ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep377

Optimizing Dyslexia Prediction Through Hyperparameter-Tuned Ensemble Machine Learning

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

Professor[2]

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

Correspondence Author: Patrick Ufomba Nwogu[1*]

Publication Date: 2026/09/22

Abstract: Dyslexia is a specific learning difficulty associated with persistent problems in accurate and fluent word recognition, decoding and spelling. Early identification can support timely educational intervention; however, computational screening models must address high-dimensional behavioural data, class imbalance and sensitivity to model configuration. This study investigates the optimization of dyslexia prediction through hyperparameter-tuned ensemble machine learning using the Dyt-desktop behavioural dataset derived from the Rello et al. dyslexia-screening study. The dataset contains 3,644 observations, 196 predictors and 392 dyslexia-positive cases, corresponding to a dyslexia prevalence of 10.76%. Three heterogeneous tree-based learners—Random Forest (RF), XGBoost (XGB) and Extra Trees (ET)—were investigated, together with Random-Forest Recursive Feature Elimination (RF-RFE) and probability-level stacking. The optimization process focused on model parameters controlling ensemble size, tree complexity, feature sampling, learning rate and regularization. Performance was assessed using accuracy, precision, recall, specificity, F1-score, ROC-AUC and Matthews Correlation Coefficient (MCC). Existing analysis of the supplied dataset showed that the untuned XGBoost model achieved 90.64% accuracy and 0.8845 ROC-AUC, while RF-RFE-XGBoost improved performance to 90.81% accuracy, 31.89% recall, 42.74% F1-score, 0.8858 ROC-AUC and 0.4122 MCC. Hyperparameter tuning further improved XGBoost to 91.22% accuracy, 70.59% precision, 30.77% recall, 98.46% specificity, 42.86% F1-score, 0.8921 ROC-AUC and 0.4285 MCC. Tuned Extra Trees achieved the highest recall among the individually tuned learners at 44.87%. The findings demonstrate that optimization improves discrimination, but that the model with the highest accuracy is not necessarily the most appropriate screening model. The results support a multi-objective optimization strategy in which ROC-AUC, recall, F1-score and MCC are considered alongside accuracy. The study provides an empirical framework for developing reproducible and sensitivity-aware machine-learning systems for early dyslexia screening.

Keywords: *Dyslexia Prediction; Machine Learning; Hyperparameter Optimization; Random Forest; Xgboost; Extra Trees; RF-RFE; Ensemble Learning; Stacking; Behavioural Data; Educational Data Mining.*

How to Cite: Patrick Ufomba Nwogu; Chigozirim Ajaegbu; Dr. Faruk Umar Ambursa; Dr. Femi Adeluyi (2026) Optimizing Dyslexia Prediction Through Hyperparameter-Tuned Ensemble Machine Learning. *International Journal of Innovative Science and Research Technology*, 11(9), 626-635. https://doi.org/10.38124/ijisrt/26sep377

I. INTRODUCTION

Dyslexia is a specific learning difficulty characterized primarily by persistent difficulties with accurate and fluent word recognition, decoding and spelling. Its effects can extend beyond reading performance to educational achievement, learner confidence and long-term academic participation. Early identification is therefore important because appropriate educational intervention can be introduced before difficulties become entrenched.

Conventional dyslexia assessment relies on specialist-administered procedures involving educational, linguistic and psychological evaluation. Although professional assessment remains essential, the growth of digital learning environments has created opportunities to complement conventional approaches with computational screening. Behavioural

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep377

information generated while learners interact with digital tasks can contain measurable indicators associated with language processing and task performance.

Rello et al. demonstrated the feasibility of this approach using an online gamified assessment. Their study involved 3,644 participants, including 392 participants with professionally diagnosed dyslexia, and used 196 features derived from demographic characteristics and interaction with 32 linguistic exercises. Their Random Forest model demonstrated that machine learning could identify dyslexia risk from behavioural interaction data.

The present study builds upon this research direction by investigating whether systematic model optimization can improve dyslexia prediction beyond a conventional base-model configuration. This question is particularly important because machine-learning performance is dependent not only on algorithm selection but also on the configuration of algorithmic hyperparameters.

Hyperparameters determine characteristics such as the number of trees, tree depth, learning rate, feature subsampling and regularization. Poorly selected values may result in underfitting, excessive variance or inefficient learning. Hyperparameter optimization therefore represents an important stage in developing robust predictive systems.

The problem is further complicated by the strong class imbalance in the Dyt-desktop dataset. Only 392 of 3,644 participants are dyslexia-positive, while 3,252 are non-dyslexic. Thus, a classifier predicting every participant as non-dyslexic would obtain approximately 89.24% accuracy while identifying no dyslexia cases.

Consequently, accuracy alone is insufficient for evaluating a dyslexia-screening model. Recall, precision, F1-score, ROC-AUC and MCC provide additional information about minority-class detection. Precision-recall analysis is particularly important in imbalanced classification because ROC curves can sometimes give an overly optimistic visual impression of classifier performance when the positive class is rare.

This study therefore investigates hyperparameter-tuned ensemble machine learning for dyslexia prediction using the Dyt-desktop behavioural dataset.

➢ *Research Problem*

Although previous work has established the feasibility of machine-learning-based dyslexia screening, three methodological problems remain.

First, predictive performance can be highly sensitive to hyperparameter configuration. Second, the severe imbalance between dyslexia-positive and dyslexia-negative observations makes conventional accuracy-based optimization potentially misleading. Third, heterogeneous learners may make different errors, suggesting that ensemble strategies can potentially improve screening performance.

The central problem addressed in this study is therefore:

How can hyperparameter optimization and heterogeneous ensemble learning be used to improve the robustness and effectiveness of machine-learning-based dyslexia prediction?

➢ *Research Objectives*

The objectives are to:

- evaluate Random Forest, XGBoost and Extra Trees for dyslexia prediction;
- establish baseline performance for the three models;
- investigate RF-RFE feature selection;
- optimize the principal hyperparameters of the base learners;
- compare tuned and untuned models;
- evaluate whether heterogeneous ensemble learning provides complementary predictive information;
- determine the most appropriate performance measures for imbalanced dyslexia prediction; and
- establish a reproducible optimization framework for future dyslexia-screening research.

➢ *Research Questions*

- RQ1: Does hyperparameter optimization improve the performance of individual dyslexia classifiers?
- RQ2: Which of RF, XGBoost and Extra Trees provides the strongest optimized prediction?
- RQ3: Does RF-RFE feature selection improve predictive performance?
- RQ4: Can heterogeneous stacking improve detection of the minority dyslexia class?
- RQ5: Which performance metrics provide the most meaningful assessment of dyslexia screening under class imbalance?

II. RELATED WORK

➢ *Machine Learning for Dyslexia Detection*

Machine learning has increasingly been investigated as a computational approach to identifying learning difficulties. Dyslexia is particularly suitable for behavioural modelling because reading and language-related tasks produce measurable patterns involving accuracy, response behaviour and interaction.

Rello et al. developed a gamified online dyslexia screening approach based on behavioural interaction data. The authors collected data from 3,644 participants and used 196 features generated from demographic variables and 32 linguistic exercises. Their Random Forest approach achieved useful dyslexia-class detection and demonstrated that machine learning can support screening without requiring specialized hardware.

An important feature of the Rello et al. study is that it treated the machine-learning model as a screening

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep377

mechanism rather than a diagnostic replacement. This distinction is retained in the present research.

Recent developments in artificial intelligence have expanded dyslexia research into multiple modalities, including behavioural information, eye tracking, handwriting, EEG and neuroimaging. Behavioural approaches have an important practical advantage because the necessary information can potentially be collected through ordinary digital learning activities.

➢ *Random Forest*

Random Forest was introduced by Breiman as an ensemble method based on multiple randomized decision trees.

The fundamental prediction can be represented as:

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

Where B represents the number of trees and $T_b(X)$represents the prediction generated by tree b.

RF is attractive for structured behavioural data because it can model nonlinear relationships and interactions without requiring strong assumptions concerning the distribution of predictor variables.

➢ *XGBoost*

XGBoost is a scalable gradient-boosting framework introduced by Chen and Guestrin.

Unlike Random Forest, which builds trees largely independently and aggregates them, gradient boosting constructs trees sequentially, with subsequent learners focusing on reducing the errors of preceding learners.

The model can be represented as:

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

where f_k represents an individual tree and K is the number of boosting iterations.

The sequential optimization mechanism makes XGBoost particularly suitable for nonlinear structured datasets.

➢ *Extra Trees*

Extra Trees, or Extremely Randomized Trees, introduces additional randomization into the tree construction process. Geurts et al. demonstrated that randomized tree ensembles can provide strong predictive performance while reducing dependence on individual observations and split configurations.

The diversity introduced by Extra Trees makes it an appropriate candidate for heterogeneous ensemble learning alongside RF and XGBoost.

➢ *Hyperparameter Optimization*

Hyperparameters are model-design parameters that are selected before or during model training rather than estimated directly from the observations. Examples include tree depth, number of estimators, learning rate, minimum samples per leaf and feature-sampling parameters.

Bergstra and Bengio demonstrated that random search can be more efficient than grid search because only a subset of hyperparameters may have substantial influence on model performance.

This finding is relevant to dyslexia prediction because the dataset contains many predictors and multiple potentially influential model parameters. Systematic random search can therefore provide a computationally efficient strategy for exploring the model configuration space.

➢ *Ensemble Learning and Stacking*

Ensemble learning combines multiple predictive models to exploit complementary error structures.

Stacking extends this principle by using the predictions of several base learners as inputs to a second-level model. Wolpert introduced stacked generalization as a method for improving generalization by learning how to combine the outputs of multiple generalizers.

For the present study, the base learners are:

$$RF,\ XGB,\ ET$$

and their probability outputs form:

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

A logistic-regression meta-classifier can then estimate:

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

To avoid information leakage, the meta-features should be generated using out-of-fold predictions rather than predictions generated from models trained on the same observations.

➢ *Evaluation under Class Imbalance*

Class imbalance creates a particular problem for dyslexia prediction. Accuracy can be dominated by the majority class. In the Dyt-desktop dataset, the majority class alone accounts for 89.24% of observations.

MCC is useful in this context because it incorporates all four elements of the confusion matrix and provides a more balanced measure of binary classification performance. The literature has highlighted the advantages of MCC over accuracy and F1-score for imbalanced binary classification.

Accordingly, the present study reports accuracy, precision, recall, specificity, F1-score, ROC-AUC and MCC.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep377

III. MATERIALS AND METHODS

➢ *Dataset*

The empirical analysis used the Dyt-desktop.csv dataset associated with the behavioural dyslexia prediction study.

Table 1 The Dataset Contains

Characteristic	Value
Total observations	3,644
Predictors	196
Target variable	Dyslexia
Dyslexia cases	392
Non-dyslexia cases	3,252
Dyslexia prevalence	10.76%
Non-dyslexia prevalence	89.24%
Linguistic tasks	32
Measures per task	6

The 196 predictors consist of demographic and behavioural variables. Demographic variables include gender, native-language status and other-language status, while behavioural variables include clicks, hits, misses, score, accuracy and miss rate across the linguistic tasks.

The original study describes the same general structure: four demographic features followed by 192 performance features derived from six measures across 32 questions.

➢ *Target Variable*

The dependent variable is binary:

$$Y = \begin{cases} 1 & \text{Dyslexia} \\ 0 & \text{Non-dyslexia} \end{cases}$$

The positive class represents participants identified as having dyslexia.

➢ *Data Preprocessing*

The preprocessing framework comprised:

- separation of predictor variables from the target;
- identification of categorical and numerical variables;
- conversion of numerical fields to appropriate numerical representations;
- median imputation for missing numerical observations;
- most-frequent imputation for categorical observations;
- categorical encoding;
- class-weighted model training;
- stratified cross-validation;
- fold-specific feature selection; and
- generation of out-of-fold predictions for stacking.

The earlier analysis explicitly specified that preprocessing and RF-RFE selection should be performed inside the training procedure to reduce information leakage.

➢ *RF-RFE Feature Selection*

RF-RFE was incorporated to reduce dimensionality.

Let the initial feature set be:

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

where $p = 196$.

A Random Forest is trained and feature importance values are calculated. The least informative predictors are progressively removed until a reduced subset $F^*$is obtained.

The resulting transformation can be expressed as:

$$X_{196} \rightarrow RF\ Ranking \rightarrow Feature\ Elimination \rightarrow X_k$$

The selected features are subsequently supplied to the downstream learners.

Feature selection is particularly relevant because redundant behavioural variables may increase model complexity without contributing equivalent predictive information.

➢ *Base Learners*

Three heterogeneous ensemble algorithms were evaluated:

- Random Forest;
- XGBoost; and
- Extra Trees.

The selection was deliberate. RF represents bagging-based tree aggregation, XGBoost represents sequential boosting, and ET provides additional tree randomization. Their different learning mechanisms create the potential for complementary prediction errors.

➢ *Hyperparameter Optimization*

Hyperparameter optimization was conducted to identify improved configurations for each base learner.

The principal parameter families considered were:

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep377

- *Random Forest*

✓ number of estimators;
✓ maximum tree depth;
✓ minimum samples required for splitting;
✓ minimum samples per leaf;
✓ maximum features;
✓ class weighting.

- *XGBoost*

✓ number of estimators;
✓ maximum tree depth;
✓ learning rate;
✓ subsampling;
✓ column subsampling;
✓ minimum child weight;
✓ regularization parameters.

- *Extra Trees*

✓ number of estimators;
✓ maximum depth;
✓ minimum samples per split;
✓ minimum samples per leaf;
✓ maximum features;
✓ class weighting.

Randomized hyperparameter search was used because random search can explore a larger effective parameter space with a fixed computational budget than conventional exhaustive grid search.

The existing analysis used stratified two-fold cross-validation during the randomized search, with ROC-AUC as the optimization criterion.

➢ *Ensemble Stacking*

After individual learners were optimized, their probability outputs were considered for heterogeneous stacking.

The architecture was:

$$Dyt - desktop \rightarrow Preprocessing \rightarrow RF - RFE \rightarrow \begin{cases} RF \\ XGBoost \rightarrow P_{RF}, P_{XGB}, P_{ET} \\ ET \end{cases} \rightarrow Logistic\ Meta - Classifier \rightarrow Prediction$$

The stacking approach is based on the principle of stacked generalization.

➢ *Evaluation Metrics*

Seven measures were used.

- *Accuracy*

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

- *Precision*

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

- *Recall*

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

- *Specificity*

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

- *F1-score*

$$F1 = \frac{2(Precision)(Recall)}{Precision + Recall}$$

- *ROC-AUC*

ROC-AUC evaluates the ability of a classifier to discriminate between the positive and negative classes over different decision thresholds.

- *Matthews Correlation Coefficient*

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

MCC is particularly useful for imbalanced binary classification because it incorporates TP, TN, FP and FN simultaneously. Because the positive class represents only 10.76% of observations, recall and MCC are particularly important in addition to accuracy.

IV. EXPERIMENTAL RESULTS

➢ *Baseline Base-Learner Results*

The previously established analysis of the Dyt-desktop dataset produced the following baseline results.

Table 1 Baseline Performance of Individual Learners

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

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep377

XGBoost was the strongest baseline learner according to accuracy, recall, F1-score, ROC-AUC and MCC.

RF and ET produced extremely high specificity but substantially lower recall. This indicates that both models were conservative in assigning participants to the dyslexia-positive class.

➢ *Effect of RF-RFE*

The application of RF-RFE produced the following results.

Table 2 Performance After RF-RFE Feature Selection

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 + 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-RFE improved XGBoost across several important measures. Accuracy increased from 90.64% to 90.81%, recall from 29.08% to 31.89%, F1 from 40.07% to 42.74%, ROC-AUC from 0.8845 to 0.8858, and MCC from 0.3912 to 0.4122.

This provides evidence that feature selection can improve the signal-to-noise ratio of the predictor space.

➢ *Hyperparameter-Tuned Models*

The principal contribution of the present analysis is the optimization of the three base learners.

Table 3 Hyperparameter-Tuned Model Performance

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

The optimized results show a clear distinction between overall discrimination and minority-class sensitivity.

Tuned XGBoost achieved the highest accuracy, precision, specificity, F1-score, ROC-AUC and MCC. Its accuracy increased to 91.22% and ROC-AUC to 0.8921.

Tuned Extra Trees, however, achieved the highest recall at 44.87%. This demonstrates that the optimal algorithm depends on the intended screening objective.

➢ *Improvement from Hyperparameter Optimization*

Table 4 Baseline-to-Tuned Comparison

Model	Baseline Accuracy	Tuned Accuracy	Change	Baseline AUC	Tuned AUC	Change
RF	89.65%	90.81%	+1.16 pp	0.8773	0.8763	−0.0010
XGBoost	90.64%	91.22%	+0.58 pp	0.8845	0.8921	+0.0076
ET	89.54%	87.11%	−2.43 pp	0.8709	0.8685	−0.0024

The results reveal an important finding: hyperparameter optimization does not guarantee improvement across every model and every metric.

For XGBoost, optimization produced simultaneous improvements in accuracy and ROC-AUC. For RF, accuracy improved while ROC-AUC remained almost unchanged. For ET, the tuned model produced substantially higher recall but lower accuracy and ROC-AUC.

This confirms that optimization should be formulated according to the actual application objective rather than treated as an automatic process for maximizing every metric.

➢ *Stacking Results*

The earlier heterogeneous stacking experiment generated the following performance:

Table 5 Stacking Ensemble Performance

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 model produced considerably higher dyslexia-class recall than any individual model. Its recall of 71.17% was substantially greater than the 44.87% obtained by tuned Extra Trees.

However, the increase in recall was accompanied by reductions in accuracy and specificity.

This illustrates the fundamental trade-off between sensitivity and false-positive control in screening applications.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep377

V. DISCUSSION

➢ *Hyperparameter Optimization Improves Model Discrimination*

The results provide evidence that hyperparameter tuning can improve dyslexia prediction, particularly for XGBoost.

The baseline XGBoost model achieved 90.64% accuracy and 0.8845 ROC-AUC. Following optimization, accuracy increased to 91.22% and ROC-AUC increased to 0.8921.

The improvement in ROC-AUC is particularly important because it suggests that optimization improved the model's ranking ability rather than simply shifting its classification threshold.

The result is consistent with the general principle that hyperparameter selection can have substantial effects on predictive performance. Bergstra and Bengio demonstrated that random search can efficiently identify high-performing parameter configurations because only a subset of parameters typically has a strong effect on model quality.

➢ *XGBoost as the Strongest Accuracy-Oriented Learner*

Among the individually optimized models, XGBoost was the strongest overall performer.

Its 91.22% accuracy exceeded the majority-class baseline of approximately 89.24%. More importantly, its ROC-AUC of 0.8921 represented the highest discrimination score among the evaluated individual models.

Its MCC of 0.4285 also indicates stronger overall binary classification association than the baseline learners.

Therefore, where the application emphasizes discrimination and a balanced combination of predictive measures, tuned XGBoost is the strongest individual candidate.

➢ *Extra Trees and Sensitivity-Oriented Screening*

Tuned Extra Trees produced a different pattern.

Although its accuracy decreased to 87.11%, recall increased to 44.87%.

This distinction is important because a dyslexia-screening application may place greater importance on identifying potentially affected learners than on maximizing overall accuracy.

A false negative means that a participant who may require further assessment is classified as low risk. In a screening context, such an error may be more consequential than a false positive that leads to additional assessment.

Therefore, tuned Extra Trees could have practical value when sensitivity is prioritized.

➢ *The Importance of Stacking*

The stacking ensemble generated the highest recall, F1-score and MCC in the existing analysis.

Its 71.17% recall indicates that heterogeneous combination substantially increased the ability to identify dyslexia-positive participants.

The result supports the underlying motivation for ensemble learning: different models learn different decision structures and may produce complementary errors.

The principle is consistent with stacked generalization, in which a second-level learner is trained to combine predictions generated by first-level learners.

However, stacking also reduced specificity to 87.45%.

Consequently, the stacking model should not be described simply as "better" than the individual models. Rather, it is more sensitivity-oriented.

➢ *Accuracy is not Sufficient*

The dataset's 10.76% positive-class prevalence makes accuracy particularly problematic.

For example, the baseline RF model achieved 89.65% accuracy but only 5.87% recall. This means that its high accuracy was achieved while identifying very few dyslexia-positive cases.

Similarly, Extra Trees achieved 89.54% accuracy but only 3.83% recall.

These results demonstrate why dyslexia screening research should report multiple metrics.

Precision-recall analysis is especially useful when the positive class is relatively rare because it focuses directly on the relationship between correctly identified positives and positive predictions.

➢ *RF-RFE Improves the Predictive Representation*

RF-RFE-XGBoost produced 90.81% accuracy and 0.8858 ROC-AUC compared with 90.64% and 0.8845 for baseline XGBoost.

Although the improvements are relatively small, the direction of improvement across recall, F1 and MCC is notable.

The results suggest that reducing the predictor space can remove some redundant or weak information while retaining features useful for dyslexia prediction.

However, feature selection should be conducted inside the cross-validation procedure. Performing feature selection on the complete dataset before validation can introduce information leakage and produce overly optimistic estimates.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep377

➢ *Comparison with the Original Dyslexia Study*

The present study extends the methodological direction of Rello et al.

The original research used Random Forest with weighted attributes and 10-fold cross-validation. The authors reported that a threshold of 0.24 provided a useful balance between the relevant classification errors and emphasized that the model should be treated as a screening mechanism rather than a diagnostic system.

The current study differs by systematically evaluating three heterogeneous tree-based learners, applying RF-RFE feature selection and investigating hyperparameter optimization and stacking.

Thus, the contribution is not simply the application of another classifier to the same data. Rather, it examines how model configuration, feature representation and heterogeneous ensemble structure influence dyslexia-screening performance.

➢ *Practical Implications*

The findings have several implications for digital educational screening.

First, behavioural machine learning can potentially provide a scalable first-stage screening mechanism.

Second, a single performance metric should not determine model selection.

Third, hyperparameter optimization can improve the predictive discrimination of individual models.

Fourth, stacking may be valuable where sensitivity is prioritized.

Finally, any resulting system should be presented as a screening and decision-support tool rather than a diagnostic instrument.

This is consistent with the original Rello et al. study, which explicitly cautioned that machine-learning screening should not replace professional dyslexia assessment.

➢ *Proposed Optimized Ensemble Framework*

Based on the experimental findings, an optimized dyslexia prediction architecture can be represented as:

$$\boxed{Dyt - desktop \rightarrow Data\ Cleaning \rightarrow Encoding \rightarrow Class\ Handling \rightarrow RF - RFE}$$

followed by:

$$\boxed{RF\ XGBoost\ ET}$$

with hyperparameter optimization applied independently:

$$RF \rightarrow RF^*$$

$$XGBoost \rightarrow XGB^*$$

$$ET \rightarrow ET^*$$

The optimized probability outputs are then:

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

Which are supplied to the stacking meta-classifier:

$$M \rightarrow Logistic\ Regression \rightarrow P(Dyslexia)$$

A decision threshold can subsequently be selected according to the intended application.

For general classification, the threshold may be 0.50. For screening, however, a lower threshold can increase sensitivity at the expense of specificity. The original dyslexia study provides an important precedent by selecting a threshold of approximately 0.24 rather than relying exclusively on 0.50.

This suggests that future optimization should consider not only hyperparameters but also decision-threshold optimization.

➢ *Model Selection Strategy For Dyslexia Screening*

The results suggest three different model-selection strategies.

➢ *Strategy 1: Accuracy-oriented*

If the principal objective is general classification performance:

Tuned XGBoost is preferred because it achieved:

✓ 91.22% accuracy;
✓ 70.59% precision;
✓ 98.46% specificity;
✓ 42.86% F1;
✓ 0.8921 ROC-AUC; and
✓ 0.4285 MCC.

➢ *Strategy 2: Sensitivity-oriented*

If the principal objective is to minimize missed dyslexia cases:

Tuned Extra Trees or stacking should be considered.

Tuned ET achieved 44.87% recall, whereas stacking achieved 71.17% recall.

➢ *Strategy 3: Balanced screening*

For a practical screening system, the recommended approach is: "Optimized Base Learners"+"Out-of-Fold Stacking"+"Threshold Optimization " rather than selecting a model exclusively according to accuracy.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep377

VI. LIMITATIONS

Several limitations should be acknowledged.

First, the empirical analysis is based on a single behavioural dataset. Although the dataset is relatively large for dyslexia research, external validation on an independent cohort is necessary before general deployment.

Second, the dataset originates from a specific linguistic and educational context. Consequently, model performance may not transfer directly to other languages, age groups or educational systems.

Third, the dyslexia-positive class is substantially smaller than the negative class. This creates uncertainty around minority-class estimates and makes precision-recall analysis important.

Fourth, the current optimization results should be interpreted as empirical results from the established analysis rather than as evidence that the reported hyperparameter configuration is globally optimal.

Fifth, stacking performance depends strongly on the method used to generate meta-features. Leakage-free out-of-fold predictions are therefore essential.

Sixth, model explainability remains important. SHAP provides a theoretically grounded approach for explaining individual predictions and identifying influential predictors in complex machine-learning models.

Finally, professional dyslexia assessment remains necessary. The proposed system should be considered an early screening mechanism rather than an autonomous diagnostic system.

VII. FUTURE RESEARCH

Future work should focus on six areas.

- *Nested Cross-Validation*

Hyperparameter optimization should ideally be embedded inside nested cross-validation so that model-selection bias is minimized.

- *Threshold Optimization*

Rather than fixing the classification threshold at 0.50, future studies should optimize the threshold according to the screening objective.

- *Precision-Recall Analysis*

Because dyslexia-positive observations are relatively rare, future studies should report precision-recall curves and area under the precision-recall curve in addition to ROC-AUC.

- *Explainable AI*

SHAP or related explainability methods should be applied to determine which behavioural variables contribute most strongly to dyslexia predictions.

This would increase the transparency of the proposed screening framework.

- *External Validation*

The optimized model should be evaluated on an independent dataset collected from a different population, educational environment and, ideally, linguistic context.

- *Fairness and Subgroup Analysis*

Performance should be evaluated across relevant demographic and linguistic groups to identify potential disparities in sensitivity, specificity and false-positive rates.

VIII. CONCLUSION

This study investigated the optimization of dyslexia prediction through hyperparameter-tuned ensemble machine learning using the Dyt-desktop behavioural dataset containing 3,644 observations and 196 predictors.

Three heterogeneous tree-based algorithms—Random Forest, XGBoost and Extra Trees—were evaluated, together with RF-RFE feature selection and heterogeneous stacking.

The baseline results demonstrated that XGBoost was the strongest individual learner, achieving 90.64% accuracy and 0.8845 ROC-AUC. RF-RFE further improved XGBoost to 90.81% accuracy, 31.89% recall, 42.74% F1-score, 0.8858 ROC-AUC and 0.4122 MCC.

Hyperparameter optimization produced additional improvement. Tuned XGBoost achieved 91.22% accuracy and 0.8921 ROC-AUC, representing the strongest overall individual configuration. It also obtained 70.59% precision, 30.77% recall, 98.46% specificity, 42.86% F1-score and 0.4285 MCC.

However, tuned Extra Trees achieved the highest recall among the individually tuned models at 44.87%, while the previously developed stacking configuration achieved 71.17% recall, 51.71% F1-score and 0.4644 MCC.

The results therefore demonstrate that model optimization is inherently objective-dependent. A model that maximizes accuracy may not be the model that maximizes dyslexia detection. For screening applications, sensitivity, F1-score, MCC and precision-recall characteristics should be considered alongside accuracy and ROC-AUC.

The principal contribution of this study is a methodological framework that integrates RF-RFE feature selection, hyperparameter optimization, heterogeneous ensemble learning and screening-oriented evaluation. The framework provides a reproducible foundation for developing machine-learning-based dyslexia screening systems from behavioural learning data.

Nevertheless, the proposed framework should be regarded as an AI-assisted screening and decision-support approach rather than a diagnostic replacement for

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep377

professional dyslexia assessment. External validation, threshold optimization, explainability, fairness analysis and prospective evaluation are required before practical deployment.

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

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

[7]. 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.

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

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

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