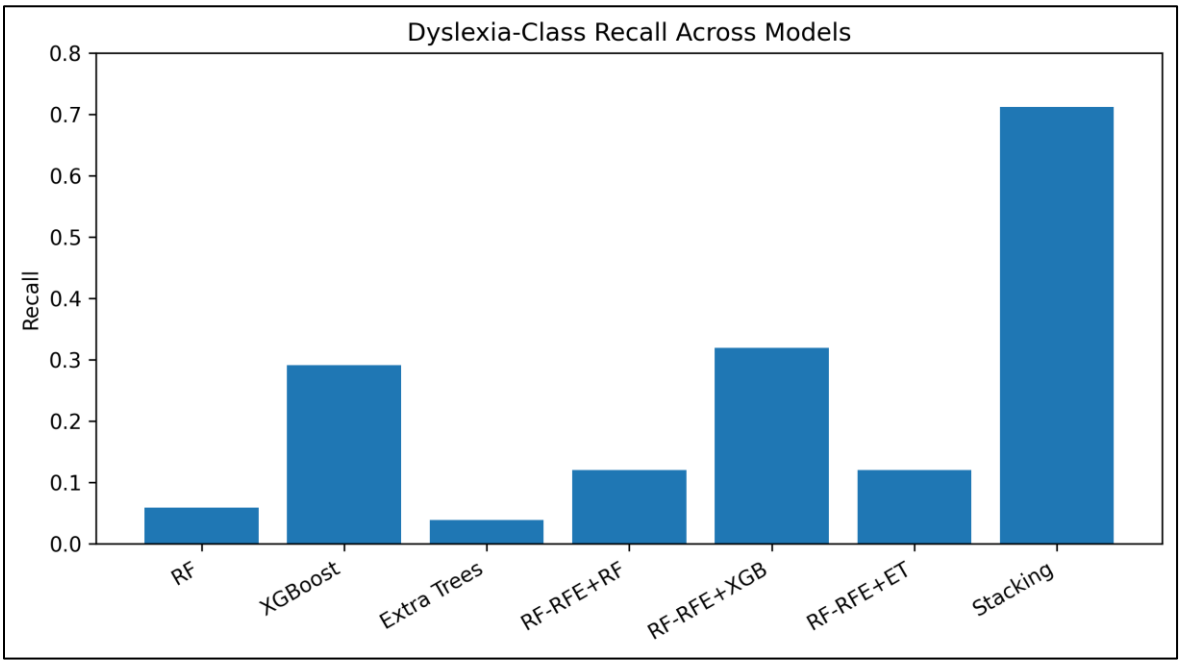

Fig 3. Dyslexia-Class Recall Across Evaluated Models.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep376

Table 5 Hyperparameter Optimization

Tuned model	Accuracy	Precision	Recall	Specificity	F1	ROC-AUC	MCC
Tuned RF	90.81%	65.71%	29.49%	98.16%	40.71%	0.8763	0.3997
Tuned ET	87.11%	40.70%	44.87%	92.17%	42.68%	0.8685	0.3549
Tuned XGBoost	91.22%	70.59%	30.77%	98.46%	42.86%	0.8921	0.4285

Randomized hyperparameter search used stratified two-fold cross-validation and ROC-AUC as the optimization criterion. The tuned XGBoost model was the strongest optimized individual classifier. The tuned Extra Trees model achieved the highest recall among the tuned individual learners.

VI. DISCUSSION

The results demonstrate that dyslexia prediction performance depends strongly on the evaluation objective. The class distribution makes accuracy potentially misleading: predicting the majority class alone would yield approximately 89.24% accuracy. The base RF and ET models achieved high specificity but very low recall, indicating that they frequently classified dyslexia-positive cases as non-dyslexic.
Feature selection improved the performance of individual models, especially XGBoost. The RF-RFE-XGBoost configuration achieved 90.81% accuracy, 31.89% recall, 42.74% F1-score, 0.8858 ROC-AUC and 0.4122 MCC. This supports the hypothesis that reducing irrelevant or redundant predictors can improve the predictive representation.

The stacking ensemble provided the most screening-oriented result. Its recall of 71.17% indicates that it identified a substantially greater proportion of dyslexia-positive participants. However, specificity decreased to 87.45%. This sensitivity-specificity trade-off is not necessarily undesirable in a screening context, where the cost of missing a potentially dyslexic child may be greater than the cost of referring some false-positive cases for additional assessment.

The findings are consistent with the general rationale of heterogeneous ensemble learning: models with different inductive biases can contribute complementary information. RF and ET provide randomized tree-based perspectives, while XGBoost provides sequential boosting. Their predictions can therefore contain information that a single classifier does not capture.

VII. IMPLICATIONS

For educational screening, a high-recall ensemble could serve as an initial risk-identification layer before formal professional assessment. Digital learning platforms could potentially incorporate behavioural indicators into a screening score. However, deployment should remain human-in-the-loop and should not present an ML prediction as a clinical diagnosis.

For researchers, the study demonstrates the importance of reporting sensitivity, specificity, F1, ROC-AUC and MCC rather than accuracy alone. For practitioners, threshold selection should be aligned with the cost of false negatives and false positives in the intended setting.

VIII. LIMITATIONS

The study is a secondary analysis of a single public dataset and is therefore constrained by the population, behavioural tasks and labels available in the source data. The substantial class imbalance may affect model calibration and threshold behaviour. External validation on independent datasets is required before generalization. The model should be regarded as a screening-support system rather than a diagnostic instrument.

IX. FUTURE RESEARCH

Future studies should perform external validation across different populations and educational environments; optimize decision thresholds according to screening objectives; investigate cost-sensitive learning; assess probability calibration; and incorporate explainable AI methods such as SHAP to identify the behavioural variables driving predictions. Fairness analysis across demographic subgroups should also be conducted.

X. CONCLUSION

This study evaluated feature selection and heterogeneous ensemble learning for robust dyslexia prediction using the Rello et al. public behavioural dataset. RF-RFE improved individual model performance, with RF-RFE-XGBoost achieving 90.81% accuracy and 0.8858 ROC-AUC. Hyperparameter optimization further strengthened XGBoost, which achieved 91.22% accuracy and 0.8921 ROC-AUC. Most importantly, the RF-XGB-ET stacking ensemble achieved 71.17% recall, 51.71% F1-score and 0.4644 MCC, demonstrating substantially improved identification of the dyslexia class. The results show that the strongest screening model is not necessarily the model with the highest accuracy. For early-risk screening, recall and balanced metrics should receive greater emphasis. The proposed feature-selection and ensemble framework therefore provides a promising research foundation for scalable dyslexia-risk prediction, subject to external validation, threshold optimization and professional oversight.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep376

REFERENCES

[1]. Alkhurayyif, Y., & Sait, A. R. W. (2024). A review of artificial intelligence-based dyslexia detection techniques. Diagnostics, 14(21), 2362. https://doi.org/10.3390/diagnostics14212362.

[2]. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13, 281–305.

[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.

[7]. 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.

[8]. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[9]. 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.

[10]. Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241–259.

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

[12]. Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. Advances in Neural Information Processing Systems, 25.