© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1722829

Exploratory Analysis of Ensemble Machine Learning for Dyslexia Detection Using RF-RFE and Stacked Learning

PATRICK UFOMBA NWOGU[1], PROF. CHIGOZIRIM AJAEGU[2], DR. FARUK UMAR AMBURSA[3], DR. FEMI ADELUYI[4]

[1, 2, 3, 4]*Doctorate, Researcher in Management Information Systems, National Open University of Nigeria*

Abstract*—This study develops and empirically evaluates an ensemble machine-learning framework for dyslexia detection using Random-Forest Recursive Feature Elimination (RF-RFE), Random Forest (RF), XGBoost, Extra Trees (ET), and probability-level logistic-regression stacking. The analysis uses the supplied Dyt-desktop.csv dataset, which contains 3,644 observations, 196 predictors and one binary dyslexia outcome. The target distribution is imbalanced: 392 dyslexia cases (10.76%) and 3,252 non-dyslexia cases (89.24%). Five-fold stratified out-of-fold evaluation was used. Class-weighted RF, XGBoost and ET were compared with RF-RFE variants, while the proposed ensemble combined out-of-fold probability predictions from RF, XGBoost and ET through logistic regression. At the default 0.50 threshold, XGBoost produced the strongest individual accuracy (90.64%), while RF-RFE + XGBoost achieved 90.81% accuracy, 64.77% precision, 31.89% recall, 97.91% specificity, F1=0.4274, ROC-AUC=0.8858 and MCC=0.4122. The proposed stacking ensemble produced ROC-AUC=0.8839, recall=71.17%, F1=0.5171 and MCC=0.4644, but lower accuracy (85.70%) because its class-weighted meta-classifier generated substantially more positive predictions. These findings demonstrate that model ranking changes according to the evaluation objective: RF-RFE + XGBoost is strongest on accuracy and ROC-AUC, whereas the proposed ensemble provides the strongest minority-class recall, F1 and MCC at the tested threshold. The results support threshold optimization and precision-recall analysis as essential components of dyslexia screening.

Keywords*— dyslexia; machine learning; ensemble learning; RF-RFE; Random Forest; XGBoost; Extra Trees; stacking; feature selection; learning analytics.

I. INTRODUCTION

Dyslexia is a specific learning disorder associated with persistent difficulties in accurate and fluent word recognition, spelling and decoding. Machine learning has increasingly been investigated as a means of supporting early screening by learning predictive patterns from behavioural, linguistic and interaction data. Rello et al. reported a large online gamified dyslexia dataset and demonstrated that Random Forest could distinguish participants with and without professionally diagnosed dyslexia [1].

The present work extends the earlier ensemble-ML research developed in this project by using the supplied dataset to perform a reproducible comparative experiment involving RF, XGBoost, ET, RF-RFE variants and a logistic-regression stacking ensemble. The central premise is that feature selection can reduce redundant predictors and that heterogeneous tree learners may provide complementary probability estimates.

II. DATASET AND DATA QUALITY

The supplied Dyt-desktop.csv file contains 3,644 rows and 197 columns: 196 predictors plus the Dyslexia target. The predictors include Gender, Nativelang, Otherlang, Age and repeated performance variables for 32 tasks. Each task contains Clicks, Hits, Misses, Score, Accuracy and Missrate. The observed class counts are 3,252 non-dyslexia and 392 dyslexia cases. Thus, dyslexia prevalence is 392/3,644 = 10.76%.

During inspection, the file contains several numeric-looking entries represented as text and some unusual decimal encodings in performance fields. Numeric-looking columns were coerced to numeric where at least 95% of values were parseable; remaining categorical variables were one-hot encoded. Missing numeric values were median-imputed and categorical values were most-frequent imputed. The models used class weighting to reduce minority-class neglect.

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1722829

III. METHODOLOGY

A stratified five-fold cross-validation procedure was used with random_state=42. All preprocessing was fitted on the training portion of each fold. Baseline RF, XGBoost and ET models used 250 trees. RF used class_weight='balanced'. XGBoost used 250 estimators, max_depth=5, learning_rate=0.05, subsample=0.90 and colsample_bytree=0.90. ET used 250 trees with balanced class weighting.

For RF-RFE models, feature selection was performed independently inside each training fold using a Random Forest estimator. The retained subset was limited to 60 transformed predictors or approximately one-half of the transformed feature space, whichever was smaller, with a minimum of 10. This fold-specific procedure prevents test-fold information from entering feature selection.

The proposed ensemble used out-of-fold probabilities from RF, XGBoost and ET as meta-features. A class-weighted logistic-regression classifier was then fitted to these probability vectors using a second stratified five-fold procedure. The default decision threshold was 0.50. No threshold was tuned after seeing the final results.

IV. MATHEMATICAL FORMULATION

For learner i, let P_RF, P_XGB and P_ET denote predicted probabilities for dyslexia. The meta-feature matrix is M=[P_RF, P_XGB, P_ET]. The logistic meta-classifier estimates: $P(Y=1)=1/(1+\exp[-(\beta0+\beta1P\_RF+\beta2P\_XGB+\beta3P\_ET)])$. The final class is 1 when $P(Y=1)\geq0.50$ and 0 otherwise.

Evaluation used accuracy, precision, recall/sensitivity, specificity, F1, ROC-AUC and Matthews correlation coefficient (MCC). Because the positive class is only 10.76% of the sample, MCC, recall, F1 and precision-recall behaviour are particularly important.

V. EMPIRICAL RESULTS

Model	Accuracy	Precision	Recall	Specificity	F1	ROC-AUC
0.8965	0.7419	0.0587	0.9975	0.1087	0.8773	0.1896
0.9064	0.6441	0.2908	0.9806	0.4007	0.8845	0.3912
0.8954	0.7895	0.0383	0.9988	0.0730	0.8709	0.1593
0.9007	0.7344	0.1199	0.9948	0.2061	0.8813	0.2705
0.9081	0.6477	0.3189	0.9791	0.4274	0.8858	0.4122
0.8996	0.6912	0.1199	0.9935	0.2043	0.8771	0.2597
0.8570	0.4061	0.7117	0.8745	0.5171	0.8839	0.4644

Table 1. Five-fold out-of-fold empirical results obtained from the supplied Dyt-desktop.csv dataset.

The Random Forest baseline produced 89.65% accuracy and ROC-AUC=0.8773 but only 5.87% recall at the 0.50 threshold. XGBoost improved recall to 29.08% and ROC-AUC to 0.8845. Extra Trees produced high specificity (99.88%) but very low recall (3.83%). RF-RFE + XGBoost produced the best accuracy (90.81%) and ROC-AUC (0.8858) among the evaluated configurations, while improving recall to 31.89%.

The proposed stacking ensemble produced the highest recall (71.17%), F1 (0.5171) and MCC (0.4644) of all tested models. Its accuracy was 85.70% and specificity was 87.45%. This trade-off reflects the use of class-weighted logistic regression at the meta-level and illustrates why threshold-dependent metrics must be interpreted jointly.

VI. RESULTS INTERPRETATION

The results provide evidence that no single metric adequately summarizes dyslexia screening performance. If the operational objective is maximum overall accuracy at a fixed 0.50 threshold, RF-RFE + XGBoost is the leading configuration. If the objective is to identify as many dyslexia cases as possible, the proposed ensemble is substantially stronger in this experiment, with recall of 71.17%. Its MCC of 0.4644 is also the highest observed, indicating the most balanced association between predictions and actual classes among the tested configurations.

The ROC-AUC values are relatively close (0.8709–0.8858), indicating that ranking performance is more similar across models than the thresholded

classification metrics suggest. This is important: the large difference in recall between models does not necessarily mean an equally large difference in underlying ranking ability. Threshold selection therefore deserves dedicated analysis.

VII. DISCUSSION

The findings support the original motivation for heterogeneous ensemble learning. XGBoost benefited from RF-RFE, suggesting that reducing redundant predictors can improve discrimination. Random Forest and Extra Trees showed high specificity but weak recall at the default threshold. In contrast, the stacking model converted complementary probability information into a more sensitive classifier.

The class imbalance is central to interpretation. With 89.24% non-dyslexia prevalence, an all-negative classifier would obtain about 89.24% accuracy while identifying no dyslexia cases. The empirical RF and ET accuracies close to this value therefore cannot be interpreted as strong screening performance without considering recall and MCC.

VIII. RECOMMENDED ADDITIONAL ANALYSES

Optimize the decision threshold inside the training/validation process and report sensitivity-specificity trade-offs.
Report precision-recall AUC because dyslexia is the minority class.
Produce bootstrap confidence intervals for all metrics.
Compare models with paired out-of-fold predictions using appropriate statistical tests.
Perform RF-RFE stability analysis to determine which predictors are consistently selected.
Generate SHAP explanations for the final ensemble and the strongest individual model.
Conduct subgroup analysis by age, gender and language variables.
Perform external validation on an independent dyslexia cohort before deployment.

IX. LIMITATIONS

The present analysis is based on one supplied dataset and one cross-validation design. The hyperparameters were deliberately kept moderate and were not exhaustively optimized. RF-RFE was operationalized with a fixed retention rule for computationally practical comparison; the optimal feature count should be selected through nested validation in a final confirmatory study. The proposed stacking result is empirical for this dataset and protocol, but it should not be generalized to other populations without external validation.

The dataset should also be treated as a screening resource rather than a diagnostic substitute. Educational or clinical deployment requires professional oversight, privacy safeguards, fairness assessment and independent validation.

X. CONCLUSION

Using the supplied Dyt-desktop.csv dataset, this study empirically demonstrates the value of combining feature selection and heterogeneous ensemble learning for dyslexia prediction. RF-RFE + XGBoost achieved the highest accuracy (90.81%) and ROC-AUC (0.8858), whereas the proposed RF/XGBoost/ET stacking ensemble achieved the highest recall (71.17%), F1 (0.5171) and MCC (0.4644). The difference illustrates the central screening trade-off between specificity and sensitivity. The next methodological priority is therefore threshold optimization, precision-recall analysis, confidence intervals, explainability 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. https://doi.org/10.1371/journal.pone.0241687.
[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.
[4] Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine Learning, 63, 3–42.
[5] Alkhurayyif, Y., & Sait, A. R. W. (2024). A review of artificial intelligence-based dyslexia

detection techniques. Diagnostics, 14(21), 2362.

[6] 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.