

# Feature-Efficient and Interpretable Dyslexia Detection via Soft Voting Ensemble Learning

Almaan Khan and Shujaat Khan\*

Cite <https://doi.org/10.64589/juri/231551>

Submitted: March 11, 2026 Revised: May 07, 2026 Accepted: May 13, 2026

## ABSTRACT

Early dyslexia prediction is difficult because of its varied behavioral signals and the need for timely, reliable screening in educational settings. This study proposes a soft-voting ensemble framework that integrates multiple boosting classifiers (AdaBoost, Gradient Boosting, XGBoost, and CatBoost) to improve predictive performance while maintaining interpretability. The model is evaluated using stratified 10-fold cross-validation on a publicly available dataset derived from gamified cognitive assessments, with minimum Redundancy Maximum Relevance feature selection and SHapley Additive exPlanations analysis used to enhance transparency. Experimental results reveal that the ensemble outperforms individual models and recent methods, achieving 91.6% accuracy, 90.7% precision, 91.6% recall, and a 90.9% F1-score using 165 selected features with class weighting, thereby supporting its practicality for explainable early dyslexia screening.

**Keywords:** dyslexia detection, ensemble learning, boosting classifiers, soft voting

## 1. INTRODUCTION

Dyslexia is a common neurodevelopmental learning disorder affecting an estimated 10–15% of the global population, impairing the acquisition of reading, spelling, and phonological processing skills despite normal intelligence and adequate educational exposure<sup>1</sup>. In Saudi Arabia, its reported prevalence is approximately 7%<sup>2</sup>. If left undiagnosed, dyslexia can result in academic underachievement, reduced self-esteem, and social exclusion, making early detection essential for timely intervention and educational support.

Traditionally, dyslexia is diagnosed by trained professionals through comprehensive psychometric evaluations encompassing a broad spectrum of cognitive and linguistic abilities, including phonological processing, word reading accuracy, reading fluency, text comprehension, spelling, working memory, and processing speed. Although clinically reliable, such evaluations are time-consuming, costly, and difficult to scale for large populations. Their dependence on specialized expertise further limits their feasibility in low-resource settings<sup>3</sup>. These constraints have motivated the search for automated, scalable alternatives.

In recent years, artificial intelligence (AI) and machine learning (ML) techniques have received increasing attention for dyslexia detection. Existing approaches utilize diverse data modalities, from linguistic and cognitive measures to neurophysiological signals such as electroencephalography (EEG), eye-tracking, and behavioral interaction data<sup>2,4</sup>. Although many of these models report promising performance, a considerable number rely on single classifiers, including support vector machines

(SVM), random forests (RF), and neural networks. Such standalone models may be prone to overfitting, particularly when trained on noisy or imbalanced datasets<sup>2,5,6</sup>.

Ensemble learning, which combines multiple base learners to enhance predictive accuracy and generalization, offers a compelling alternative. Techniques such as bagging, boosting, and stacking have demonstrated strong performance across biomedical and educational applications<sup>7</sup>, with boosting-based methods such as AdaBoost, Gradient Boosting (GB), and XGBoost being particularly effective because of their sequential error-correction mechanism<sup>5</sup>. However, most dyslexia-related ensemble studies are limited to individual boosting models or simple aggregation strategies like majority voting<sup>8</sup>, failing to fully leverage classifier diversity. Moreover, interpretability—an essential requirement in educational diagnostics—is often insufficiently addressed.

To overcome these limitations, this paper introduces a robust ensemble framework for early dyslexia detection that integrates multiple boosting classifiers through a soft voting mechanism. The proposed approach is evaluated on a publicly available behavioral dataset using 10-fold cross-validation. To enhance transparency and interpretability, we incorporate minimum Redundancy Maximum Relevance (mRMR) feature selection and employ SHapley Additive exPlanations (SHAP)<sup>9</sup> analysis to explain model predictions.

The remainder of the paper is organized as follows. Section 2 reviews related literature. Section 3 presents the proposed methodology. Section 4 details the experimental results and analysis. Section 5 concludes the study and outlines directions for future research.

## 2. LITERATURE REVIEW

Several ML techniques have been proposed for dyslexia detection, utilizing diverse data modalities such as eye-tracking signals, EEG recordings, gamified interaction logs, and linguistic features<sup>10–13</sup>. Although conventional classifiers (e.g., SVM, RF, and logistic regression) have achieved encouraging results<sup>14,15</sup>, they have several practical limitations.

To enhance generalization, ensemble learning strategies have been explored. Boosting-based methods are especially effective because of their iterative focus on correcting misclassified instances<sup>16–18</sup>. Nevertheless, many existing studies restrict themselves to single boosting models and fail to exploit integrative ensemble strategies that could leverage model diversity for improved robustness<sup>8</sup>.

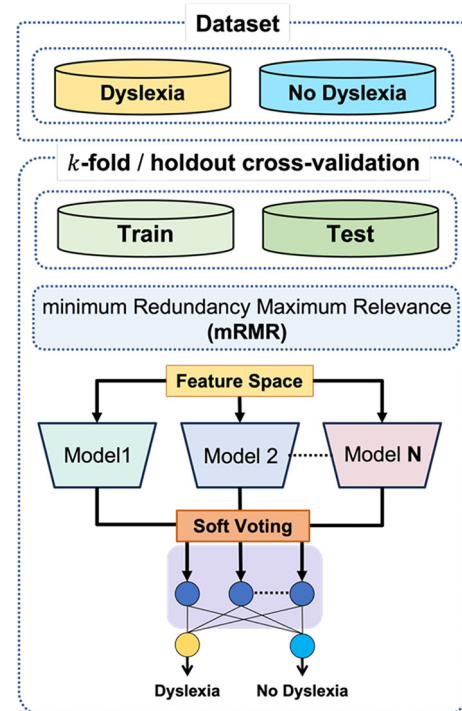
Interpretability remains an underexplored dimension in dyslexia-related ML research. Although dimensionality reduction and feature ranking techniques like correlation attribute evaluation and principal component analysis (PCA) are widely used<sup>19</sup>, few studies employ more advanced methods such as mRMR, which balances feature relevance against redundancy, or model explanation tools such as SHAP<sup>9</sup>, which provide instance-level feature attributions, despite their increasing adoption in biomedical AI applications. Reproducibility is another concern, because many studies do not share open-source code or follow standardized evaluation procedures, limiting the reliability of comparisons.

To bridge these gaps, this study proposes a robust, interpretable ensemble framework for dyslexia detection. Its main contributions are as follows: (i) a soft-voting ensemble of heterogeneous boosting models (AdaBoost, GB, XGBoost, and CatBoost), where CatBoost enables native handling of categorical features without one-hot encoding; (ii) a feature-efficient learning framework using mRMR, which jointly optimizes feature relevance and redundancy, reducing the feature space from 196 to 165 while maintaining stable performance; (iii) an interpretable modeling pipeline integrating SHAP analysis, providing both global and instance-level explanations for model predictions; (iv) a comprehensive, reproducible benchmarking framework demonstrating consistent improvements in F1-score and balanced performance compared to prior methods; and (v) open-source implementation to support transparency and reproducibility (<https://github.com/Shujaat123/Dyslexia>).

## 3. METHODOLOGY

The architecture of the proposed ensemble-based classification framework for dyslexia detection is depicted in Figure 1.

**3.1. Dataset Description.** The dataset employed was derived from Rello et al.'s research<sup>11</sup> exploring the use of interactive digital assessments to evaluate the risk of dyslexia in children. It comprises detailed interaction logs of participants who completed an online gamified test designed to measure cognitive and linguistic abilities associated with dyslexia. In addition to demographic attributes such as age, sex, and native and other spoken languages, the dataset includes behavioral metrics recorded



**Figure 1.** Overview of the proposed ensemble-based dyslexia detection framework

across 32 interactive trials, each capturing detailed performance-related measures reflecting participants' responses and interaction patterns during the assessment tasks: number of user clicks (Clicks1–Clicks32), correct responses (Hits1–Hits32), incorrect responses (Misses1–Misses32), and calculated task-specific metrics such as score, accuracy, and miss rate (Score1–Score32, Accuracy1 – Accuracy32, Missrate1–Missrate32).

Each record in the dataset also includes a binary label (Dyslexia) indicating whether the participant is at risk of dyslexia. The dataset comprises 3644 samples, with 392 labeled as at-risk (Dyslexia = 1) and 3252 as not-at-risk (Dyslexia = 0), indicating a significant class imbalance (approximately 89:11), which can influence model learning and evaluation. The dataset contains 196 features: 192 task-level interaction features and 4 demographic attributes. The “Dyslexia” column is excluded from the feature set and serves as the target label. The original data, licensed for academic use, was accessed via the supplementary dataset published with the article.

**3.2. Proposed Soft Voting Ensemble of Boosting Classifiers.** The mRMR method selects a subset of features that maximizes relevance to the target variable while minimizing redundancy among features. Formally, the objective is:

$$\max_S \left[ \frac{1}{|S|} \sum_{f_i \in S} I(f_i; y) - \frac{1}{|S|^2} \sum_{f_i, f_j \in S} I(f_i; f_j) \right]$$

where  $I(\cdot; \cdot)$  denotes mutual information.

The proposed model integrates four boosting-based classifiers AdaBoost<sup>16</sup>, GB<sup>17</sup>, XGBoost<sup>18</sup>, and CatBoost<sup>20</sup> into a soft voting ensemble to enhance predictive performance.  $\mathcal{D} =$

$\{(x_i, y_i)\}_{i=1}^N$  denotes the dataset, where  $x_i \in \mathbb{R}^d$  represents the feature vector and  $y_i \in \{0, 1\}$  is the class label.

$M$  denotes the number of base classifiers, where each classifier  $f_m(x)$  outputs the probability of class membership. The final prediction of the soft-voting ensemble is defined as:

$$\hat{y} = \operatorname{argmax}_{k \in \{0,1\}} \left( \frac{1}{M} \sum_{m=1}^M P_m(y = k|x) \right)$$

where  $P_m(y = k|x)$  represents the predicted probability of class  $k$  by the  $m$ -th classifier.

This ensemble strategy leverages the strengths of individual models, reduces variance, and improves overall robustness.

**3.3. Evaluation Protocol.** To assess the performance of the proposed model, we employed stratified 10-fold cross-validation to preserve class proportions across folds, which is particularly important given the dataset's class imbalance. The results were averaged across all folds to ensure robustness. The model was evaluated using the following standard classification metrics: accuracy, precision, recall, F1-score. Using the F1-score as the primary metric ensures that model performance is not biased toward the majority class. To prevent data leakage, feature selection using mRMR was performed within each training fold during cross-validation, ensuring that no information from the validation data influenced feature selection or model training.

## 4. RESULTS AND DISCUSSION

Extensive comparisons were performed with state-of-the-art classifiers, analyzing cross-validation results and assessing the stability and robustness of the proposed ensemble method.

**4.1. Exploratory Data Visualization.** To examine whether behavioral feature groups exhibit discriminative patterns between dyslexic and non-dyslexic participants, PCA serves

as a visualization tool to reduce each group of 32 features into a single component (PC1). The resulting distributions of PC1 values are then visualized using histograms to compare class-wise separability.

To visualize group-level differences in high-dimensional behavioral features, PCA is applied as a dimensionality reduction tool. Figure 2 presents the class-wise distributions of the first principal component (PC1) for each behavioral feature group, allowing a visual assessment of separability between dyslexic and non-dyslexic participants. The observable separation across multiple feature groups suggests that these behavioral features contain discriminative information, motivating their use in the subsequent classification framework.

To quantify these observations and compare PC1 values between the dyslexic and non-dyslexic groups, independent two-sample t-tests were conducted. All results indicate statistically significant differences: User Clicks ( $t = 10.73, p < 0.001$ ), Correct Hits ( $t = 8.20, p < 0.001$ ), Incorrect Misses ( $t = 4.64, p < 0.001$ ), and Task Scores ( $t = 8.49, p < 0.001$ ). This reinforces the discriminative power of behavioral features for dyslexia detection.

Figure 3 depicts the distribution of key demographic features, stratified by dyslexia status. Although sex is approximately balanced, males exhibit a marginally higher prevalence of dyslexia. The majority of participants report Spanish as their native language, with bilingualism (indicated by "Otherlang") present in a substantial portion of both groups. Age distribution reveals a slightly younger mean among the dyslexic group, consistent with the goal of early screening. These insights validate the relevance of both behavioral interaction logs and demographic variables in characterizing dyslexia risk, justifying the integration of both types of features in subsequent classification models.

To identify a refined and compact feature subset for model interpretability, the mRMR method was employed. A series of

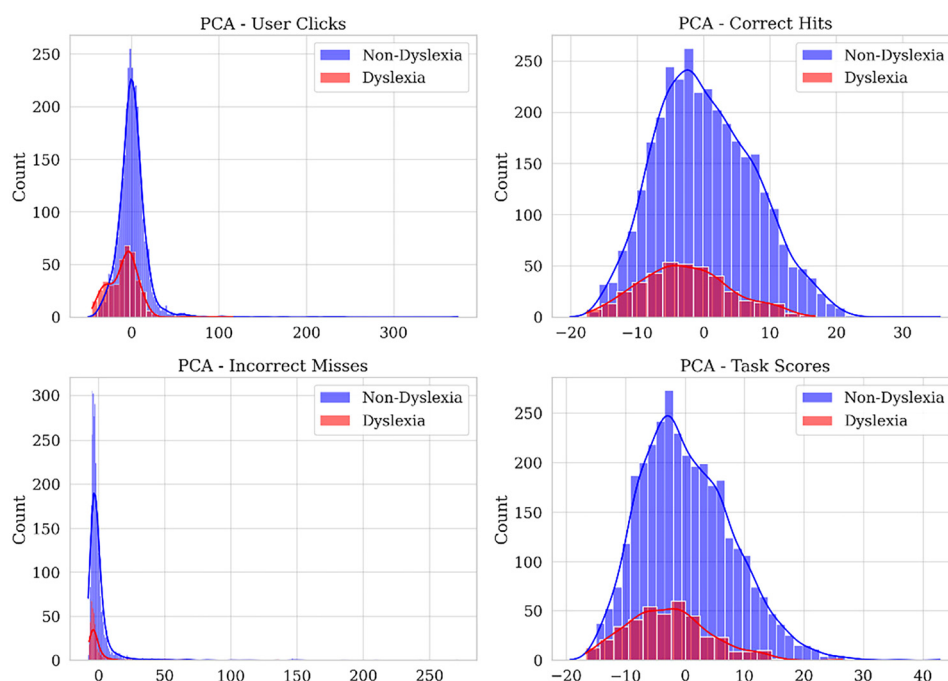


Figure 2. Distribution of the first principal component (PC1) for each behavioral feature group

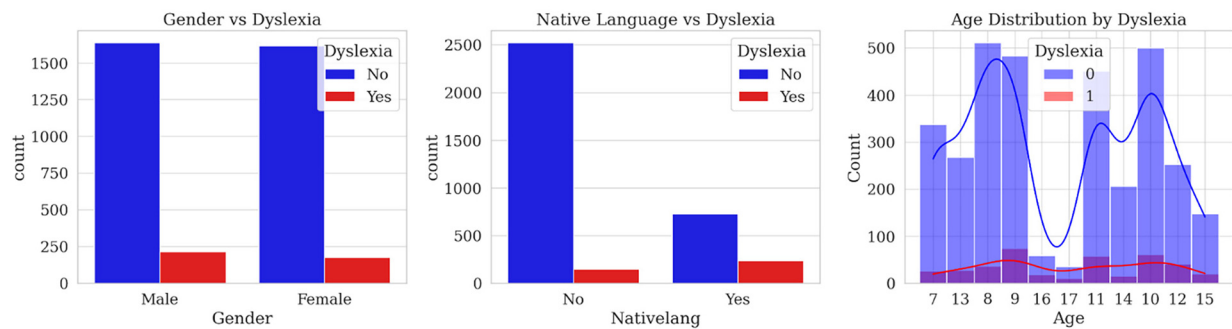


Figure 3. Demographic feature distributions by dyslexia status

10-fold stratified cross-validation experiments were conducted by selecting the top  $k$  features, where  $k \in \{5, 10, 15, \dots, all\}$ . For each subset, the F1-score was recorded.

Figure 4 (left) reveals that model performance improves as more features are added, peaking at around 165 features, beyond which performance slightly declines. When all features are included, the F1-score reverts to a level similar to that achieved with 165 features, suggesting that the top 165 mRMR-selected features are sufficient to design an effective classifier without introducing unnecessary complexity.

Figure 4 (right) compares the Area Under the Receiver Operating Characteristic (AUC-ROC) performance for models trained with all features versus the selected 165 features. Interestingly, both configurations achieved an AUC of 89%, indicating that the top 165 features sufficiently capture the essential patterns in the data, with both models demonstrating comparable sensitivity and specificity.

#### 4.2. Baseline Classifiers and Experimental Setup.

The proposed method was compared with two baseline classifiers—Logistic Regression (LR) and XGBoost—using stratified 10-fold cross-validation to evaluate its effectiveness. The results reveal a progressive improvement in F1-score from

83.19% (LR) to 89.77% (XGBoost). The proposed soft-voting ensemble outperforms all baselines, achieving an F1-score of 90.48%, which further improves to 90.51% when using the reduced set of 165 mRMR-selected features. This corresponds to a statistically significant 0.74% improvement over XGBoost (paired t-test p-value =  $9.3 \times 10^{-5}$ ), demonstrating the effectiveness of the proposed approach in achieving superior and balanced performance.

The computational efficiency of the proposed method was analyzed in comparison with the XGBoost baseline on a MacBook Pro M3 (16 GB RAM), evaluated over 10 cross-validation folds for training and 1,000 runs for inference. The proposed ensemble requires 16.97 seconds of training time compared to 1.21 seconds for XGBoost, making it approximately 14× slower during training. For inference, the ensemble requires a mean time of 40.33 ms (median: 34.43 ms), compared to 27.25 ms (median: 23.54 ms) for XGBoost, corresponding to approximately 1.48× higher inference cost. Despite this increase, the inference time remains within real-time constraints, demonstrating a fair accuracy–efficiency trade-off for practical deployment.

Table 1 summarizes the key hyperparameters used for each component of the proposed framework, along with the rationale for each setting.

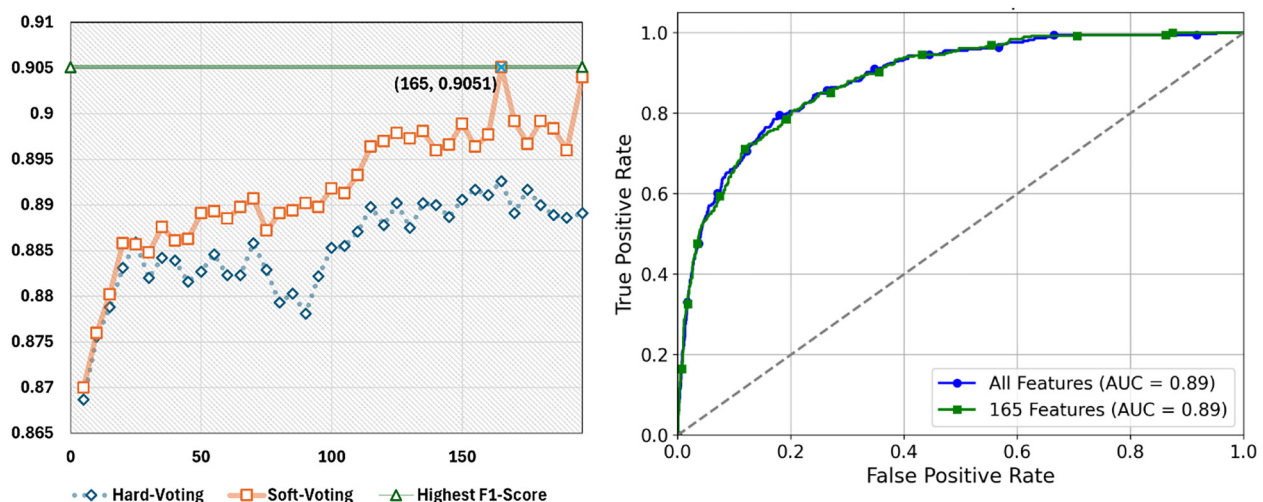


Figure 4. (left) F1-score variation with increasing number of mRMR-selected features; (right) Area Under the Receiver Operating Characteristic curves for models trained with all features versus top 165 mRMR features

**Table 1.** Hyperparameter configuration for each component of the proposed ensemble framework

| Model                          | Parameter        | Value + Rationale                                                                |
|--------------------------------|------------------|----------------------------------------------------------------------------------|
| XGBoost                        | n_estimators     | 100 (default) - sufficient for tabular data at this scale                        |
|                                | max_depth        | 6 (default) - controls tree depth to limit overfitting                           |
|                                | learning_rate    | 0.3 (default) - default step size for gradient updates                           |
|                                | scale_pos_weight | neg/pos $\approx$ 8.1 - set to majority/minority ratio to handle class imbalance |
| BaggingClassifier              | estimator        | DecisionTreeClassifier(class_weight="balanced") - balanced DT base learner       |
|                                | n_estimators     | 10 (default) - default number of bootstrap aggregations                          |
|                                | max_samples      | 1.0 (default) - each bootstrap uses 100% of training samples                     |
|                                | max_features     | 1.0 (default) - each bootstrap uses all features                                 |
| HistGradientBoostingClassifier | max_iter         | 100 (default) - default number of boosting rounds                                |
|                                | learning_rate    | 0.1 (default) - default shrinkage rate for gradient updates                      |
|                                | class_weight     | inverse-frequency dict - computed as total / (n_classes $\times$ class_count)    |
| CatBoostClassifier             | iterations       | 1000 (default) - default number of boosting iterations                           |
|                                | depth            | 6 (default) - default symmetric tree depth                                       |
|                                | learning_rate    | auto (default) - CatBoost auto-selects based on dataset size                     |
|                                | class_weights    | inverse-frequency list - computed as total / (n_classes $\times$ class_count)    |
| VotingClassifier               | voting           | soft - averages predicted probabilities across base learners                     |
|                                | weights          | [1, 1, 1, 1] - uniform weights, all base models contribute equally               |
| mRMR                           | method           | MIQ - Mutual Information Quotient, balances relevance and redundancy             |
|                                | n_features       | 165 - optimal from sweep over 5–195 features (peak weighted F1)                  |
| StratifiedKfold                | n_splits         | 10 - standard 10-fold, preserves class distribution per fold                     |
|                                | shuffle          | True - random shuffling before splitting                                         |

**4.3. Comparison with Current State-of-the-Art Methods.** Table 2 presents the evaluation using 10-fold cross-validation. Prior works including those by Luz Rello et al.<sup>11</sup>, Vanitha et al.<sup>19</sup> and Khan et al.<sup>21</sup> demonstrate limited balance between sensitivity and precision, with several classifiers yielding high variance across metrics. The proposed model consistently outperforms all baselines, achieving an F1-score of 0.909, with strong accuracy (0.916), precision (0.907), and recall (0.916), representing a substantial improvement relative to Rello et al. (e.g., F1 = 0.454), while also achieving balanced precision and recall, unlike several prior methods that exhibit significant metric imbalance. Importantly, many prior methods exhibit high accuracy but extremely low recall for the dyslexic class, indicating bias toward the majority class in this imbalanced dataset. In contrast, the proposed model achieves balanced precision and recall, resulting in a substantially higher F1-score. The consistency of results across cross-validation experiments suggests that this improvement is not due to overfitting.<sup>1</sup>

Table 3 evaluates the impact of class imbalance handling strategies on the proposed ensemble across accuracy, precision, recall, and F1-score. Among the five strategies evaluated—No Handling (Baseline), Class Weighting, Cost-Sensitive Learning, Synthetic Minority Over-sampling Technique (SMOTE),

and Adaptive Synthetic Sampling (ADASYN)—Class Weighting achieved the best overall performance. Cost-Sensitive Learning ranked second, closely followed by No Handling. Resampling-based methods—SMOTE and ADASYN—produced slightly lower results, suggesting that adjusting class weights directly is more effective than synthetic oversampling for this dataset. The superiority of Class Weighting may be attributed to the class imbalance (approximately 89:11), where re-weighting the loss function is sufficient to bias the model toward the minority class without introducing noise from synthetic samples. These results confirm Class Weighting as the optimal imbalance-handling strategy. The corresponding metrics are reported as the primary performance figures throughout this paper.

**4.4. Feature Importance Analysis using SHAP.** Figure 5 presents the SHAP summary bar plot, ranking the top 25 features by their mean absolute SHAP values. Interestingly, Age was the most influential feature, followed by Nativelang, and trial-specific interaction features such as Clicks24, Clicks28, and Missrate8. The prominence of behavioral features—particularly

<sup>1</sup> Re-implemented by the authors on the dyslexia dataset; Khan et al.<sup>16</sup> originally proposed DeepLSE for anticancer peptide classification. Results in Table 1 reflect this re-implementation.

**Table 2.** Comparison of the proposed ensemble model with existing state-of-the-art methods

| Year | Method                                 | Classifier(s) Used                         | Accuracy      | Precision     | Recall        | F1-Score      |
|------|----------------------------------------|--------------------------------------------|---------------|---------------|---------------|---------------|
| 2020 | Rello et al. <sup>11</sup>             | Random Forest                              | 0.791         | 0.316         | 0.806         | 0.454         |
| 2023 | G. Vanitha & M. Kasthuri <sup>19</sup> | C4.5                                       | 0.868         | 0.376         | 0.342         | 0.358         |
|      |                                        | Logistic Regression                        | 0.898         | 0.542         | 0.347         | 0.423         |
| 2024 | Khan et al. <sup>21</sup>              | DeepLSE                                    | 0.901         | 0.556         | 0.614         | 0.577         |
|      | <b>Proposed</b>                        | <b>Soft-Voting Ensemble (165 Features)</b> | <b>0.9155</b> | <b>0.9072</b> | <b>0.9155</b> | <b>0.9086</b> |

**Table 3.** Effect of class imbalance handling strategies on ensemble performance (10-fold cross-validation)

| Strategy                | Accuracy      | Precision     | Recall        | F1-Score      |
|-------------------------|---------------|---------------|---------------|---------------|
| No Handling (Baseline)  | 0.9103        | 0.8972        | 0.9103        | 0.8935        |
| Cost-Sensitive Learning | 0.9144        | 0.9046        | 0.9144        | 0.9078        |
| SMOTE                   | 0.9078        | 0.8939        | 0.9078        | 0.8961        |
| ADASYN                  | 0.9062        | 0.8903        | 0.9062        | 0.8927        |
| <b>Class Weighting</b>  | <b>0.9155</b> | <b>0.9072</b> | <b>0.9155</b> | <b>0.9086</b> |

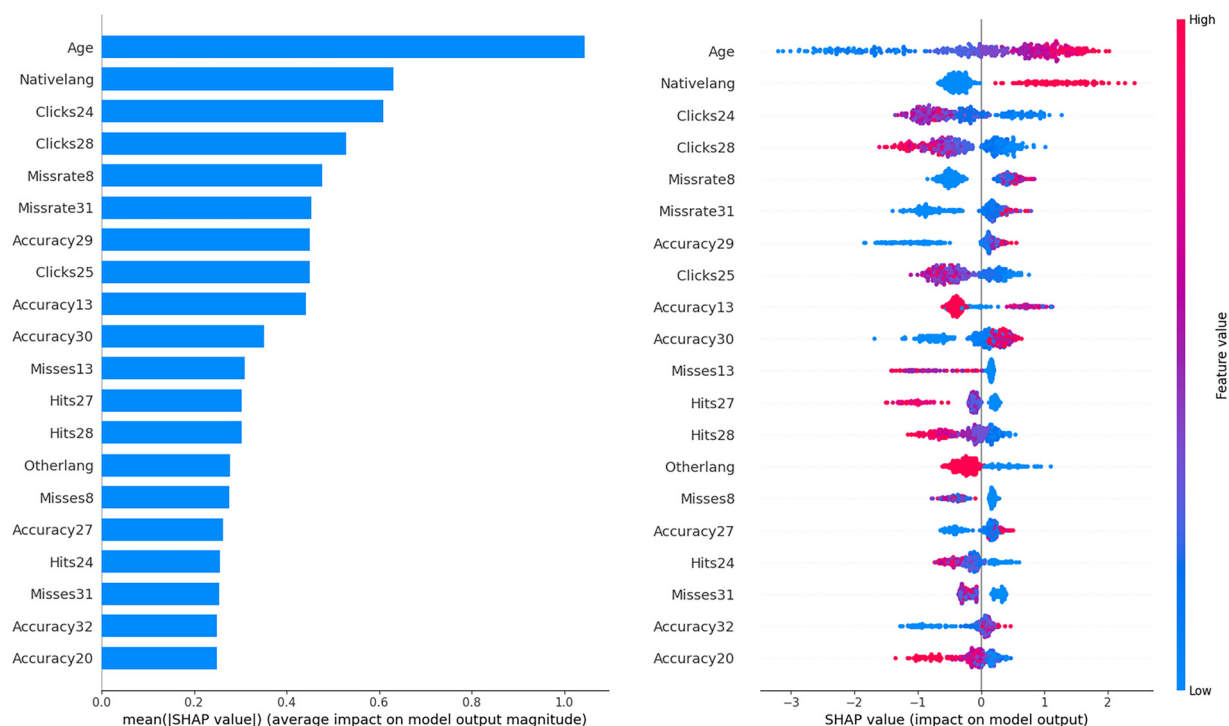
click counts, accuracy, and miss rates—underscores the critical role of temporal task engagement in the model’s decision-making process. These results indicate that the model relies heavily on trial-level dynamics and engagement patterns.

The beeswarm visualization in Figure 5 reveals that higher Age values (red points) tend to increase the likelihood of positive classification, suggesting that older individuals are more likely to be identified as dyslexic by the model. Similarly, the presence of Nativelang contributes positively to classification. Conversely, high miss rates in certain trials (e.g., Missrate8, Missrate31)

are associated with a negative contribution, reflecting error-prone behavior. Positive SHAP values for features like Clicks24, Clicks28, and Accuracy29 further underscore the importance of user interaction and task success in driving accurate predictions.

## 5. CONCLUSIONS

This paper presents a robust ensemble-based classification framework for early dyslexia detection. By integrating four boosting classifiers through a soft voting strategy, the proposed



**Figure 5.** SHAP-based interpretability analysis of the proposed ensemble model. (Left) Average magnitude of feature contributions across all test samples, where higher bars indicate greater overall impact; (right) each dot represents a single instance, with horizontal position indicating impact on model output and color representing feature value (red = high, blue = low)

method achieves strong, balanced performance across key metrics, including accuracy, precision, recall, and F1-score. The results from 10-fold cross-validation demonstrate its superiority over individual base learners and several recent state-of-the-art methods. These findings highlight the proposed model's potential as a reliable, scalable tool to support clinical and educational practitioners in early screening and intervention planning. A limitation of this study is its reliance on a single publicly available dataset, which may restrict the generalizability of the results. Future research should focus on validating the proposed framework on additional datasets and real-world educational settings to further assess its robustness and applicability.

## AFFILIATIONS AND AUTHOR DETAILS

### Undergraduate Author

**Almaan Khan** – Department of Information and Computer Science, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia;  0009-0002-6978-6040  
Email: s202335270@kfupm.edu.sa

### Corresponding Author

**Shujaat Khan** – Research Mentor, Assistant Professor, Department of Computer Engineering, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia;  0000-0001-9676-6817  
Email: shujaat.khan@kfupm.edu.sa

## ACKNOWLEDGEMENTS

Authors would like to acknowledge the support from the King Fahd University of Petroleum & Minerals (KFUPM) under Early Career Research Grant no. EC241027.1

## REFERENCES

- (1) Shaywitz, S. E. & Shaywitz, B. A. Dyslexia (Specific Reading Disability). *Biol. Psychiatry* **57**, 1301–1309 (2005).
- (2) AlGhamdi, A. S. Novel Ensemble Model Recommendation Approach for the Detection of Dyslexia. *Children* **9**, 1337 (2022).
- (3) Al-Fahoum, S., Alshraideh, S. & Bani-Salameh, A. A novel efficient model for dyslexia diagnosis using ML and feature fusion. *International Journal of Intelligent Engineering and Systems* **14**, 19–30 (2021).
- (4) A., J., S., J., H., A. & M.H., Y. Machine learning for dyslexia identification in children: A review. *Iranian Journal of Science and Technology, Transactions of Electrical Engineering* **47**, 1–18 (2023).
- (5) Alam, Md. K., Uddin, Md. F. & Haque, A. K. M. M. Ensemble-Based Classifiers for Detecting Dyslexia Using Behavioral Data. *Smart Health* **24**, 100247 (2022).
- (6) He, H. & Garcia, E. A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **21**, 1263–1284 (2009).
- (7) Gasmi, K. & others. Optimal Ensemble Learning Model for Dyslexia Prediction Based on an Adaptive Genetic Algorithm. *IEEE Access* **12**, 64754–64764 (2024).
- (8) Kaisar, S. & Chowdhury, A. Integrating oversampling and ensemble-based machine learning techniques for an imbalanced dataset in dyslexia screening tests. *ICT Express* **8**, 563–568 (2022).
- (9) Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)* 4765–4774 (2017).
- (10) Rauschenberger, M., Baeza-Yates, R. & Rello, L. Technologies for dyslexia. *Web Accessibility: A Foundation for Research* 603–627 (2019).
- (11) Rello, L., Baeza-Yates, R., Ali, A., Bigham, J. P. & Serra, M. Predicting risk of dyslexia with an online gamified test. *PLoS One* **15**, e0241687 (2020).
- (12) Alkhurayif, Y. & Sait, A. R. W. Deep learning-driven dyslexia detection model using multi-modality data. *PeerJ Comput. Sci.* **10**, e2077 (2024).
- (13) Aldehim, G., Rashid, M., Alluhaidan, A. S., Sakri, S. & Basheer, S. Deep learning for dyslexia detection: A comprehensive CNN approach with handwriting analysis and benchmark comparisons. *Journal of Disability Research* **3**, 20240010 (2024).
- (14) Benfatto, M. N. & others. Screening for dyslexia using eye tracking during reading. *PLoS One* **11**, e0165508 (2016).
- (15) Prabha, A. J. & Bhargavi, R. Prediction of dyslexia from eye movements using machine learning. *IETE J. Res.* **68**, 814–823 (2022).
- (16) Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**, 119–139 (1997).
- (17) Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
- (18) Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (2016). doi:10.1145/2939672.2939785.
- (19) Vanitha, G. & Kasthuri, M. Performance Analysis of Feature Selection and Classification Methods for Predicting Dyslexia. *Indian J. Sci. Technol.* **16**, 2663–2669 (2023).
- (20) Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: unbiased boosting with categorical features. in *Advances in Neural Information Processing Systems* (eds. Bengio, S. et al.) vol. 31 (Curran Associates, Inc., 2018).
- (21) Khan, S. Deep-Representation-Learning-Based Classification Strategy for Anticancer Peptides. *Mathematics* **12**, (2024).