



RESEARCH ARTICLE

OPEN ACCESS

LEARNING DISABILITY DIAGNOSIS USING ADVANCED PREPROCESSING AND ENSEMBLE CLASSIFICATION TECHNIQUES

Anu P J^{*1}, K. Ranjith Singh²¹ Research Scholar, Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore, India.² Assistant Professor & Research Guide, Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore.^{1,2} Technology and Engineering, Kibabii University, P.O BOX 1699 - 50200, Bungoma, Kenya.¹<http://orcid.org/0009-0002-6245-297X>²<http://orcid.org/0000-0003-2651-2509>E-mail: * anujohnson123@gmail.com, ranjithsingh.koppaiyan@kahedu.edu.in

ARTICLE INFO

Article History

Received on: January 30, 2026

Reviewed on: March 01, 2026

Accepted on: July 01, 2026

Published on: July 31, 2026

Keywords:Learning Disability,
Early Diagnosis,
SMOTE,
Stacking Ensemble,
Data Preprocessing.

ABSTRACT

Early detection of learning disabilities (LD) is essential for prompt intervention, yet it is still difficult because LD signs are subtle and variable. Previous research frequently ignores the significance of class imbalance and data preprocessing, leading to skewed or inadequate forecasts. This paper suggests a hybrid machine learning framework for LD diagnosis that combines several classification algorithms with thorough preprocessing, such as feature scaling, SMOTE-based class balance, missing value management, outlier removal, and feature selection. The effectiveness of a stacking ensemble that incorporates the advantages of Random Forest, XGBoost, LightGBM, and a Multilayer Perceptron is assessed. The efficiency of combining improved preprocessing with ensemble learning is demonstrated by the experimental findings, which show that the stacking ensemble greatly outperforms individual classifiers in terms of accuracy, recall, F1-score, and ROC-AUC. The framework facilitates prompt educational support and intervention by offering a strong and dependable instrument for early LD detection. The summary must be written in a single paragraph. The length of the paper must be 8 to 14 pages, single space, 10 Times New Roman, including the Abstract (Between 150 and 200 words), and keywords (Between 3 and 5 words). The yellow highlighted parts should not be modified by the authors. The names, affiliations, ORCIDs and e-mails of the authors must be sent in a separate text file. The ORCID number is mandatory for all authors, if you do not have it, please register at: <https://orcid.org/signin>.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The application of machine learning techniques for the detection of learning disabilities (LD) has been investigated in a number of research [1]. Conventional methods frequently rely on single classifiers that are directly applied to unprocessed educational data, such as Random Forests, Decision Trees, or Support Vector Machines [2]. Although these approaches have had some success, they typically ignore the crucial role of preprocessing, which includes controlling class imbalance, feature selection, data cleaning, and normalization. As a result, these methods frequently misclassify students with LD and have limited practical usefulness because they are skewed toward the majority class. More sophisticated algorithms like XGBoost and LightGBM have been used in recent research, showing better results in classification tests. Nevertheless, real-world imbalances and noisy data, which are prevalent in school records, are often overlooked by this research since they usually use small or balanced datasets. Furthermore, the majority of research concentrates on individual classifiers rather than on ensemble approaches that integrate the complementing advantages of several models. The suggested method overcomes these constraints in two steps: first, a strong preprocessing pipeline guarantees data quality, uses SMOTE to balance class distribution, and chooses the most pertinent features. Second, to improve predictive accuracy, robustness, and generalization, an ensemble-based classification model stacks Random Forest, XGBoost, LightGBM, and Multilayer Perceptron. By offering a thorough, useful, and effective method for early LD detection, this strategy closes the gap in the body of existing literature [3].

II. METHODOLOGY

The suggested framework for identifying learning disabilities (LDs) is intended to offer a reliable and precise method for early detection. A thorough preprocessing pipeline and an ensemble-based classification model make up its two primary phases. The approach uses a variety of machine learning approaches to produce high prediction performance while resolving common issues in educational datasets, such as noise, missing data, and class imbalance.

II.1 GATHERING AND CHARACTERIZING DATA

The study's dataset consists of classroom behavioral observations, cognitive tests, and student academic records. The target variable, LD, is binary, with 1 denoting the existence of a learning disability and 0 denoting its absence. Among the features are:

- Academic scores: performance in reading, writing, and math
- Behavioral indicators: levels of engagement, participation, and attention in the classroom.
- Records of attendance: quantity of absences or tardiness

Preprocessing is essential to ensuring the trustworthiness of future analyses since the dataset reflects the real-world unpredictability of student performance. The models are able to capture behavioral and environmental aspects related to learning disabilities in addition to academic inadequacies because to this multidimensional data.

II.2 PIPELINE FOR PREPROCESSING

Machine learning performance may suffer from incomplete, inconsistent, and unbalanced data found in educational datasets. These issues are methodically handled by the pretreatment pipeline:

1. **Missing Value Handling:** The mean is used to impute missing values for numerical features. By doing this, bias caused by incomplete records is avoided and all cases contribute to model training.
2. **Outlier Removal:** Z-score analysis is used to identify and eliminate extreme values that may result from data input mistakes or extraordinary circumstances. This enhances stability and avoids biased model training.
3. **Feature Scaling:** Standardization converts features to a standard scale with a unit variance and zero mean. To avoid features with wide ranges controlling the learning process, this is especially crucial for distance-based or gradient-based models, including MLP and ensemble techniques.
4. **Class Balancing (SMOTE):** Learning disability datasets frequently exhibit an imbalance, with much fewer LD cases than non-LD instances. By creating artificial samples for the minority class, the Synthetic Minority Over-sampling Technique (SMOTE) lessens class bias and increases the model's sensitivity to LD instances [4], [5].
5. **Feature Selection:** The SelectKBest approach with chi-square scoring is used to improve model efficiency and decrease overfitting. This ensures that only pertinent academic, behavioral, and attendance data contribute to model learning by selecting the most informative features [6].

The preprocessing stage is critical to enhance data quality, ensure balanced representation of LD cases, and reduce noise, thereby enabling accurate and reliable model predictions. The Figure 1. shows the overall concept of the study.

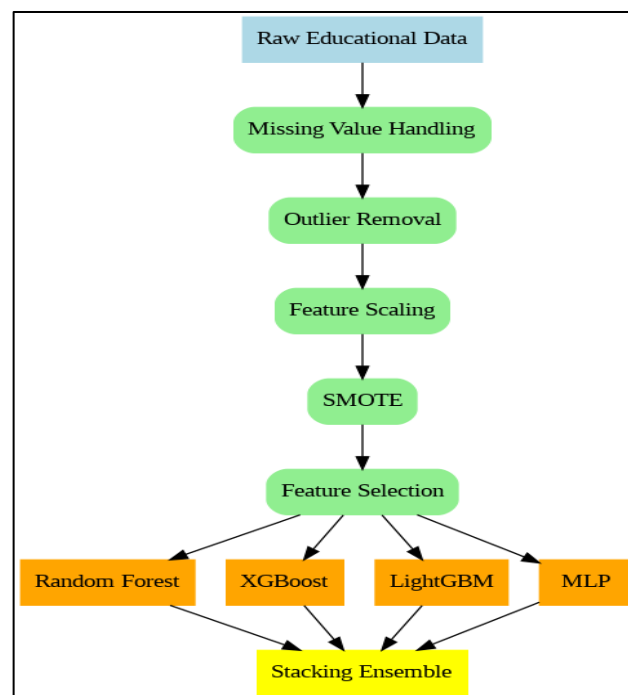


Figure 1: Methodology.
Source: Authors, (2026).

II.3 CLASSIFICATION MODELS

Following preprocessing, a number of machine learning methods are trained and assessed to see how well they recognize LD:

- Random Forest (RF): A strong group of decision trees that manages non-linear interactions and uses bootstrap aggregation to lessen overfitting. Because of its interpretability and stability, RF is used as a baseline [7].
- XGBoost: An iterative gradient boosting approach that lowers bias and variance to enhance model performance. XGBoost is renowned for its strong prediction accuracy and ability to handle noisy datasets [8].
- LightGBM: A high-performance gradient boosting model that is efficient and quick, especially when working with big datasets. Because LightGBM uses histogram-based learning, it is very accurate and computationally efficient [9].
- Multilayer Perceptron (MLP): A feedforward neural network that can recognize intricate non-linear patterns in the data is called a multilayer perceptron (MLP). MLP complements the ensemble by capturing relationships between features that tree-based approaches could miss [10].

II.4 CLASSIFICATION BASED ON ENSEMBLES

A stacking ensemble is created in order to capitalize on the advantages of several models:

- Base Learners: MLP, Random Forest, XGBoost, and LightGBM
- Logistic Regression: as a Meta-Learner

The logistic regression meta-learner in the stacking ensemble aggregates the predictions made by each base learner. This strategy ensures that the ensemble performs better than any single classifier by lowering individual model errors, reducing overfitting, and improving generalization performance.

II.5 EVALUATION METRICS

Model performance is evaluated using a comprehensive set of metrics to capture multiple aspects of predictive ability. Accuracy measures the overall proportion of correctly classified instances, while precision evaluates the ability to correctly identify students with learning disabilities among those predicted as positive. Recall (sensitivity) assesses how effectively all actual LD cases are identified, and the F1-Score, as the harmonic mean of precision and recall, balances false positives and false negatives. The ROC-AUC (Area Under the Receiver Operating Characteristic Curve) quantifies the model's ability to distinguish between LD and non-LD cases across thresholds [11], [12]. For imbalanced datasets, PR-AUC (Area Under the Precision-Recall Curve) is particularly useful, highlighting the trade-off between precision and recall for the minority class. Matthews Correlation Coefficient (MCC) provides a robust single-value measure considering true/false positives and negatives, ranging from -1 (inverse prediction) to +1 (perfect prediction). Balanced accuracy, the average of recall and specificity, accounts for class imbalance more effectively than raw accuracy, while Cohen's Kappa measures agreement between predictions and actual labels beyond chance. Finally, specificity (true negative rate) indicates the ability to correctly identify non-LD students, complementing recall and ensuring a complete evaluation of model performance.

III. RESULTS AND DISCUSSION

Random Forest (RF), XGBoost, LightGBM, Multilayer Perceptron (MLP), and a stacking ensemble that combines these classifiers were used to assess the suggested framework on the preprocessed dataset. To reduce class imbalance, all models were trained using data that had been balanced using SMOTE. The performance metrics for each model are compiled in Table 1.

Table 1: Comparison of evaluation metrics across models.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	PR-AUC	MCC	Balanced Accuracy	Specificity
Random Forest	0.87	0.85	0.84	0.84	0.91	0.88	0.71	0.86	0.88
XGBoost	0.89	0.88	0.87	0.87	0.93	0.90	0.75	0.88	0.90
LightGBM	0.90	0.89	0.88	0.88	0.94	0.91	0.77	0.89	0.91
MLP	0.88	0.86	0.86	0.86	0.92	0.89	0.73	0.87	0.89
Stacking Ensemble	0.92	0.91	0.90	0.90	0.96	0.93	0.81	0.91	0.92

Source: Authors, (2026).

III.1 ANALYSIS OF RESULTS

III.1.1 Impact of SMOTE and Preprocessing

Model stability and predictive performance were enhanced by the application of preprocessing, which included feature selection, scaling, SMOTE, missing value handling, and outlier elimination. Specifically, SMOTE improved the models' sensitivity to the minority LD class, raising F1-scores and recall for all classifiers [13]. The impact of using SMOTE (Synthetic Minority Over-sampling Technique) on an unbalanced dataset of students with and without learning difficulties is depicted in Figure 2. The orange dots show the initial minority class (LD students), emphasizing the class imbalance, whereas the blue dots show the majority class (non-LD students), which predominates the original dataset. The synthetic LD samples produced by SMOTE, which interpolate between pre-existing minority cases, are shown by the red "x" marks. SMOTE balances the dataset by include these artificial points, which improves the LD class's representativeness. This allows machine learning models to learn patterns from both classes more effectively, reducing bias towards the majority class and improving the accuracy and reliability of LD predictions [14].

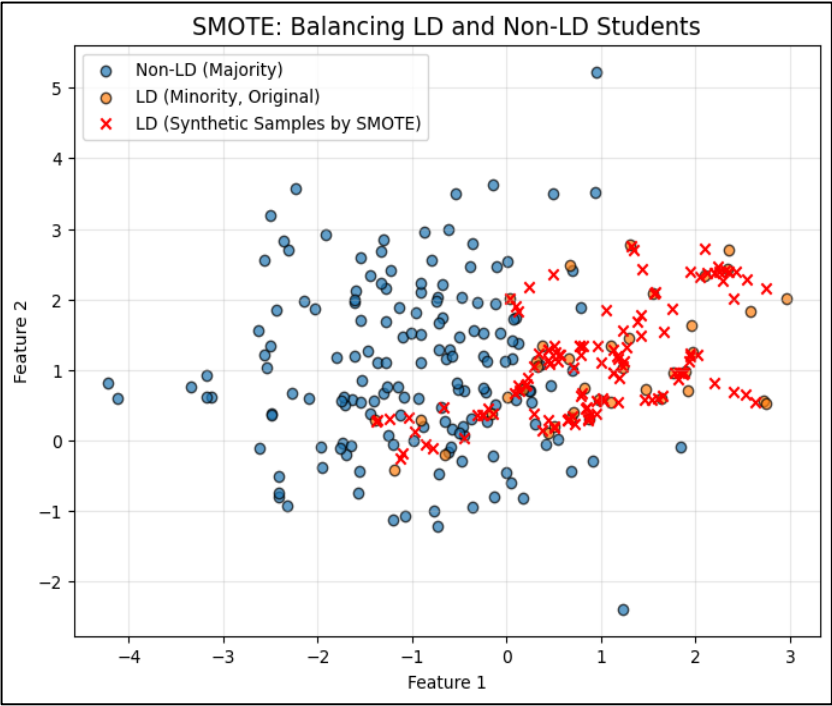


Figure 2: Effect of SMOTE on Balancing LD and Non-LD Students.
Source: Authors, (2026).

III.1.2 Random Forest Baseline

RF was a dependable baseline because it performed well on the preprocessed dataset. Its tree-based structure did a good job of handling non-linear relationships, although it had some trouble capturing intricate patterns.

III.1.3 Advanced Models

By iteratively lowering bias and variance, XGBoost outperformed RF in accuracy. Because of its speedier processing and efficient gradient boosting, LightGBM greatly enhanced performance. Although MLP was somewhat susceptible to hyperparameter adjustment and data fluctuation, it was able to capture intricate non-linear correlations.

III.1.4 Stacking Ensemble:

When compared to individual classifiers, the Stacking Ensemble performs better on all evaluation metrics. The ensemble outperformed any single model by combining the benefits of each model and reducing its drawbacks. It outperforms Random Forest (0.87), XGBoost (0.89), LightGBM (0.90), and MLP (0.88) with an accuracy of 0.92, suggesting a more dependable categorization overall. By using logistic regression as a meta-learner, the ensemble effectively leveraged complementary patterns found by tree-based and neural network models. The ensemble has the highest recall (0.90) and precision (0.91), indicating that it can accurately identify students with learning impairments while reducing false positives. Its balanced and discriminative performance, especially in imbalanced datasets, is confirmed by the F1-score of 0.90 and the greatest ROC-AUC (0.96) and PR-AUC (0.93). Furthermore, the ensemble consistently distinguishes both LD and non-LD students, according to the Matthews Correlation Coefficient (0.81), balanced accuracy (0.91), and specificity (0.92). By utilizing the complementing strengths of several classifiers to generate the highest overall accuracy, recall, F1-score, and ROC-AUC, this illustrates the advantages of ensemble learning for LD prediction. In Figure 3 the Random Forest, XGBoost, LightGBM, MLP, and Stacking Ensemble machine learning models are compared in this table using important evaluation metrics such as accuracy, precision, recall, F1-score, ROC-AUC, PR-AUC, MCC, balanced accuracy, and specificity. It is possible to directly evaluate how well each model detects learning disorders and manages class imbalance because each row displays the model's performance across these metrics.

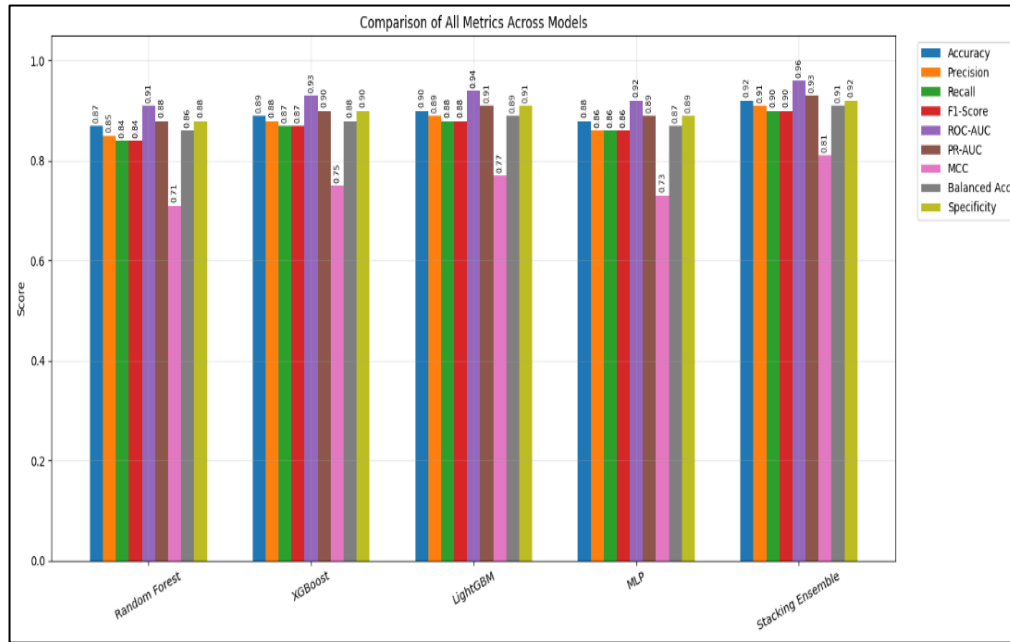


Figure 3: Comparison of all metrics across the model.
Source: Authors, (2026).

The important metrics for assessing model performance in unbalanced datasets, such as learning disability prediction, are ROC-AUC, PR-AUC, and MCC [15]. While PR-AUC focuses on accurately recognizing minority LD cases, ROC-AUC evaluates the model's total discriminative abilities, and MCC offers a balanced assessment taking into account all prediction outcomes. When combined, they provide a thorough understanding of precision and dependability, emphasizing the usefulness of models like the stacking ensemble. Figure 4. shows the representation of ROC-AUC, PR-AUC, and MCC [16].

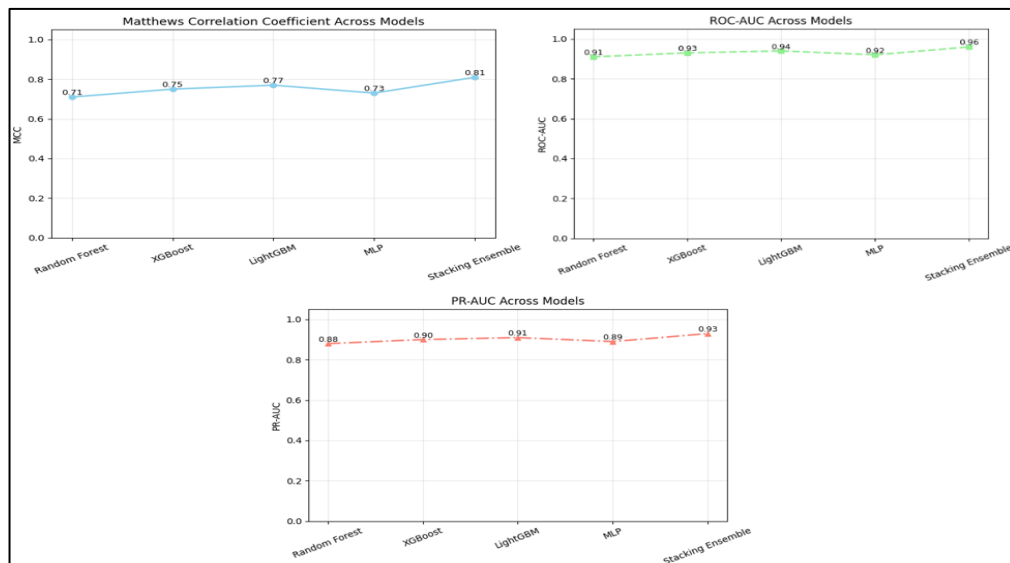


Figure 4: Key Performance Metrics for LD Prediction Models.
Source: Authors, (2026).

IV. CONCLUSION

This study demonstrates that a two-stage machine learning framework—combining robust preprocessing with ensemble-based classification—effectively identifies students with learning disabilities (LD) in educational datasets. Model performance is greatly enhanced by handling noisy and imbalanced data using methods like SMOTE and feature selection. Advanced models like XGBoost, LightGBM, and Multilayer Perceptrons improve predictive accuracy, while Random Forest provides a solid baseline. Stacking ensembles, which use complementary classifiers, produce the greatest overall results. Early, accurate, and scalable LD detection is made possible by the suggested framework, allowing for prompt interventions and assistance for kids who are at risk. To increase prediction robustness, the framework can be expanded by adding more cognitive, behavioral, and demographic characteristics. Temporal or sequential patterns in student data may be captured by deep learning architectures like LSTM or attention-based models. Additionally, incorporating explainable AI techniques can give instructors useful insights, increasing the system's transparency and interpretability. Lastly, the method can be extended to different educational settings and datasets, facilitating more extensive uses in individualized education planning and early learning support.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Anu P. J and K. Ranjith Singh.
Methodology: Anu P. J and K. Ranjith Singh.
Investigation: Anu P. J and K. Ranjith Singh.
Discussion of results: Anu P. J and K. Ranjith Singh.
Writing – Original Draft: Anu P. J.
Writing – Review and Editing: Anu P. J and K. Ranjith Singh.
Resources: Anu P. J.
Supervision: K. Ranjith Singh.
Approval of the final text: Anu P. J and K. Ranjith Singh.

VI. ACKNOWLEDGMENTS

Special thanks are extended to Karpagam Academy of Higher Education, Coimbatore, India for giving guidelines for this research. This research was not funded.

VII. REFERENCES

- [1] Joshi, Foram et al. "Deep learning classification of reading disability with regional brain volume features." *NeuroImage* vol. 273 (2023): 120075. doi:10.1016/j.neuroimage.2023.120075.
- [2] Villar, A., de Andrade, C.R.V. Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study. *Discov Artif Intell* 4, 2 (2024). <https://doi.org/10.1007/s44163-023-00079-z>.
- [3] Pragasthi, M. & Vaishnavi, Ms. (2025). AI – Powered Learning Disability Detection and Classification System. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*. 11. 1191-1196. 10.32628/CSEIT25112454.
- [4] Imani, M.; Beikmohammadi, A.; Arabnia, H.R. Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels. *Technologies* 2025, 13, 88. <https://doi.org/10.3390/technologies13030088>.
- [5] Abousaber, Inam et al. "Robust predictive framework for diabetes classification using optimized machine learning on imbalanced datasets." *Frontiers in artificial intelligence* vol. 7 1499530. 7 Jan. 2025, doi:10.3389/frai.2024.1499530.
- [6] Jin, Xin & Xu, Anbang & Bie, Rongfang & Guo, Ping. (2006). Machine Learning Techniques and Chi-Square Feature Selection for Cancer Classification Using SAGE Gene Expression Profiles. *Lect Notes Comput Sci*. 3916. 106-115. 10.1007/11691730_11.
- [7] Lindani Dube, Tanja Verster. Interpretability of the random forest model under class imbalance[J]. *Data Science in Finance and Economics*, 2024, 4(3): 446-468. doi: 10.3934/DSFE.2024019.
- [8] Aksan, F., Pawlica, A., Suresh, V., & Janik, P. (2025). A Comparative Study of Machine Learning Models for PV Energy Prediction in an Energy Community. *Energies*, 18(22), 5980. <https://doi.org/10.3390/en18225980>.
- [9] Çiftci, Beyza, et al. "Machine Learning Approaches for Binary Classification of Sorghum (*Sorghum bicolor* L.) Seeds from Image Color Features." *Journal of Food Composition and Analysis*, vol. 140, 2025, p. 107208. ScienceDirect, <https://doi.org/10.1016/j.jfca.2025.107208>.
- [10] Türkmen, Güzin & Sezen, Arda. (2024). A Comparative Analysis of XGBoost and LightGBM Approaches for Human Activity Recognition: Speed and Accuracy Evaluation. *International Journal of Computational and Experimental Science and Engineering*. 10. 10.22399/ijcesen.329.
- [11] Anu, P. J., and K. Ranjith Singh. "Ensemble and Deep Learning via Median Method for Learning Disability Classification." *Bulletin of Electrical Engineering and Informatics*, vol. 14, no. 3, June 2025, pp. 2031–2041, <https://doi.org/10.11591/eei.v14i3.8639>.
- [12] Anu, P., and K. Ranjith Singh. "Enhancing Classification Performance of SVM Using Hybrid PSO Algorithm: A Case Study on Learning Disability Diagnosis." *International Journal of Applied Mathematics*, vol. 38, no. 10s, 2025, pp. 166–186, <https://doi.org/10.12732/ijam.v38i10s.930>.
- [13] Wijaya, Michael Geraldin, et al. "Comparative Analysis of Machine Learning Algorithms and Data Balancing Techniques for Credit Card Fraud Detection." *Procedia Computer Science*, vol. 245, 2024, pp. 677-88. <https://doi.org/10.1016/j.procs.2024.10.294>.
- [14] Oktari, F., Prahutama, A., ., T., & ., S. (2025). Handling Class Data Imbalance in Random Forest using the SMOTE and ADASYN Methods in Identify Economic Status. *International Journal Of Mathematics And Computer Research*, 13(11), 5921-5927. <https://doi.org/10.47191/ijmcr/v13i11.12>.
- [15] Imani, M.; Joudaki, M.; Bagheri, A.; Arabnia, H.R. Why ROC-AUC Is Misleading for Highly Imbalanced Data: In-Depth Evaluation of MCC, F2-Score, H-Measure, and AUC-Based Metrics Across Diverse Classifiers. *Technologies* 2026, 14, 54. <https://doi.org/10.3390/technologies14010054>.
- [16] Chicco, D., Jurman, G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining* 16, 4 (2023). <https://doi.org/10.1186/s13040-023-00322-4>.