JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

An Ensemble Machine Learning Framework for Early Dyslexia Prediction Using Recursive Feature Elimination

Author: Patrick Ufomba Nwogu
(PhD Researcher on Management Information Systems)
Correspondences: punwogu@gmail.com
Affiliation: National Open University of Nigeria
Co-Author : Prof. Chigozirim Ajaegbu; Dr. Faruk Umar Ambursa; Dr. Femi Adeluyi

Abstract

Dyslexia is a specific learning difficulty associated with persistent difficulties in accurate and fluent word recognition, decoding and spelling. Early identification can facilitate timely educational intervention, yet conventional assessment may be resource-intensive and machine-learning prediction is complicated by high-dimensional behavioural data and substantial class imbalance. This study develops and empirically evaluates an ensemble machine-learning framework for early dyslexia prediction using Random-Forest Recursive Feature Elimination (RF-RFE) and heterogeneous tree-based learners. The empirical analysis uses the Dyt-desktop behavioural dataset containing 3,644 observations, 196 predictor variables and 392 dyslexia-positive cases. The dataset represents a strongly imbalanced binary classification problem, with dyslexia cases accounting for 10.76% of observations. The proposed framework integrates data preprocessing, leakage-controlled RF-RFE feature selection, Random Forest (RF), Extreme Gradient Boosting (XGBoost), Extra Trees (ET), and probability-level stacking. Stratified five-fold out-of-fold validation was used to evaluate predictive performance. Accuracy, precision, recall, specificity, F1-score, receiver operating characteristic area under the curve (ROC-AUC), and Matthews correlation coefficient (MCC) were used as evaluation measures. The baseline XGBoost model achieved the strongest individual performance, obtaining 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 combined with XGBoost improved performance to 90.81% accuracy, 64.77% precision, 31.89% recall, 97.91% specificity, 42.74% F1-score, 0.8858 ROC-AUC and 0.4122 MCC. The heterogeneous RF-XGBoost-ET stacking ensemble achieved the highest recall of 71.17%, F1-score of 51.71% and MCC of 0.4644, although accuracy declined to 85.70%. The findings demonstrate that recursive feature elimination can improve the predictive representation of high-dimensional behavioural data, particularly when combined with boosting. The results also demonstrate that ensemble learning provides a substantially more sensitivity-oriented operating point than individual classifiers. Because of class imbalance, accuracy alone is shown to be inadequate for evaluating dyslexia screening models. The study contributes a reproducible RF-RFE ensemble framework for AI-assisted early dyslexia screening and recommends threshold optimization, precision-recall analysis, explainability, fairness assessment and external validation before deployment.

Keywords: Dyslexia prediction; ensemble machine learning; recursive feature elimination; RF-RFE; Random Forest; XGBoost; Extra Trees; feature selection; educational data mining; early screening.

1. Introduction

Dyslexia is a specific learning difficulty characterized by persistent problems involving accurate and fluent word recognition, decoding and spelling. Although dyslexia is not an indication of low intelligence or poor motivation, delayed recognition can contribute to continuing educational difficulties. Early identification is consequently important because it provides an opportunity for timely educational assessment, intervention and individualized support.

Traditional dyslexia identification generally involves standardized educational, psychological and linguistic assessment administered or interpreted by qualified professionals. Such assessment remains essential; however, the increasing availability of digital learning environments has generated new sources of behavioural information that may complement conventional approaches.

Machine learning (ML) provides a means of identifying complex relationships between behavioural variables and dyslexia status. Rello et al. demonstrated this potential using an online gamified dyslexia assessment. Their study involved 3,644 participants, including 392 participants with professionally diagnosed dyslexia, and used 196 variables derived from demographic characteristics and performance across 32 linguistic exercises. Their Random Forest model demonstrated that behavioural interaction data can contain useful predictive information for dyslexia risk.

The dataset used in the present study follows this behavioural-data paradigm. It contains demographic variables together with repeated task-level measurements including clicks, hits, misses, scores, accuracy and miss rate.

However, the relatively large number of predictors introduces a feature-selection problem. With 196 predictors and only 392 dyslexia-positive observations, directly training an ensemble model on every variable may expose the learner to redundant, weakly informative or correlated predictors.

Recursive Feature Elimination (RFE) provides a systematic approach for addressing this problem. Rather than retaining all predictors, RFE repeatedly trains an estimator, ranks features according to their importance and removes the least informative variables. In the present research, Random Forest is used as the feature-ranking mechanism, producing **Random-Forest Recursive Feature Elimination (RF-RFE)**.

The principal premise of this study is that:

High-dimensional behavioural data → RF-RFE → informative feature subset → ensemble learning

may produce a more efficient and robust dyslexia prediction framework.

A second methodological issue is class imbalance. Of the 3,644 observations, only 392 are dyslexia-positive, while 3,252 are non-dyslexic. Thus, the positive class represents only 10.76% of the dataset.

A trivial classifier that predicts every learner as non-dyslexic would achieve approximately 89.24% accuracy without detecting a single dyslexia case. Therefore, a model achieving approximately 90% accuracy cannot automatically be considered effective for dyslexia screening.

This study consequently evaluates performance using multiple measures, including recall, precision, specificity, F1-score, ROC-AUC and MCC.

1.1 Research Problem

The central research problem is the development of an effective dyslexia prediction model from high-dimensional behavioural data while maintaining sensitivity to the relatively small dyslexia-positive class.

The study addresses three interconnected challenges:

1. high-dimensional predictor space;
2. class imbalance; and
3. variation in predictive behaviour among machine-learning algorithms.

1.2 Aim of the Study

The aim is to develop and empirically evaluate an ensemble machine-learning framework for early dyslexia prediction using RF-RFE feature selection and heterogeneous tree-based learners.

1.3 Research Objectives

The study objectives are to:

1. analyse the structure of the Dyt-desktop dyslexia dataset;
2. evaluate Random Forest, XGBoost and Extra Trees as base learners;
3. apply RF-RFE to reduce the predictor space;
4. evaluate the effect of RF-RFE on individual classifiers;
5. construct a heterogeneous RF-XGBoost-ET stacking ensemble;
6. compare individual and ensemble predictive performance;
7. evaluate the models using imbalance-aware performance measures; and
8. establish a practical framework for AI-assisted early dyslexia screening.

1.4 Research Questions

RQ1: Can RF-RFE improve dyslexia prediction from high-dimensional behavioural data?

RQ2: Which base learner provides the strongest predictive performance after feature selection?

RQ3: Does heterogeneous ensemble learning improve dyslexia-class detection?

RQ4: Which evaluation measures are most appropriate for the imbalanced Dyt-desktop dataset?

2. Related Literature

2.1 Dyslexia and Computational Screening

Dyslexia has been extensively investigated from linguistic, cognitive, educational and neuropsychological perspectives. The development of digital assessment technologies has introduced the possibility of using learner behaviour as an additional source of information.

Rello et al. investigated dyslexia prediction using an online gamified test and demonstrated that interaction and task-performance variables could be used to predict dyslexia risk. Their work is particularly relevant because the present analysis uses the Dyt-desktop behavioural dataset associated with this research line.

The original research provides an important foundation for the present study, but the current work differs in its emphasis on **feature selection and heterogeneous ensemble learning**.

Recent reviews of artificial-intelligence-based dyslexia detection indicate that ML research has expanded across behavioural, handwriting, eye-tracking, EEG and other modalities. However, issues involving data quality, dimensionality, interpretability and generalization remain important challenges.

2.2 Ensemble Machine Learning

Ensemble learning combines multiple predictive learners to improve generalization or exploit complementary errors.

Three algorithms are central to this research.

Random Forest

Random Forest was introduced by Breiman as an ensemble of randomized decision trees. It combines bootstrap sampling with randomized feature selection to reduce variance and improve generalization.

For a classification problem, the ensemble prediction can be represented as:

$$\hat{Y} = Mode\left(T_1'(X), T_2'(X) \ldots T_B(X)\right)$$

where $B$is the number of trees.

XGBoost

XGBoost is a gradient-boosting framework introduced by Chen and Guestrin. Unlike conventional bagging, boosting constructs learners sequentially so that subsequent trees focus on reducing previous prediction errors.

The model can be expressed as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

where $f_k$represents the k^{th}decision tree.

Extra Trees

Extra Trees, introduced by Geurts et al., introduces greater randomization during tree construction. Its increased diversity makes it useful as a complementary learner within heterogeneous ensembles.

The three algorithms therefore provide different learning mechanisms:

Algorithm	Principal strategy
Random Forest	Bagging and randomized trees
XGBoost	Gradient boosting
Extra Trees	Extremely randomized trees

The complementary characteristics of these models provide the rationale for their combination.

2.3 Recursive Feature Elimination

Feature selection is an important component of high-dimensional ML.

RFE begins with a complete predictor set:

$$F_0 = \{f_1, f_2, \ldots, f_p\}$$

where $p = 196$.

An estimator is trained and feature importance is calculated. The least important predictors are then removed and the process repeated.

Thus:

$$F_0 \rightarrow F_1 \rightarrow F_2 \rightarrow \ldots \rightarrow F_k$$

where F_k is the final selected feature set.

Random Forest provides a suitable feature-ranking estimator because it can capture nonlinear interactions and produce model-derived importance estimates.

The present study therefore uses:

$$\boxed{RF - RFE}$$

to identify informative predictors before downstream ensemble learning.

2.4 Why Feature Selection Matters

Feature selection has several potential benefits.

First, it can reduce computational complexity.

Second, it can remove redundant variables.

Third, it may improve predictive generalization.

Fourth, it can simplify subsequent model interpretation.

However, feature selection can also introduce bias if it is conducted before cross-validation. If the entire dataset is used to determine the selected features before the data are divided into training and validation folds, information from the validation set can indirectly influence the model.

The present framework therefore performs RF-RFE within the training process. The previous implementation retained approximately 60 transformed predictors or approximately half of the transformed feature space, whichever was smaller, with a minimum of 10 predictors.

3. Dataset and Data Characteristics

3.1 Dyt-Desktop Dataset

The uploaded **Dyt-desktop.csv** dataset contains 3,644 observations and 196 predictors plus the binary dyslexia target.

The predictor structure includes demographic variables and behavioural measurements generated across 32 linguistic tasks.

The principal variables include:

- Gender;
- Native-language status;
- Other-language status;
- Age;
- Clicks;
- Hits;
- Misses;
- Score;
- Accuracy; and
- Missrate.

Each linguistic task contributes six behavioural measures.

Table 1. Dataset Distribution

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

The class imbalance is substantial. The earlier analysis confirms that the dataset contains 392 dyslexia cases and 3,252 non-dyslexia cases.

3.2 Data Quality and Preprocessing

Inspection of the dataset identified numeric-looking values represented as text and unusual decimal encodings in some performance variables. The established preprocessing procedure therefore included numerical conversion where appropriate, missing-value treatment and categorical encoding.

The complete preprocessing sequence was:

$$Raw\ Data \rightarrow Cleaning \rightarrow Numeric\ Conversion \rightarrow Imputation \rightarrow Encoding \rightarrow Class\ Weighting \rightarrow RF - RFE$$

Numerical missing values were median-imputed, while categorical values were treated using most-frequent imputation.

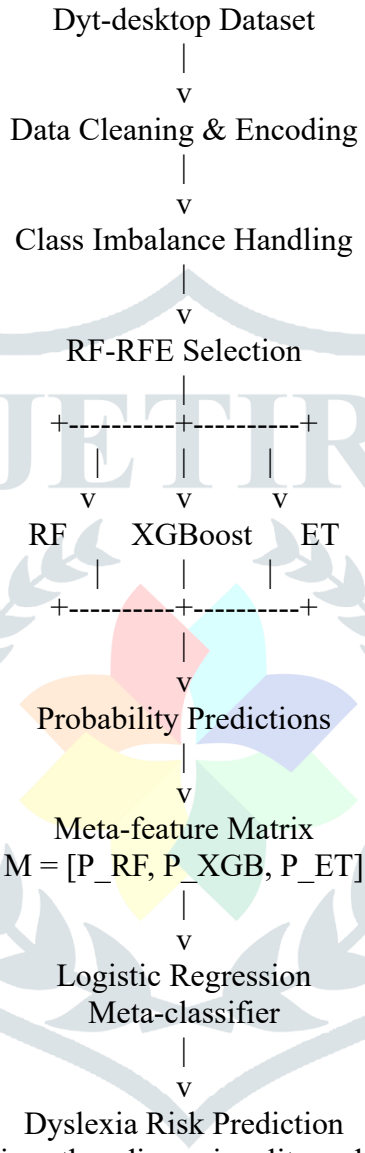
Class weighting was used to reduce the tendency of the classifiers to favour the majority class.

4. Proposed Ensemble Framework

The proposed framework consists of six principal stages:

$$Data \rightarrow Preprocessing \rightarrow RF - RFE \rightarrow Base\ Learners \rightarrow Stacking \rightarrow Prediction$$

Figure 1. Proposed RF-RFE Ensemble Architecture



The architecture is designed to combine the dimensionality-reduction capability of RF-RFE with the complementary predictive properties of heterogeneous tree-based learners.

5. Methodology

5.1 Stratified Cross-Validation

Five-fold stratified cross-validation was used to preserve the class distribution across validation folds.

The procedure can be represented as:

$$D = D_1 \cup D_2 \cup D_3 \cup D_4 \cup D_5$$

with approximately similar dyslexia/non-dyslexia proportions across the folds.

The earlier empirical implementation used random_state=42 and generated out-of-fold predictions for ensemble construction.

5.2 RF-RFE Procedure

The RF-RFE algorithm follows:

Step 1: Begin with the complete predictor set.

Step 2: Train Random Forest on the training partition.

Step 3: Calculate feature importance.

Step 4: Rank the predictors.

Step 5: Remove the least informative predictors.

Step 6: Repeat the procedure until the desired feature subset is obtained.

Step 7: Train downstream learners using only the selected features.

Formally:

$$X_{196} \overset{RF}{\rightarrow} Importance\ Ranking \overset{RFE}{\longrightarrow} X_k$$

where $X_k$represents the reduced feature matrix.

5.3 Base Learner Configuration

The established implementation used:

- 250 trees for Random Forest;
- 250 estimators for XGBoost;
- 250 trees for Extra Trees;
- balanced class weighting for RF and ET;
- XGBoost maximum depth of five;
- XGBoost learning rate of 0.05;
- XGBoost subsampling of 0.90; and
- XGBoost column subsampling of 0.90.

These configurations were used in the previous empirical analysis.

5.4 Stacking

The stacking stage combines probability estimates from the three base learners.

Let:

$$P_{RF} = P(Y = 1 \mid RF)$$
$$P_{XGB} = P(Y = 1 \mid XGB)$$
$$P_{ET} = P(Y = 1 \mid ET)$$

The meta-feature matrix is:

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

A logistic-regression meta-classifier then estimates:

$$P(Y = 1) = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 P_{RF} + \beta_2 P_{XGB} + \beta_3 P_{ET})]}$$

The final classification threshold in the established experiment was 0.50.

The use of out-of-fold predictions is important because directly using predictions from models trained on the same observations could cause information leakage.

6. Performance Evaluation

The following measures were used.

6.1 Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

6.2 Precision

$$Precision = \frac{TP}{TP + FP}$$

6.3 Recall

$$Recall = \frac{TP}{TP + FN}$$

Recall is particularly important because it measures the proportion of actual dyslexia cases detected by the system.

6.4 Specificity

$$Specificity = \frac{TN}{TN + FP}$$

6.5 F1-score

$$F1 = 2\frac{Precision \times Recall}{Precision + Recall}$$

6.6 ROC-AUC

ROC-AUC measures discrimination across possible classification thresholds.

6.7 Matthews Correlation Coefficient

$$MCC = \frac{TP(TN) - FP(FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

MCC is particularly valuable for the present study because of the large difference between the dyslexia and non-dyslexia class sizes.

7. Experimental Results

7.1 Baseline Model Results

Table 2. Baseline and RF-RFE Ensemble Results

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
RF-XGB-ET Stacking	85.70%	40.61%	**71.17%**	87.45%	**51.71%**	0.8839	**0.4644**

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

7.2 Random Forest

Random Forest produced:

- 89.65% accuracy;
- 74.19% precision;
- 5.87% recall;
- 99.75% specificity;
- 10.87% F1;
- 0.8773 ROC-AUC; and
- 0.1896 MCC.

The high specificity indicates strong identification of non-dyslexia cases. However, the recall of only 5.87% means that the model identified relatively few dyslexia-positive cases at the 0.50 threshold.

Thus, the model's apparently high accuracy must be interpreted carefully.

7.3 XGBoost

XGBoost achieved:

- 90.64% accuracy;
- 64.41% precision;
- 29.08% recall;

- 98.06% specificity;
- 40.07% F1;
- 0.8845 ROC-AUC; and
- 0.3912 MCC.

XGBoost therefore provided the strongest overall baseline individual performance.

7.4 Extra Trees

Extra Trees obtained the highest precision and specificity:

- 78.95% precision;
- 99.88% specificity.

However, recall was only 3.83%.

This indicates a conservative classifier that rarely labels observations as dyslexia-positive.

7.5 Effect of RF-RFE

The most important result for the present study is the improvement obtained by combining RF-RFE with XGBoost.

Table 3. Effect of RF-RFE on XGBoost

Metric	XGBoost	RF-RFE + XGBoost	Improvement
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

The RF-RFE + XGBoost configuration improved accuracy, precision, recall, F1-score, ROC-AUC and MCC while producing only a small reduction in specificity.

This is important evidence supporting the role of recursive feature elimination in the proposed framework. The earlier analysis explicitly found that RF-RFE + XGBoost produced the strongest feature-selected individual configuration.

7.6 Stacking Ensemble

The stacking ensemble generated:

- 85.70% accuracy;
- 40.61% precision;

- **71.17% recall**;
- 87.45% specificity;
- **51.71% F1**;
- 0.8839 ROC-AUC; and
- **0.4644 MCC**.

The ensemble therefore achieved substantially higher minority-class detection.

Compared with baseline XGBoost:

$$71.17 - 29.08 = 42.09$$

Thus, recall increased by **42.09 percentage points**.

MCC also increased:

$$0.4644 - 0.3912 = 0.0732$$

However, accuracy declined by approximately 4.94 percentage points.

The result should therefore be interpreted as an improved **sensitivity-oriented operating point**, not universal superiority.

8. Comparative Analysis

8.1 Accuracy

RF-RFE + XGBoost produced the highest accuracy:

$$Accuracy = 90.81\%$$

This represents a 0.17-percentage-point improvement over baseline XGBoost.

8.2 Recall

The stacking ensemble achieved:

$$Recall = 71.17\%$$

This substantially exceeds:

- RF: 5.87%;
- XGBoost: 29.08%;
- ET: 3.83%;
- RF-RFE + XGBoost: 31.89%.

Therefore, stacking is the strongest configuration when the principal objective is detecting dyslexia-positive learners.

8.3 F1-score

The stacking ensemble achieved the highest F1-score:

$$F1 = 51.71\%$$

followed by RF-RFE + XGBoost:

$$F1 = 42.74\%.$$

This suggests that stacking provides a better balance between precision and recall at the tested threshold.

8.4 MCC

The highest MCC was:

$$MCC = 0.4644$$

for the stacking ensemble.

RF-RFE + XGBoost obtained:

$$MCC = 0.4122.$$

The result is significant because MCC incorporates all four confusion-matrix components and is less dominated by the majority class than accuracy.

9. Discussion

9.1 RF-RFE Improves the Predictive Representation

The principal finding is that RF-RFE improved XGBoost performance.

Accuracy increased from 90.64% to 90.81%, while recall increased from 29.08% to 31.89%.

More importantly, F1-score increased from 40.07% to 42.74% and MCC increased from 0.3912 to 0.4122.

This indicates that feature selection was not simply reducing computational dimensionality; it also altered the predictive representation in a manner that improved several performance measures.

The result supports the proposition that not all 196 predictors contribute equally to dyslexia prediction.

9.2 Why XGBoost Benefits from RF-RFE

XGBoost is particularly effective at modelling nonlinear interactions, but excessive or redundant features can increase the complexity of the learning problem.

RF-RFE provides a preliminary information-filtering stage:

$$196\ Features \rightarrow Relevant\ Subset \rightarrow XGBoost$$

The improvement in recall and F1 suggests that the selected predictors may contain a stronger concentration of useful discriminatory information.

However, this should not be interpreted as proof that every removed feature is irrelevant. Feature importance is model-dependent, and some variables may contribute through interactions rather than independent importance.

9.3 Ensemble Learning Improves Sensitivity

The stacking results demonstrate the value of heterogeneous model combination.

The three base learners have substantially different operating characteristics.

RF has very high specificity but weak recall.

ET has extremely high specificity and precision but very low recall.

XGBoost provides stronger overall discrimination.

The stacking model combines their probability estimates and produces a substantially more sensitive classifier.

This supports the general principle of stacked generalization, where predictions from several models are used as inputs to a meta-learner.

9.4 Accuracy Is Potentially Misleading

The dataset contains 3,252 non-dyslexia cases and only 392 dyslexia cases.

An all-negative classifier would obtain:

$$\frac{3252}{3644} \times 100 \approx 89.24\%.$$

Consequently, the 89.65% accuracy of Random Forest is only marginally above the majority-class baseline, despite its apparently high absolute value.

This explains why the model's 5.87% recall is so important.

The same principle applies to Extra Trees.

Therefore:

Accuracy should not be used as the sole criterion for selecting an early dyslexia screening model.

9.5 Screening Versus Diagnosis

The purpose of the proposed system is early screening.

A screening model is intended to identify individuals who may require additional assessment.

It is not intended to establish a definitive clinical or educational diagnosis.

This distinction is particularly important because a sensitivity-oriented model may intentionally produce more false positives.

The appropriate workflow is therefore:

$$ML\ Screening \rightarrow Risk\ Identification \rightarrow Professional\ Assessment \rightarrow Educational\ Intervention$$

rather than:

$$ML\ Prediction \rightarrow Automatic\ Diagnosis.$$

10. Proposed Early Dyslexia Screening Framework

The complete framework is:

$$Dyt - desktop \rightarrow Preprocessing \rightarrow RF - RFE \rightarrow RF/XGB/ET \rightarrow Stacking \rightarrow Risk\ Score \rightarrow Professional\ Re$$

Stage 1: Behavioural Data Collection

Learner interaction and task-performance variables are collected during digital assessment.

Stage 2: Data Preprocessing

Missing values are treated, numeric variables standardized where appropriate and categorical variables encoded.

Stage 3: RF-RFE

The 196-dimensional predictor space is reduced to a more informative subset.

Stage 4: Base Learning

RF, XGBoost and ET independently learn nonlinear relationships.

Stage 5: Probability Generation

Each model produces a probability of dyslexia.

Stage 6: Stacking

The probability outputs form:

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

Stage 7: Meta-Learning

Logistic regression learns the optimal combination of base-model probabilities.

Stage 8: Risk Classification

The final probability is converted into a screening classification.

Stage 9: Professional Assessment

Potentially high-risk learners are referred for appropriate assessment.

11. Research Contribution

The contribution of this study can be summarized in four dimensions.

11.1 Feature-Space Contribution

The study demonstrates the usefulness of RF-RFE for reducing a 196-predictor behavioural feature space.

11.2 Algorithmic Contribution

The study combines three complementary tree-based learners:

$$RF + XGBoost + ET.$$

11.3 Ensemble Contribution

The probability outputs are combined using a logistic-regression meta-classifier.

11.4 Screening Contribution

The framework emphasizes recall, F1 and MCC in addition to conventional accuracy.

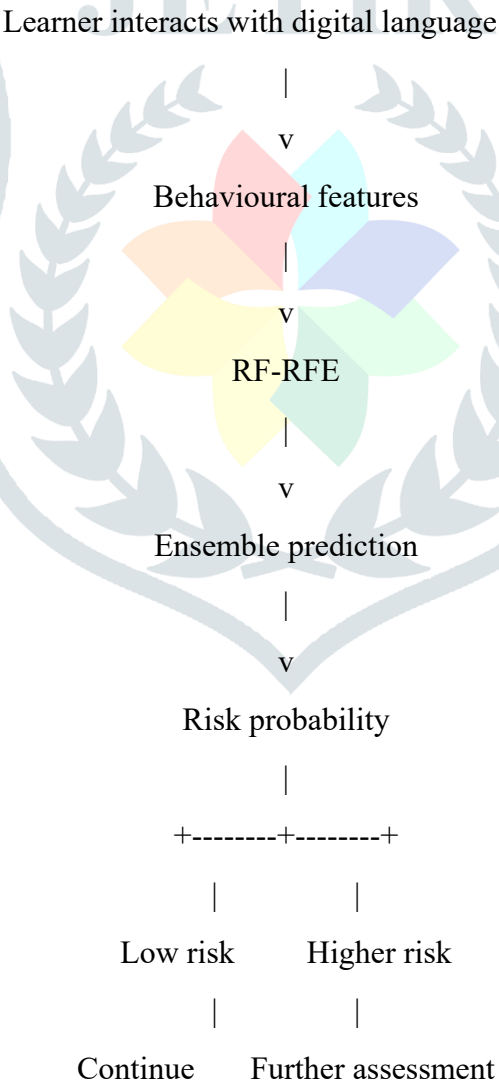
This is important because the objective of early dyslexia screening is not simply to maximize the number of correctly classified observations but to identify potentially affected learners.

12. Implications for Educational Technology

The proposed framework has potential applications in digital learning environments.

A system could potentially operate continuously during educational activities and generate a screening alert when behavioural patterns indicate elevated risk.

For example:



Such a system could potentially reduce the time between the emergence of observable difficulties and referral for assessment.

However, implementation should incorporate privacy protection, transparency, fairness and human oversight.

13. Limitations

Several limitations should be acknowledged.

13.1 Single Dataset

The study is based on the Dyt-desktop dataset. External validation is necessary before generalizing the findings to other populations.

13.2 Class Imbalance

Only 10.76% of observations belong to the dyslexia-positive class.

This makes minority-class metrics particularly important.

13.3 Threshold Dependence

The reported classification results use a 0.50 threshold.

A different threshold could substantially change recall and specificity.

13.4 Feature-Selection Stability

RF-RFE feature selection may vary across folds.

Future research should calculate feature-selection stability across repeated experiments.

13.5 No Causal Interpretation

A predictive feature should not automatically be interpreted as a cause of dyslexia.

RF-RFE identifies variables useful for prediction, not causal determinants.

13.6 External Generalizability

Differences in language, age, educational environment and assessment conditions could influence performance.

14. Recommended Validation Strategy for a Final Journal Version

For a stronger confirmatory study, the following framework is recommended:

$$Dataset \rightarrow Outer\ CV \rightarrow Training\ Fold \rightarrow Preprocessing \rightarrow RF-RFE \rightarrow Hyperparameter\ Tuning \rightarrow Inner\ CV \rightarrow Model \rightarrow Outer\ Test$$

The final study should additionally report:

1. confidence intervals;
2. precision-recall AUC;
3. ROC-AUC confidence intervals;
4. calibration;
5. confusion matrices;

6. threshold sensitivity analysis;
7. RF-RFE feature stability;
8. SHAP explanations;
9. subgroup/fairness analysis; and
10. external validation.

These additions would make the study substantially stronger for a high-quality journal submission.

15. Future Research

Future research should extend the framework in several directions.

15.1 Optimized Threshold

Instead of fixing:

$$Threshold = 0.50,$$

the threshold should be optimized according to the screening objective.

A sensitivity-oriented threshold could be selected using validation data.

15.2 Precision-Recall Analysis

Because dyslexia is the minority class, precision-recall curves should complement ROC curves.

15.3 Hyperparameter Optimization

RF, XGBoost and ET should undergo nested randomized or Bayesian hyperparameter optimization.

15.4 SHAP Explainability

SHAP can be incorporated to answer:

Which behavioural features most strongly influenced the dyslexia prediction?

The framework can therefore progress from:

$$Feature\ Selection \rightarrow Prediction$$

to:

$$Feature\ Selection \rightarrow Prediction \rightarrow Explanation.$$

15.5 External Validation

The final model should be evaluated using an independent dyslexia dataset.

15.6 Fairness

Performance should be compared across demographic and linguistic subgroups.

16. Conclusion

This study developed and empirically evaluated an ensemble machine-learning framework for early dyslexia prediction using Recursive Feature Elimination.

The analysis was based on the Dyt-desktop behavioural dataset containing **3,644 observations, 196 predictors and 392 dyslexia-positive cases**. The substantial class imbalance makes conventional accuracy insufficient as the sole evaluation criterion.

Three heterogeneous tree-based learners—Random Forest, XGBoost and Extra Trees—were investigated.

XGBoost produced the strongest individual baseline performance, achieving **90.64% accuracy and 0.8845 ROC-AUC**.

The application of RF-RFE produced an important improvement. **RF-RFE + XGBoost achieved 90.81% accuracy, 31.89% recall, 42.74% F1-score, 0.8858 ROC-AUC and 0.4122 MCC.** These results provide empirical support for the use of recursive feature elimination in high-dimensional behavioural dyslexia prediction.

The heterogeneous stacking ensemble produced an even stronger sensitivity-oriented result, achieving **71.17% recall, 51.71% F1-score and 0.4644 MCC**, although accuracy decreased to 85.70%.

The findings therefore demonstrate that no single metric adequately describes dyslexia prediction performance.

The study's principal methodological contribution is the integration of:

$$\boxed{RF - RFE + RF + XGBoost + ET + Stacking}$$

into a unified early-screening framework.

RF-RFE reduces the feature space, heterogeneous learners capture complementary patterns and stacking combines their predictions.

The empirical evidence suggests that **RF-RFE + XGBoost is the strongest individual configuration**, while the **RF-XGBoost-ET stacking ensemble is the strongest sensitivity-oriented configuration** under the tested threshold.

The framework should consequently be regarded as an **AI-assisted early-warning and screening system rather than an autonomous diagnostic instrument**.

Before deployment, future research should undertake nested validation, optimized decision thresholds, precision-recall analysis, confidence intervals, SHAP-based explainability, calibration, fairness assessment and independent external validation.

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.

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

[4] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *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] Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2), 241–259.

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

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

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

[11] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), e0118432.

[12] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21, 6.