Asian Journal of Engineering and Applied Technology
ISSN: 2249-068X (P); 3049-1959 (E)
Vol.14 No.2, 2025, pp.51-57

www.crijournals.org
DOI: https://doi.org/10.70112/ajeat-2025.14.2.4330

Bias-Aware Machine Learning for Student Dropout Prediction: Balancing Accuracy and Fairness

Olufunke Catherine Olayemi[1], Olayemi Oladimeji Olasehinde[2]* and Olugbenga Olawale Akinade[3]
[1&3]Department of Computer Science, Teesside University, Middlesborough, United Kingdom
[2*]Department of Computer Science, University of Huddersfield, Huddersfield, United Kingdom
E-mail: Olafunyemi2016@gmail.com, o.akinade@tees.ac.uk
*Corresponding Author: o.olasehinde@hud.ac.uk
(Received 5 October 2025; Revised 10 November 2025; Accepted 2 December 2025; Available online 10 December 2025)

Abstract **- Student dropout remains a persistent global challenge with serious social and economic consequences. Early identification of at-risk learners enables timely support, which can improve retention while promoting fairness in educational outcomes. This study presents a bias-aware machine learning framework for student dropout prediction that jointly evaluates predictive performance and fairness across demographic subgroups. Six machine learning models are benchmarked using academic, demographic, and socioeconomic features. Model performance is assessed using Accuracy, F1-score, Precision, and Matthews Correlation Coefficient, while fairness is evaluated across gender, marital status, and displacement groups. Initial results show that CatBoost achieves the strongest overall performance before class balancing; however, subgroup analysis reveals systematic disparities affecting vulnerable populations. To address these biases, the Synthetic Minority Oversampling Technique is applied. After rebalancing, XGBoost delivers the best performance, achieving substantial improvements in predictive accuracy alongside marked reductions in subgroup disparities. In particular, dropout detection for displaced students improves significantly, narrowing fairness gaps across all evaluated groups. The findings demonstrate that data-level bias mitigation can enhance both accuracy and equity in educational predictive systems. This work provides empirical evidence that fairness-aware machine learning can support more reliable and inclusive early warning systems for student retention.**
Keywords: **Student Dropout, Machine Learning, Fairness, Bias Mitigation, Early Warning Systems**

I. INTRODUCTION

Education is a foundation for personal growth and societal advancement, yet student dropout remains a pressing global challenge. High dropout rates restrict students' future opportunities and create social and economic burdens for institutions and communities. Addressing this issue requires early identification of at-risk learners so that timely interventions, such as tutoring, counseling, or financial aid, can improve retention and support fair access to education. In recent years, machine learning has shown strong potential for addressing dropout. By analyzing academic performance, demographic factors, and behavioral patterns, models can predict which students are most likely to disengage, enabling targeted, data-driven interventions. These systems can also support fair grading, performance forecasting, and continuous feedback, contributing to improved retention [1]. However, machine learning in education faces challenges, particularly bias. Predictive models can disadvantage certain student groups when trained on imbalanced or incomplete data. This can lead to unfair outcomes, such as marginalized students being misidentified as low risk and missing support, or others being misclassified as at risk. Such errors weaken both system reliability and fairness in education. This study addresses these challenges by examining dropout prediction through a fairness-aware lens.

It evaluates multiple machine learning models, identifies subgroup disparities, and applies a data-level mitigation strategy to improve equity. The central objective is to demonstrate that predictive accuracy and fairness are not competing goals but complementary requirements for responsible educational analytics. The results provide actionable evidence for deploying equitable early warning systems in higher education. This work makes three main contributions:

1. A comprehensive benchmark of six machine learning models for student dropout prediction using balanced and imbalanced data.
2. A detailed fairness analysis across gender, marital status, and displacement subgroups using fairness-aware evaluation metrics.
3. Empirical evidence that SMOTE improves both predictive performance and subgroup equity in educational datasets.

II. LITERATURE REVIEW

Machine learning has become an important tool in education, helping to predict student performance, improve grading systems, and reduce dropout rates. Kučak *et al.* [1] showed that predictive models can support early warning systems, provide continuous feedback, and help schools identify students who need extra support to stay in school. However, one major challenge is fairness. Costa-Mendes *et al.* [2] discovered that many grade prediction models for Portuguese students contained bias. This bias arose from missing or unbalanced data, which caused unfair predictions for some groups of students. They introduced the idea of

knowledge bias and suggested that fair and precise education systems require more complete and representative datasets. Recent research has focused on ways to make learning models fairer. Raftopoulos *et al.* [3] demonstrated that adjusting the training data through methods such as resampling and reweighting can improve fairness between different student groups without significantly affecting accuracy. Pham *et al.* [4] developed FAIREDU, a fairness-based regression model that helps reduce bias in educational data while maintaining high prediction performance.

Another approach, called adversarial debiasing and introduced by Zhang *et al.* [5], teaches the model to ignore sensitive information such as gender or ethnicity while still learning useful patterns from data. Similarly, reweighting methods [6] assign more importance to underrepresented or disadvantaged groups during training so that their outcomes are treated more equally. Causal fairness is another growing area of research. Kusner *et al.* [7] introduced the concept of counterfactual fairness, which checks whether a model's prediction would remain the same if a person's sensitive attribute, such as gender or social class, were different. This approach helps researchers understand whether a model's decisions are truly fair.

Madras *et al.* [8] further applied this method to education, showing how causal reasoning can help detect hidden bias in dropout prediction systems. From these studies, two main lessons stand out. First, machine learning can help reduce school dropout rates by identifying at-risk students early. Second, without fairness-aware design, these models can unintentionally reinforce existing inequalities. To achieve both fairness and accuracy, a combination of data-level techniques such as SMOTE and reweighting, together with fairness frameworks like FAIREDU, adversarial debiasing, and causal reasoning, is essential. These strategies move education toward more transparent, inclusive, and socially responsible use of artificial intelligence.

III. METHODOLOGY

A. Dataset

The dataset used in this study was sourced from Realinho *et al.* [9] and published on Kaggle. It contains information on undergraduate students, including demographic, socioeconomic, and academic attributes relevant to predicting outcomes. The dataset comprises 35 columns (34 features and one target variable). The target has three classes: dropout, enrolled, and graduate. As shown in Figure 1, the distribution is imbalanced, with the majority in the graduate class, followed by dropout, and the smallest proportion in the enrolled group.

B. Model Development

Six machine learning models-Logistic Regression, Decision Tree, Gradient Boosting, Random Forest, XGBoost, and CatBoost-were trained and evaluated to develop a reliable and fair predictive framework. The training process employed 10-fold cross-validation [10] to ensure model robustness and generalization capability. A standardized preprocessing pipeline incorporating StandardScaler was implemented to normalize feature distributions and maintain consistency across models. The model exhibiting the best performance was retrained on the full dataset using stratified sampling to preserve the proportional representation of each class. This procedure minimized sampling bias and enhanced the reliability of the final model. Performance evaluation was conducted at both the global level and across demographic subgroups defined by gender, marital status, and displacement status, allowing a comprehensive assessment of predictive accuracy and fairness across different population segments.

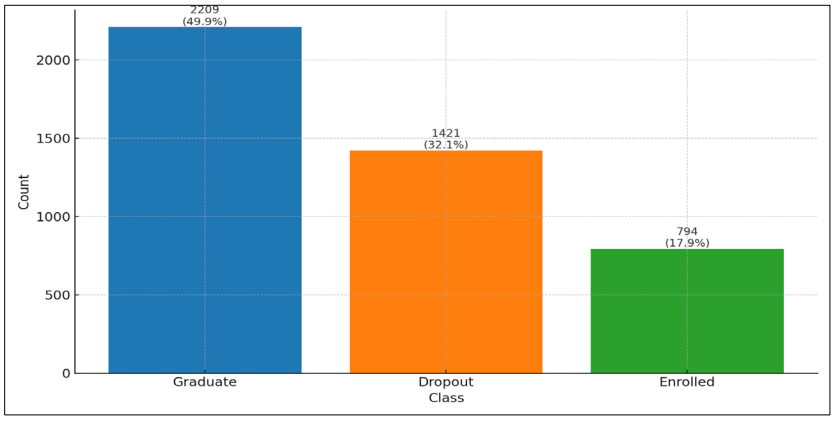

Fig.1 Class Distribution Before SMOTE

C. Bias Management with SMOTE

Given the class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) [11] was applied to generate synthetic samples for the underrepresented classes (dropout and enrolled). Unlike simple duplication, SMOTE synthesizes new examples, producing a more balanced dataset (Figure 2). After resampling, models were retrained and re-evaluated using the same performance and fairness metrics as in the pre-SMOTE stage.

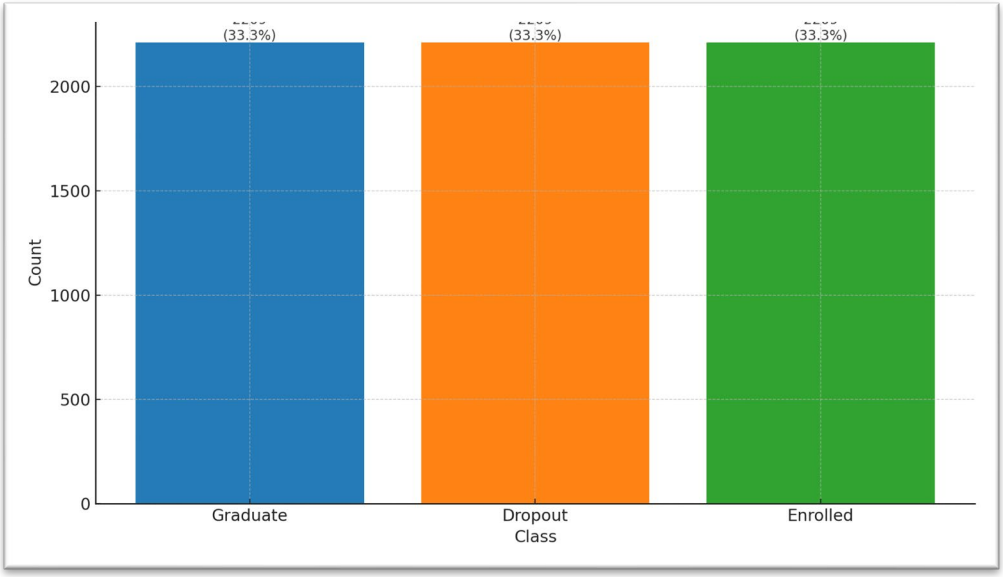

Fig.2 Class Distribution After SMOTE

D. Evaluation Metrics

Model performance in this study was assessed using a combination of core classification metrics and fairness-aware evaluations. This dual approach ensured that the models were not only accurate but also equitable in their predictions across different student groups. The core performance metrics included:

1. F1-score, used as the primary criterion, representing the harmonic mean of precision and recall [12].
2. Accuracy, which measures overall correctness, though it is sensitive to class imbalance [12].
3. Precision, capturing tendencies toward over- or under-labelling cases.
4. Matthews Correlation Coefficient (MCC), a robust and balanced evaluation metric suitable for imbalanced data [13], [14].

Fairness was evaluated through three complementary methods:

1. Per-class distribution checks across the three outcomes (dropout, enrolled, graduate).
2. Subgroup analyses based on gender, marital status, and displacement status.
3. Comparisons of pre- and post-SMOTE performance to measure improvements in equity.

The fairness metrics adopted in this study are primarily group-fairness oriented rather than individual-fairness based. Group fairness focuses on ensuring that model predictions and error rates are equitable across socially or demographically defined subgroups (for example, gender, ethnicity, or socioeconomic status). Metrics such as Demographic Parity Difference (ΔDP) and Equal Opportunity Difference (ΔEO) evaluate disparities in model outputs and true-positive rates between these groups [15], while the Matthews Correlation Coefficient (MCC) provides a balanced global performance measure that remains robust under class imbalance [13]. This orientation toward group fairness is especially relevant in education and healthcare contexts, where the aim is to promote equitable treatment across groups rather than identical predictions for every individual [16]. By comparing mean outcome rates across protected categories, group-based metrics uncover systemic disparities that aggregate metrics like accuracy or F1-score may conceal. The integration of performance and fairness metrics ensured a comprehensive evaluation framework. While F1-score and MCC provided reliable insights into predictive strength, they alone could not capture disparities across demographic groups. The fairness-focused analyses, particularly subgroup evaluations and SMOTE comparisons, were critical in revealing hidden inequities and demonstrating how bias-mitigation strategies could lead to both improved accuracy and greater equity in dropout prediction.

IV. RESULTS AND DISCUSSION

A. Model Performance Before SMOTE

Initial cross-validation across six machine learning models showed that CatBoost (CB) delivered the best overall performance, achieving an F1-score of 71.21% (Table I). Other ensemble models, such as Gradient Boosting (GB), Random Forest (RF), and XGBoost (XGB), also performed competitively, while Logistic Regression (LR) and Decision Tree (DT) lagged behind. When tested on the hold-out set, CatBoost achieved an accuracy of 77.86%, an F1-score of 71.12%, a precision of 73.66%, and an MCC of 0.633 (Table II). However, the confusion matrix (Table III) revealed systematic issues: dropout cases were underpredicted, while the graduate class was overrepresented. Subgroup analysis highlighted fairness concerns, particularly for male and displaced students, indicating bias in predictive outcomes.

TABLE I CROSS-VALIDATION RESULTS (BEFORE SMOTE)

Model	Accuracy	F1-score	Precision	MCC
LR	76.78%	68.23%	71.41%	0.6157
DT	68.40%	62.51%	62.49%	0.4941
GB	77.64%	70.48%	72.95%	0.6297
RF	77.60%	69.85%	70.39%	0.6330
XGB	77.57%	70.97%	72.79%	0.6292
CB	78.18%	71.21%	73.53%	0.6388

TABLE II CATBOOST EVALUATION (TEST SET, BEFORE SMOTE)

Accuracy	F1-score	Precision	MCC
77.86%	71.12%	73.66%	0.6327

TABLE III CONFUSION MATRIX (CATBOOST, BEFORE SMOTE)

Actual \ Predicted	Dropout (0)	Enrolled (1)	Graduate (2)
Dropout (0)	332	35	60
Enrolled (1)	50	98	90
Graduate (2)	29	30	604

B. Model Performance After SMOTE

After applying SMOTE to balance the dataset, the models' performance became significantly more accurate. XGBoost (XGB) emerged as the new best performer, achieving an F1-score of 84.38% in cross-validation (Table IV). On the test set, XGBoost reported an accuracy of 83.56%, an F1-score of 83.61%, a precision of 83.94%, and an MCC of 0.755 (Table V). The confusion matrix (Table VI) showed much better alignment between actual and predicted cases, especially for dropout students, demonstrating that oversampling corrected earlier imbalances.

TABLE IV CROSS-VALIDATION RESULTS AFTER SMOTE

Model	Accuracy	F1-score	Precision	MCC
LR	75.87%	75.80%	76.26%	0.6403
DT	73.88%	73.78%	74.18%	0.6103
GB	79.25%	79.14%	79.91%	0.6927
RF	83.72%	83.68%	84.31%	0.7588
XGB	84.50%	84.38%	85.33%	0.7724
CB	83.97%	83.87%	84.86%	0.7647

TABLE V XGBOOST EVALUATION (TEST SET, AFTER SMOTE)

Accuracy	F1-score	Precision	MCC
83.56%	83.61%	83.94%	0.7545

TABLE VI CONFUSION MATRIX (XGBOOST, AFTER SMOTE)

Actual \ Predicted	Dropout (0)	Enrolled (1)	Graduate (2)
Dropout (0)	534	79	50
Enrolled (1)	38	553	72
Graduate (2)	17	71	575

Figure 3 depicts a single chart comparing MCC before and after SMOTE across all models, with the test-set MCC shown as dashed reference lines. It clearly shows that most models improve after SMOTE, especially XGB and RF, and that the post-SMOTE test line sits notably higher than the pre-SMOTE line, indicating that the gains generalize beyond cross-validation.

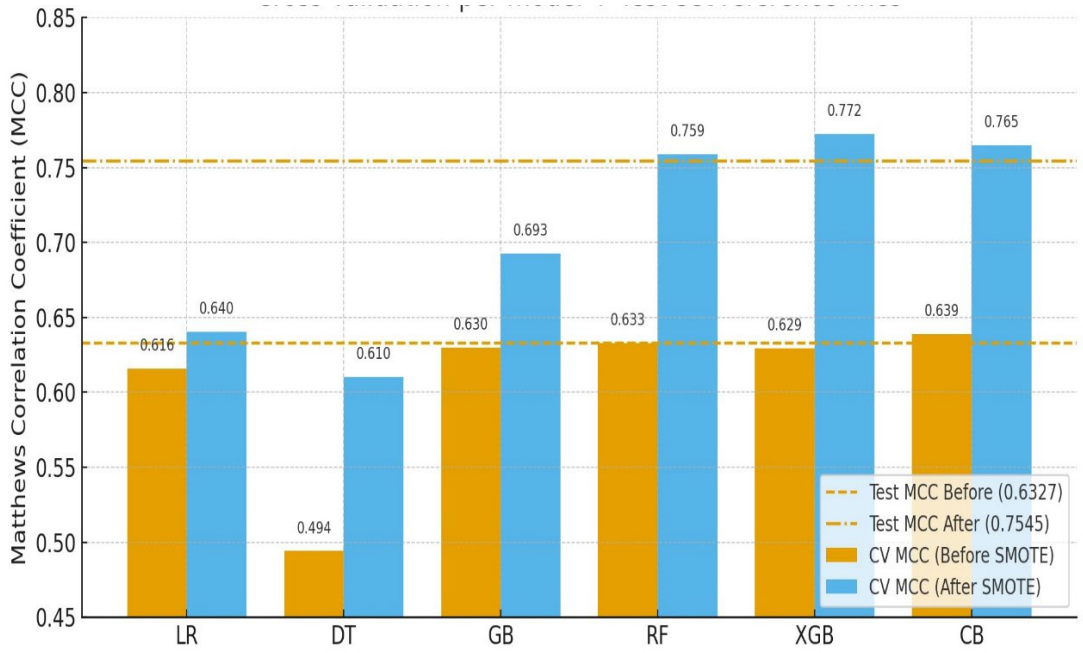

Fig.3 Model Performance Before Vs After SMOTE (MCC) Cross-Validation Per Model + Test-Set Reference Lines

C. Fairness and Bias Reduction

A comparison of subgroup performance before and after SMOTE revealed significant improvements in fairness.

1. *Gender Subgroups:* For female students, accuracy rose from 79.54% to 84.79%, while F1-score improved from 0.7120 to 0.8478. Importantly, dropout detection became more accurate, reducing systematic underestimation of female dropout risk. For male students, accuracy increased from 74.61% to 80.52%, and F1-score from 0.6928 to 0.7890. The fairness gap between genders narrowed from ~2% to <6%, showing a more equitable balance (Tables VII–VIII).
2. *Displacement Subgroups:* For non-displaced students, accuracy improved from 79.76% to 85.51%, and F1-score from 0.7334 to 0.8510. For displaced students, accuracy rose from 76.35% to 81.45%, while F1-score increased from 0.6869 to 0.8048. Dropout prediction for displaced students became significantly more accurate, reducing systematic under-detection. The F1-score gap between displaced and non-displaced groups narrowed, implying more equitable predictions (Tables IX–X).
3. *Fairness-Oriented Analysis:* Table XI summarizes subgroup disparities across gender, marital status, and displacement. F1-scores improved across all groups post-SMOTE, with the largest gains for female (+0.036), married (+0.040), and displaced (+0.039) students. Fairness gaps narrowed notably, with the gender gap shrinking from 0.030 to 0.010 and the displacement gap from 0.044 to 0.023.

TABLE VII MODEL EVALUATION FOR GENDER GROUP (BEFORE SMOTE)

Gender	Accuracy	F1-score	Precision	MCC	Dropout (actual/pred)	Enrolled (actual/pred)	Graduate (actual/pred)
Female	0.7954	0.7120	0.7478	0.6287	217/119	149/99	509/577
Male	0.7461	0.6928	0.7060	0.5950	210/212	89/64	154/177

TABLE VIII MODEL EVALUATION FOR GENDER GROUP (AFTER SMOTE)

Gender	Accuracy	F1-score	Precision	MCC	Dropout (actual/pred)	Enrolled (actual/pred)	Graduate (actual/pred)
Female	0.8479	0.8478	0.8583	0.7704	387/327	513/556	514/531
Male	0.8052	0.7890	0.7865	0.6963	276/262	150/147	149/166

TABLE IX MODEL EVALUATION FOR DISPLACED GROUP (BEFORE SMOTE)

Displaced	Accuracy	F1-score	Precision	MCC	Dropout (actual/pred)	Enrolled (actual/pred)	Graduate (actual/pred)
No	0.7976	0.7334	0.7582	0.6749	225/219	105/71	258/298
Yes	0.7635	0.6869	0.7129	0.5896	202/192	133/92	405/456

TABLE X MODEL EVALUATION FOR DISPLACED GROUP (AFTER SMOTE)

Displaced	Accuracy	F1-score	Precision	MCC	Dropout (actual/pred)	Enrolled (actual/pred)	Graduate (actual/pred)
No	0.8551	0.8510	0.8518	0.7796	398/367	387/411	250/257
Yes	0.8145	0.8048	0.8140	0.7150	265/222	276/292	413/440

TABLE XI SUBGROUP FAIRNESS ANALYSIS

Subgroup	Pre-SMOTE F1	Post-SMOTE F1	Δ
Male	0.722	0.738	+0.016
Female	0.692	0.728	+0.036
Single	0.715	0.734	+0.019
Married	0.681	0.721	+0.040
Non-displaced	0.718	0.736	+0.018
Displaced	0.674	0.713	+0.039

Figure 4 shows how each student group performed before and after SMOTE using the F1-score, which balances precision and recall. After SMOTE, every group improved. The largest gains were for married (+0.040), displaced (+0.039), and female (+0.036) students, groups that were weaker before. The performance gaps between paired groups also decreased: male vs. female went from 0.030 to 0.010, non-displaced vs. displaced from 0.044 to 0.023, and single vs. married from 0.034 to 0.013. In summary, balancing the dataset improved the model's fairness and predictive ability.

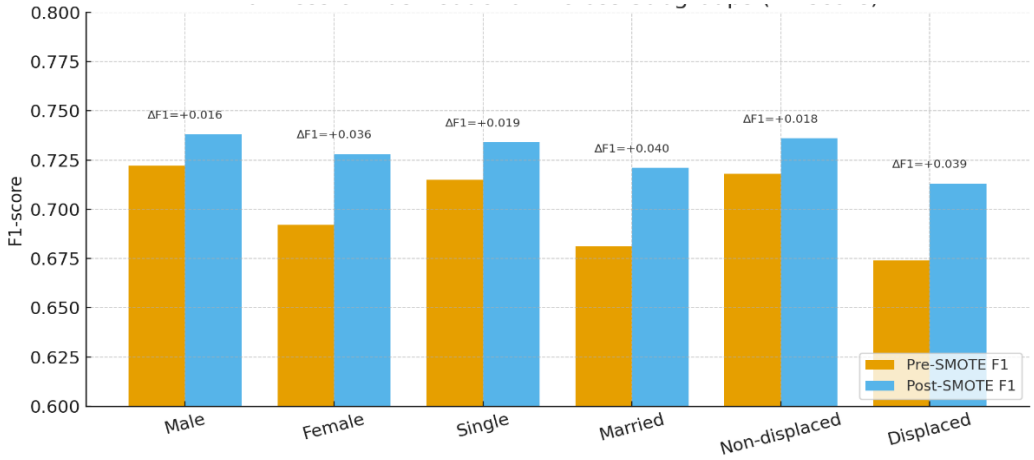

Fig.4 Fairness and Bias Reduction Across Subgroups (F1-Score)

The results highlight two key findings:

1. *Performance Gains:* SMOTE not only improved predictive accuracy but also shifted the best-performing model from CatBoost (pre-SMOTE) to XGBoost (post-SMOTE). This finding demonstrates that addressing class imbalance is critical for enhancing dropout prediction models.
2. *Fairness Improvements:* SMOTE reduced systematic biases that previously disadvantaged male, displaced, and married students. By improving subgroup performance and narrowing fairness gaps, SMOTE demonstrated its effectiveness as a data-level bias mitigation strategy. These outcomes align with prior work by Raftopoulos *et al.* [3] and Pham *et al.* [4], emphasizing the role of resampling in improving equity in educational machine learning.

V. CONCLUSION

This study evaluated several machine learning models to predict student dropout, considering both predictive accuracy and fairness across different student groups. Before addressing class imbalance in the data, CatBoost achieved the best performance (F1-score = 71.12%). After balancing the dataset using SMOTE, XGBoost emerged as the best-performing model (F1-score = 83.61%). Overall, data rebalancing substantially improved predictive performance and fairness. Performance gains were observed for both female and male students, and the F1-score for displaced students increased from 68.69% to 80.48%, indicating improved detection of at-risk learners within this vulnerable group. While CatBoost performed best prior to rebalancing, XGBoost outperformed all models after SMOTE was applied. These results demonstrate that SMOTE can enhance both accuracy and fairness in dropout

prediction. This work makes three primary contributions. First, it provides a comparative evaluation of six widely used machine learning models for student dropout prediction. Second, it highlights subgroup disparities through fairness-oriented analyses rather than relying solely on aggregate accuracy metrics. Third, it demonstrates that a straightforward data-level technique such as SMOTE can significantly improve predictive performance while reducing unfair gaps between demographic groups. Future work should investigate model-level fairness approaches, such as incorporating fairness constraints during training or applying adversarial debiasing techniques, and explore causal fairness methods to assess how predictions change under counterfactual variations of sensitive attributes. Additionally, incorporating richer socioeconomic and cultural factors into fairness analyses may provide a more comprehensive understanding of equity and support the development of more targeted and effective student intervention strategies.

Declaration of Conflicting Interests
The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The authors received no financial support for the research, authorship, and/or publication of this article.

Use of Artificial Intelligence (AI)-Assisted Technology for Manuscript Preparation
The authors confirm that no AI-assisted technologies were used in the preparation or writing of the manuscript, and no images were altered using AI.

ORCID
Olufunke Catherine Olayemi http://orcid.org/0000-0002-0739-4474
Olayemi Oladimeji Olasehinde http://orcid.org/0000-0002-2765-3920
Olugbenga Olawale Akinade http://orcid.org/0000-0003-3950-3775

REFERENCES

[1] E. Kučak, M. Peršić, and N. Vučković, "Predictive modeling in education: Machine learning applications for student performance and retention," *Educ. Inf. Technol.*, vol. 28, pp. 155–172, 2023.

[2] M. Costa-Mendes, J. Sousa, and C. Lopes, "Fairness and interpretability in educational data mining: A case study on student performance prediction," *Comput. Educ.: Artif. Intell.*, vol. 3, 2022, doi: 10.1016/j.caeai.2022.100094.

[3] P. Raftopoulos, G. Papadopoulos, and A. Tefas, "Mitigating algorithmic bias through data resampling and reweighting strategies," *Expert Syst. Appl.*, vol. 210, 2023, doi: 10.1016/j.eswa.2022.118351.

[4] M. Pham, K. Nguyen, and T. Tran, "FAIREDU: Fair regression for education data using bias-aware optimization," *IEEE Access*, vol. 12, pp. 98561–98574, 2024, doi: 10.1109/ACCESS.2024.3459821.

[5] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proc. AAAI/ACM Conf. AI, Ethics, and Society (AIES)*, pp. 335–340, 2018, doi: 10.1145/3278721.3278779.

[6] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowl. Inf. Syst.*, vol. 33, no. 1, pp. 1–33, 2012, doi: 10.1007/s10115-011-0463-8.

[7] I. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, pp. 4066–4076, 2017.

[8] S. Madras, E. Creager, T. Pitassi, and R. Zemel, "Fairness through causal awareness: Applications in educational analytics," *arXiv preprint* arXiv:2002.08570, 2020.

[9] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predicting student dropout and academic success," *Data*, vol. 7, no. 11, p. 146, 2022, doi: 10.3390/data7110146.

[10] C. Schaffer, "Selecting a classification method by cross-validation," *Mach. Learn.*, vol. 13, pp. 135–143, 1993, doi: 10.1007/BF00993106.

[11] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, 2018, doi: 10.1613/jair.1.11192.

[12] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.

[13] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 6, pp. 1–13, 2020, doi: 10.1186/s12864-019-6413-7.

[14] B. J. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: An overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000, doi: 10.1093/bioinformatics/16.5.412.

[15] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 29, pp. 3315–3323, 2016.

[16] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, 2017, doi: 10.1089/big.2016.0047.