© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1722828

Comparative Analysis of Base Machine-Learning Models for Dyslexia Prediction Using Behavioral Learning Data

PATRICK NWOGU[1], PROF. CHIGOZIRIM AJAEGU[2], DR. FARUK UMAR AMBURSA[3], DR. FEMI ADELUYI[4]
[1, 2, 3, 4]*National Open University of Nigeria*

Abstract*—Dyslexia is a specific learning disorder that affects reading, decoding, spelling and fluent word recognition. Early identification can facilitate timely educational intervention. This study comparatively evaluates three machine-learning (ML) base classifiers—Random Forest (RF), XGBoost and Extra Trees (ET)—for dyslexia prediction using behavioural learning data. The supplied Dyt-desktop dataset contains 3,644 observations, 196 predictors and 392 dyslexia-positive cases. Stratified five-fold cross-validation was applied with class-weighted classifiers. Performance was assessed using accuracy, precision, recall, specificity, F1-score, ROC-AUC and Matthews Correlation Coefficient (MCC). XGBoost achieved the strongest overall base-model performance, with 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 and 31.89% recall. The previously developed probability-level ensemble achieved 71.17% recall and 0.4644 MCC, although its accuracy decreased to 85.70%. The findings demonstrate that accuracy alone is inadequate for evaluating dyslexia prediction under substantial class imbalance. XGBoost is the strongest individual base learner, while ensemble learning provides greater sensitivity when screening is prioritized.

Keywords*— dyslexia; learning disability; machine learning; Random Forest; XGBoost; Extra Trees; RF-RFE; ensemble learning; educational data mining.

I. INTRODUCTION

Dyslexia is a specific learning disorder characterized by persistent difficulties in accurate and fluent word recognition, decoding and spelling [1]. Early identification is important because timely educational support can reduce the long-term consequences of unidentified reading difficulties.

The increasing use of digital learning environments generates large amounts of learner-performance data that can be analysed using machine learning. Rello et al. [1] demonstrated that behavioural information collected through online assessment can be used to predict dyslexia risk. Tree-based ML methods are particularly appropriate for structured educational data because they can model nonlinear relationships and interactions among predictors.

Random Forest [2], XGBoost [3] and Extra Trees [4] use different tree-learning strategies and can therefore produce complementary predictions. The earlier work underlying this study combined these models with Random-Forest Recursive Feature Elimination (RF-RFE) and probability-level stacking. The present study focuses specifically on the comparative performance of the individual base models, providing the empirical basis for determining whether ensemble learning is justified.

The objectives are to: (i) compare RF, XGBoost and ET; (ii) evaluate the effect of RF-RFE; and (iii) determine the most appropriate model for dyslexia screening.

II. DATASET AND METHODOLOGY

The supplied Dyt-desktop.csv dataset contains 3,644 observations and 196 predictors. The variables include demographic information and performance measures such as clicks, hits, misses, score, accuracy and miss rate across 32 linguistic tasks.

Table 1. Dataset distribution

Class	Frequency	Percentage
Dyslexia	392	10.76%
Non-dyslexia	3,252	89.24%
Total	3,644	100%

The substantial class imbalance means that a majority-class classifier could obtain approximately 89.24% accuracy without identifying dyslexia cases. Therefore, multiple evaluation metrics are essential.

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1722828

The experimental pipeline consisted of data preprocessing, categorical encoding, numerical imputation, class weighting and stratified five-fold cross-validation. Out-of-fold predictions were used for unbiased performance estimation.

RF, XGBoost and ET were used as the base learners. RF and ET used 250 trees, while XGBoost used 250 estimators with a learning rate of 0.05 and maximum depth of five.

RF-RFE was subsequently applied to reduce the feature space and retain the most informative predictors.

Performance was measured using:

[Accuracy]

[Precision]

[Recall]

[Specificity]

[F1]

MCC was also reported because it provides a balanced measure for imbalanced binary classification.

III. RESULTS

3.1 Base-Model Comparison

Table 2. Performance of base ML 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

The results show that XGBoost is the strongest overall base learner. It obtained the highest accuracy, recall, F1-score, ROC-AUC and MCC.

XGBoost achieved 90.64% accuracy compared with 89.65% for RF and 89.54% for ET. More importantly, its recall of 29.08% was substantially higher than RF (5.87%) and ET (3.83%).

ET achieved the highest precision (78.95%) and specificity (99.88%), indicating that it produces relatively few false-positive classifications. However, its very low recall means that most dyslexia-positive observations remain unidentified.

These results demonstrate that similar accuracy values can conceal substantial differences in minority-class performance.

3.2 RF-RFE Results

Table 3. Effect of RF-RFE

Model	Accuracy	Recall	F1	ROC-AUC	MCC
RF	89.65%	5.87%	10.87%	0.8773	0.1896
RF-RFE + RF	90.07%	11.99%	20.61%	0.8813	0.2705
XGBoost	90.64%	29.08%	40.07%	0.8845	0.3912
RF-RFE + XGBoost	90.81%	31.89%	42.74%	0.8858	0.4122
ET	89.54%	3.83%	7.30%	0.8709	0.1593
RF-RFE + ET	89.96%	11.99%	20.43%	0.8771	0.2597

RF-RFE improves each base learner. The strongest improvement occurs for XGBoost, where accuracy increases from 90.64% to 90.81%, recall from 29.08% to 31.89%, and MCC from 0.3912 to 0.4122.

Thus:

[ ]

is the strongest individual configuration in the experiment.

3.3 Ensemble Comparison

The earlier research combined RF, XGBoost and ET probability outputs using a logistic-regression meta-classifier.

Table 4. Best base models versus ensemble

Model	Accuracy	Recall	F1	ROC-AUC	MCC
Random Forest	89.65%	5.87%	10.87%	0.8773	0.1896

XGBoost	90.64%	29.08%	40.07%	0.8845	0.3912
RF-RFE + XGBoost	90.81%	31.89%	42.74%	0.8858	0.4122
Ensemble	85.70%	71.17%	51.71%	0.8839	0.4644

The ensemble substantially increases recall:
[71.17%-29.08%=42.09]
percentage points above XGBoost.
MCC also increases from 0.3912 to 0.4644. However, accuracy decreases from 90.64% to 85.70%. The ensemble therefore represents a sensitivity-oriented model, rather than a model that maximizes accuracy.

IV. DISCUSSION AND CONCLUSION

The comparative analysis demonstrates that XGBoost is the strongest individual base model for the supplied dyslexia dataset. Its ROC-AUC of 0.8845 and MCC of 0.3912 indicate better overall discrimination and class-balanced performance than RF and ET.

The results also reveal the limitations of accuracy in imbalanced dyslexia datasets. Although RF and ET obtain approximately 90% accuracy, their recalls are only 5.87% and 3.83%, respectively. Consequently, these models could miss a large proportion of dyslexia-positive learners.

RF-RFE improves XGBoost performance, producing 90.81% accuracy, 31.89% recall, 42.74% F1-score, 0.8858 ROC-AUC and 0.4122 MCC. Feature selection therefore provides a modest but consistent improvement.

The probability-level ensemble demonstrates a different advantage. Its recall of 71.17% is substantially greater than that of the individual classifiers, although this comes with reduced accuracy and specificity. This finding supports the use of ensemble learning where the primary objective is early screening and minimizing false negatives.

Overall, the study establishes the following hierarchy: The models should be regarded as screening and educational decision-support tools rather than diagnostic instruments. Future research should include threshold optimization, precision-recall curves, SHAP-based explainability, confidence intervals, fairness analysis and independent external validation.

REFERENCES

[1] 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.
[2] Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.
[3] 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.
[4] Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63, 3–42.
[5] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
[6] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.