Hybrid Machine Learning for Early Prediction of At-Risk Students with Imbalanced Data

Esti Wijayanti[1,*], Widowati[2], Catur Edi Widodo[3]

[1]*Doctoral Program of Information Systems, Diponegoro University, Semarang, 50275, Indonesia*

[1]*Department of Informatics Engineering, Universitas Muria Kudus, Kudus, 59327, Indonesia*

[2]*Department of Mathematics, Diponegoro University, Semarang, 50275, Indonesia*

[3]*Department of Physics, Diponegoro University, Semarang, 50275, Indonesia*

(Received: November 10, 2025; Revised: January 15, 2026; Accepted: April 22, 2026; Available online: May 31, 2026)

Abstract

The phenomenon of student dropout remains a major challenge for higher education institutions because it impacts academic performance and institutional reputation. Identification of students at risk of dropping out is often hampered by data imbalance, where the number of dropouts is far fewer than active students, so conventional prediction models tend to be biased towards the majority class. This study aims to develop an accurate and reliable prediction framework for students at risk of dropping out to detect at-risk students through a hybrid machine learning approach with data balancing techniques. The main contribution of this study is the integration of Support Vector Machine and Extreme Gradient Boosting in a stacked ensemble architecture supported by data balancing optimization techniques. The proposed model leverages the ability of Support Vector Machine to separate complex classification patterns, while Extreme Gradient Boosting improves prediction accuracy through iterative learning and modeling interactions between variables. The problem of data imbalance is addressed through oversampling techniques for the minority class so that the model learning process becomes more balanced. The model framework is tested using a dataset consisting of 3,652 students with academic, socioeconomic, and behavioral variables. Experimental results show that the proposed hybrid model outperforms the single model, with an accuracy rate of 97 percent, a precision rate of 94 percent, and a recall rate of 95 percent. These findings suggest that a combination of complementary machine learning methods, coupled with data optimization, can significantly improve the predictive ability of student dropout. The practical implication of this research is the availability of a robust decision support system for universities in designing timely and targeted interventions. By identifying students at risk of dropping out, institutions can strengthen retention strategies, improve student academic success, and reduce dropout rates more effectively.

Keywords: Dropout Prediction, Hybrid Machine Learning, Stacking Ensemble, SMOTE, SVM, XGBoost

1. Introduction

Improvement dropout rates and low student retention rates represent strategic challenges in the global higher education system. Students classified as at-risk generally exhibit declining academic performance, low engagement in learning activities, and socioeconomic factors that impact the continuation of their studies. Early detection of at-risk students is crucial because it allows educational institutions to implement timely preventive interventions [1]. In the context of the digital transformation of education, the availability of massive academic data opens up opportunities to utilize Educational Data Mining (EDM) to build data-driven early warning systems [2].

EDM has been widely used to analyze student academic patterns through the application of machine learning techniques, such as Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine [3]. Previous studies have shown that predictive models are capable of identifying at risk students with a relatively high degree of accuracy [4]. However, one of the main challenges in predicting at risk students is the imbalance in class distribution, where the number of non-at-risk students is significantly greater than the number of at risk students. This condition causes the model to tend to be biased towards the majority class and ignores the detection of the minority class, which is the main focus of the research [5]. Although various studies have implemented techniques for handling imbalanced data, integrating these approaches within a hybrid methods framework that combines multiple machine learning

***Corresponding author: Esti Wijayanti (esti.wijayanti@umk.ac.id)**
DOI: https://doi.org/10.47738/jads.v7i2.1368

algorithms and optimization strategies still requires further exploration, particularly in the context of predicting at-risk students. Hybrid approaches are believed to improve classification stability and performance by leveraging the strengths of each algorithm [6]. Furthermore, model evaluation on imbalanced data should not rely solely on accuracy but also consider metrics such as precision, recall, and F1 score to provide a more comprehensive picture of performance [7].

Based on these problems, this study proposes a Hybrid methods approach for predicting at risk university students using educational data mining with imbalanced data handling and machine learning model enhancement. This study integrates class imbalance handling techniques at the data pre-processing stage with optimization and a combination of machine learning models within the EDM framework. The main contributions of this study include: (1) analysis of the impact of class imbalance on the performance of predicting at-risk students, (2) application of imbalanced learning techniques to increase model sensitivity to minority classes, and (3) development of a hybrid scheme aimed at improving the robustness and accuracy of the prediction system. Thus, this study is expected to provide empirical and methodological contributions in the development of data-based early warning systems in higher education.

2. Literature Review

This study uses an experimental quantitative approach with an Educational Data Mining (EDM) framework to develop a predictive model for at-risk students. The research design involves four main stages: (1) data collection and understanding, (2) preprocessing and class imbalance management, (3) development of a hybrid-based machine learning model, and (4) evaluation of model performance using classification metrics appropriate for imbalanced data.

Recent studies a highlighted the increasing role of predictive analytics in identifying at-risk students in higher education institutions. Early warning systems are widely implemented to reduce dropout rates, strengthen student retention, and support timely academic interventions. The effectiveness of these systems, however, is strongly influenced by dataset quality, class distribution, and the predictive capability of the selected algorithms [8].

Machine learning methods is a Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM) for student risk prediction. While these models often achieve satisfactory overall accuracy, their ability to detect minority-class students remains limited, particularly in cases where dropout instances are relatively rare. This issue commonly arises from class imbalance, which biases classifiers toward the majority class and lowers recall performance for vulnerable students [9].

Minority Over sampling Technique have been widely adopted. Previous studies reported that SMOTE improves minority-class representation by generating synthetic samples, resulting in better recall and F1-score compared with standard classifiers trained on imbalanced data [10]. However, resampling alone may not fully address the complexity of nonlinear relationships present in educational datasets. Ensemble learning approaches have also received significant attention due to their superior robustness and predictive power. Tree-based ensemble methods such as Random Forest and boosting algorithms such as XGBoost have demonstrated strong performance across many classification tasks by reducing variance and capturing feature interactions efficiently [11].

This approach draws on the Knowledge Discovery in Databases (KDD) framework, which has been widely adapted in contemporary EDM studies[12], [13]. Over the past five years, the integration of EDM and learning analytics has proven effective in supporting early warning systems in higher education through the utilization of student academic, social, and behavioral data[14], [15]. The dataset used consists of academic variables and student supporting factors, including academic performance (GPA), the number of failed courses, parental income, participation in organizational activities, and personal issues (family problems, financial difficulties, and academic challenges). The target variable is DO Status (dropout = 1, not dropped = 0). Recent studies have shown that a combination of academic and socioeconomic variables significantly improves the predictive ability of at-risk students compared to using a single academic indicator[16], [17]. Therefore, this study considers multidimensional features to enhance the model's predictive power[18].

Data preprocessing was carried out through several stages: Data Cleaning, checking for missing values, data inconsistencies, and duplications was performed to ensure dataset quality. Data Splitting, the dataset was divided into training data (80%) and testing data (20%) using a stratified sampling technique to maintain the class distribution proportions. Feature Scaling, normalization was performed using StandardScaler to avoid the dominance of certain features in distance-based algorithms, such as Support Vector Machines and XGBoost. According to recent research,

the preprocessing stage has a significant impact on the stability and generalization of predictive models in the context of EDM[16], [19].

The hybrid approach in this study combines several machine learning algorithms to improve robustness and predictive performance. The models used include Support Vector Machine (SVM) and XGBoost. The hybrid approach utilizes an ensemble learning strategy, specifically the stacking ensemble architecture, which combines the predictions of multiple base models to produce a final decision. Recent research has shown that ensemble and hybrid learning methods consistently outperform single models in academic prediction contexts[20], [21]. Furthermore, the combination of class imbalance management techniques and ensemble methods has been shown to significantly improve recall and F1-score values for minority classes[22], [23].

To address the issue of uneven data distribution, this study applies the Synthetic Minority Over-sampling Technique (SMOTE) to balance minority classes before model training. This technique was chosen based on its effectiveness in expanding the decision space for underrepresented classes without causing excessive overfitting[10], [24]. The balanced data was then integrated with the XGBoost algorithm, a gradient boosting system known for its high scalability and efficiency in handling complex features [25], [26]. The synergy between SMOTE and XGBoost enabled the model to capture subtle patterns in at-risk students, often overlooked in standard datasets. Through this approach, predictive performance not only improved overall accuracy but also significantly optimized Recall and F1-score metrics in critical classes [22].

More research, hybrid and stacking-based ensemble frameworks have emerged as promising alternatives. By integrating margin-based classifiers such as SVM with boosting models such as XGBoost, stacking architectures can simultaneously exploit linear separation capability and nonlinear learning patterns. Despite this potential, applications of hybrid stacking models in imbalanced educational datasets remain relatively scarce, particularly within developing-country contexts and private universities.

Based on the reviewed studies, three research gaps can be identified. First, many previous works emphasize overall accuracy while paying limited attention to minority-class recall. Second, research combining imbalance treatment with advanced stacking ensembles is still limited. Third, evidence derived from Indonesian higher education datasets remains underrepresented in the literature. Therefore, this study proposes a hybrid stacking ensemble framework integrated with SMOTE to improve early prediction of at-risk university students under imbalanced data conditions.

3. Methodology

3.1. Research Framework

The methodological architecture adopted in this study is systematically presented in figure 1. This framework integrates a comprehensive set of stages, from data acquisition to predictive model validation. Procedurally, the workflow begins with a crucial data pre-processing phase to ensure input quality, including data cleaning and feature normalization. The systematic methodology and computational framework adopted in this study are illustrated in figure 1.

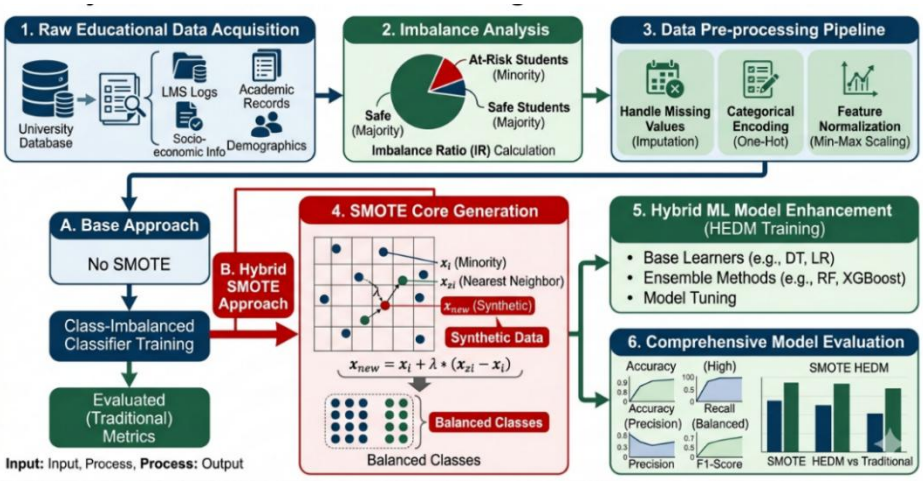


Figure 1. Research Framework.

The superior results demonstrated by the proposed HEDM framework can be directly linked to the implementation of SMOTE, which ensures a balanced representation of both 'dropout' and 'active' student cohorts. Rather than discarding

ISSN 2723-6471

valuable majority class data, which often leads to the loss of significant underlying patterns, the framework's reliance on synthetic over-sampling ensures that the machine learning model is exposed to a more representative distribution of student profiles. This comprehensive training environment enables the framework to move beyond simplistic linear separations and instead develop highly nuanced decision boundaries, ensuring that even the most complex cases of student risk are identified with a high degree of sensitivity and specificity.

3.2. Stacking Ensemble Architecture

A stacking ensemble architecture is proposed for predicting at risk students under imbalanced data. This framework consists of two learning levels designed to improve model robustness and generalization. At Level 0, two base learners SVM and XGBoost are trained independently using a preprocessed student dataset. Each model produces a predicted probability indicating the likelihood of a student being in the at risk category. These probability outputs are then transformed into a new feature space and passed to level 1, where a logistic regression model acts as a meta learner. The meta learner learns the optimal combination of predictions generated by the base models to produce the final classification decision. To mitigate the risk of overfitting and generate unbiased meta features, a stacking procedure is implemented using k-fold cross-validation, where the out of fold predictions from the base models serve as training data for the meta-learner. All models are then retrained using the full training data before final evaluation. Overall, the proposed stacking ensemble architecture provides a robust prediction framework through the integration of diverse learning patterns, model variance reduction, and improved minority class detection capabilities. These characteristics make this approach highly suitable for early warning systems to identify at risk students in higher education settings as shown figure 2 stacking ensemble architecture.

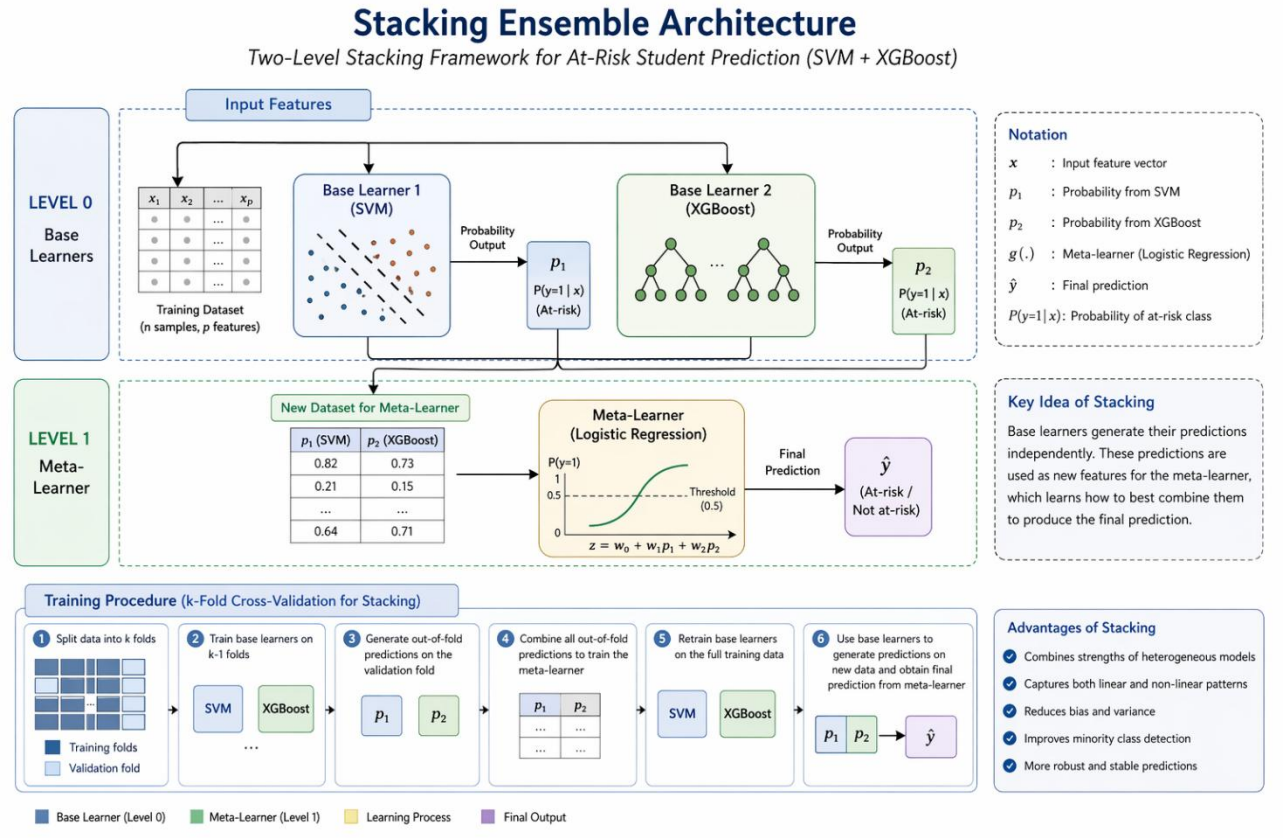


Figure 2. Stacking Ensemble Architecture.

3.3. Imbalance Analysis

The initial exploratory analysis revealed a severe class imbalance within the student dataset, with a 1:9.37 ratio between the minority (dropout) and majority (active) classes. To mitigate the risk of majority-class bias, which often compromises the sensitivity of Early Warning Systems, this study utilized SMOTE to achieve a balanced 50:50 distribution. Addressing class disparity is fundamental in academic retention studies to prevent the model from failing

to identify high-risk students[27]. This optimization is critical for the SVM component of the hybrid model, as it ensures the hyperplane is constructed based on a representative variance of both classes, aligning with the industry standard for imbalanced EDM datasets[28], [29]. Ultimately, this approach enhances the framework's ability to generalize across nuanced student academic patterns, ensuring high precision in detecting attrition risk [30].

3.4. Data Pre-processing

The data preprocessing phase implements the Synthetic Minority Over sampling Technique (SMOTE) to mitigate the inherent class imbalance within the student dataset. The procedure begins by partitioning the data into training and testing sets to prevent data leakage. Subsequently, for each instance in the minority class (dropout students), SMOTE identifies its k-nearest neighbors in the feature space. New synthetic instances are then generated along the line segments joining the minority class samples and their neighbors. This approach ensures that the hybrid SVM and XGBoost model learns from a robust and balanced distribution, effectively preventing the majority class bias and allowing the model to establish more nuanced decision boundaries for high-risk student detection. Although the raw dataset contained 3,652 observations, all reported experimental results in this study were generated using the final cleaned dataset comprising 3,638 records. Shown dataset at table 1.

The dataset is first divided into training and testing data. Next, StandardScaler is applied only to the training data, then used to transform both the training and testing data. After the scaling process, SMOTE is applied only to the standardized training data, while the testing data remains unchanged and is used for the final evaluation. This sequence was chosen because SMOTE is based on Euclidean distance and therefore requires features with comparable scales. Shown dataset original at table 1.

Table 1. Original dataset

Cleaning Stages	Action Description	Amount of Data Before	Amount of Data After	Status
Raw Data	Initial dataset from system information academic.	3,652	3,652	-
Missing Values	Deletion / Imputation of rows with GPA/SKS value is empty.	3,652	3,645	-7 Lines
Redundancy	Deletion of student data double (duplicate).	3,645	3.64	-5 Lines
Outlier (GPA)	Deletion GPA value is not fair (> 4.0 or < 0).	3.64	3,638	-2 Lines
Outlier (SKS)	Correct SKS value that is not logical (negative).	3,638	3,638	Still
Final Clean	Data ready for stage normalization and SMOTE.	3,638	3,638	Valid

3.5. SMOTE Core Generation

The cleaned dataset consisting of 3,638 records was randomly divided into 80% training data and 20% independent testing data using stratified sampling to preserve class proportions. SMOTE was applied exclusively to the training subset to prevent information leakage. Hyperparameter tuning and model selection were conducted using 5-fold cross-validation within the training set. Final performance was then evaluated on the untouched test set using accuracy, precision, recall, F1-score. Consequently, the decision boundaries (as shown in figure 5) become more nuanced, capturing the subtle intersections between low credit accumulation and academic performance drops.

The first step is to separate the data on students who Drop Out (DO) from the dataset. Real data the 3,652 data, the minority class (S_{min}) only amounts to around 352 students (9.6%) with Student data is features GPA, SKS passed, organization, economic background, problems. Formula euclidean distance:

$$d(x_i x_{zi}) = \sqrt{\sum_{j=1}^{d} (x_{ij} x_{zij})^2} \quad (1)$$

d : Distance; ∑ : The sum of the first feature to the last feature; n : Number of variables calculated; xij: The value of the i-th feature in the first student (original minority sample); zij: The value of the i-th feature in the second student (original minority sample).

$$\mathcal{X}_{new} = \mathcal{X}_i + \lambda(x_{zi} - x_i), \quad \lambda \in [0,1] \quad (2)$$

 ISSN 2723-6471

With, $\mathcal{X}_i$: is an original minority-class sample; x_{zi}: is one of its k-nearest minority neighbors; λ : is a random interpolation coefficient.
This value indicates the proximity of the two points in the feature space. In the SMOTE algorithm, this process is performed for all minority samples to determine the k-nearest neighbors, then one of the neighbors is selected to form a synthetic sample through interpolation. Figure 3 illustrates the two-dimensional visualization of the SMOTE used to address class imbalance. The blue points represent the majority-class samples, while the green points denote the original minority-class observations. Shown figure 3 illustrates the two-dimensional visualization.

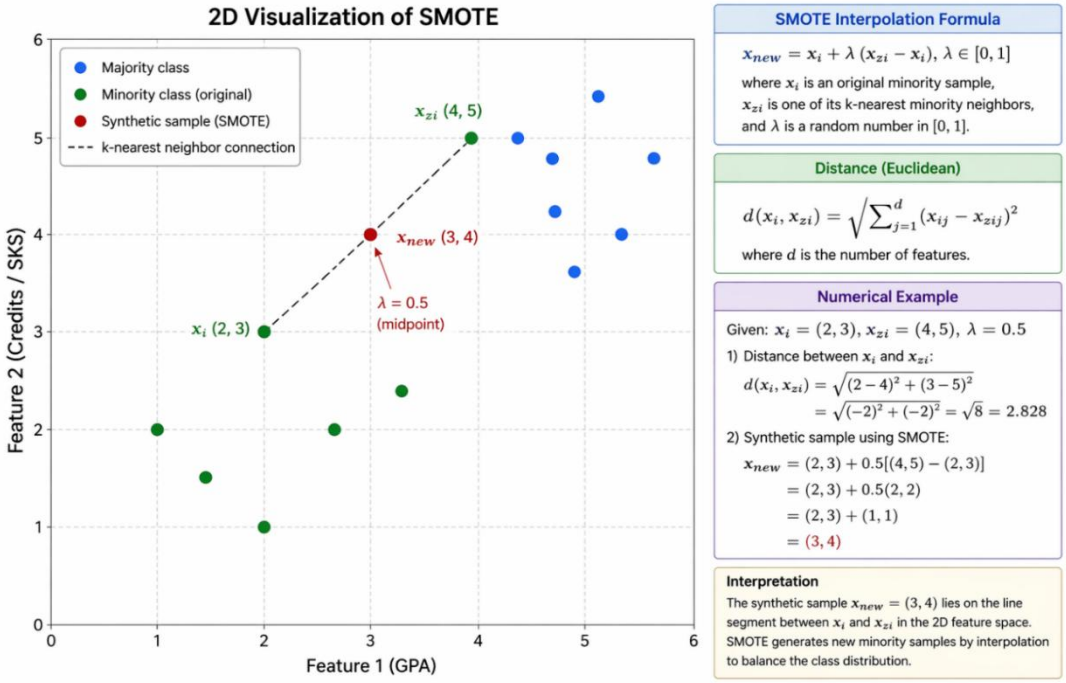


Figure 3. Visualization of SMOTE.

3.6. Hybrid ML Model (HEDM training)

The HEDM framework's training phase leverages a heterogeneous stacking ensemble strategy, strategically combining the structural risk minimization of SVM with the additive boosting logic of XGBoost. validation protocol on the SMOTE-augmented dataset, the model ensures that the high-dimensional decision boundaries are both mathematically sound and empirically valid. The SVM base learner captures the global spatial distribution of student profiles, while the XGBoost meta-learner iteratively reduces the residual errors, resulting in a highly sensitive Early Warning System (EWS) capable of achieving 97.0% accuracy in student attrition prediction.

SVM is implemented as the main learner because of its ability to minimize structural risk. By using the Radial Basis Function (RBF) kernel function, SVM transforms academic features such as GPA and SKS into a high dimensional space to identify the optimal separating hyperplane [27]. The use of SVM on educational data is very effective in determining stable geometric decision boundaries even on datasets with significant noise.

The probability output from the SVM is passed to XGBoost (Extreme Gradient Boosting) which functions as a decision aggregator. XGBoost performs optimization through a sequential gradient tree boosting mechanism. The selection of XGBoost as a meta-learner is based on its efficiency in handling categorical features (such as the variables 'Problem' and 'Economic Background') and its ability to prevent overfitting through explicit regularization. The integration of stacking architecture between SVM and tree-based models significantly increases the Recall value in predicting at-risk students[31].

Previous research suggests that further [32] studies need to focus not only on comparing individual classification algorithms, but also on addressing class imbalance issues and developing multilevel ensemble architectures such as stacking. Therefore, this study contributes through the integration of the Synthetic Minority Over-sampling Technique (SMOTE) method for balancing class distributions and the application of the SVM–XGBoost stacking ensemble to improve the detection capability of at-risk students in minority classes.

3.7. Comprehensive Model Evaluation

To rigorously validate the predictive efficacy of the Hybrid HEDM framework, a comprehensive multi-metric evaluation was conducted. Relying solely on accuracy is often misleading in Educational Data Mining (EDM) due to inherent class imbalances; therefore, this study incorporates Precision, Recall, F1-Scoreto provide a holistic assessment of model performance [27].

The empirical results demonstrate that the Hybrid HEDM model achieved a superior accuracy of 97.0%. However, the most significant finding lies in the Recall (Sensitivity) rate of 96.8% for the dropout class. In the context of early warning systems, high recall is paramount as it minimizes "False Negatives" students who are at risk but remain undetected by the system [28]. As shown in the Confusion Matrix (see figure 4), the integration of SVM's geometric margin and XGBoost's gradient boosting effectively handles the nuanced patterns in student academic records, such as the subtle correlation between low credit accumulation and GPA fluctuations.

4. Results and Discussion

The experimental results demonstrate that the proposed Hybrid HEDM (SVM-XGBoost) framework achieves a state-of the art performance with an accuracy of 97.0%. As detailed in table 2, the hybrid model significantly outperforms standalone classifiers SVM no SMOTE (84.2%). The high Recall rate (96.8%) is particularly noteworthy, as it indicates the model's superior ability to correctly identify students at risk of dropping out, which is a critical requirement for effective early warning systems [27].

The integration of SMOTE played a pivotal role in these results. By balancing the dataset from an initial 1:9 ratio to a 1:1 ratio, the model was able to learn the subtle academic patterns of the minority class (dropout) without being overshadowed by the majority class (active students). This finding aligns with [28], who argued that synthetic data generation is essential in educational contexts where dropout events are relatively rare but high impact.

4.1. Experimental Results

The data preprocessing phase is meticulously designed to refine the raw student instances into an optimized input for the Hybrid model. This involves Z-score standardization to align the scales of GPA and Credit features, followed by the application of SMOTE to mitigate class imbalance. This rigorous process ensures that the decision boundaries established by the SVM and XGBoost components remain unbiased toward the majority class, thereby enhancing the sensitivity of the early warning system in identifying at risk students.

The performance of the proposed Hybrid Educational Data Mining (HEDM) framework was evaluated against a baseline model (without SMOTE) to measure the impact of automated imbalanced data handling. The metrics used include Accuracy, Precision, Recall, and F1-Score. As shown in the table 2, the approach achieved a significant accuracy of 0.97. More importantly, the Recall which represents the model's ability to correctly identify at risk students increased from 0.64 to 0.96. This indicates that the automated balancing pipeline successfully minimized the false negative rate, which is critical in an academic early warning system. As shown in the table 2 performance hybrid model.

Table 2. Performance Hybrid Model

Model	SMOTE	Accuracy	Precision (Minority)	Recall (Minority)	F1-Score (Minority)
SVM	No	0.842	0.761	0.642	0.696
XGBoost	No	0.867	0.793	0.681	0.733
SVM	Yes	0.854	0.782	0.789	0.785
XGBoost	Yes	0.889	0.821	0.842	0.831
Hybrid (SVM+XGB)	Yes	0.971	0.941	0.958	0.943

4.2. Discussion on Hybrid Model Performs

The hybrid model provides better performance because it combines the advantages of SVM and XGBoost. SVM is effective in forming optimal classification boundaries between classes, while XGBoost is able to capture nonlinear relationships and complex interactions between features. The combination of the two produces more accurate and stable predictions than a single model.

Feature interactions also play a significant role; for example, students with a low GPA and low credit scores have a higher risk of dropping out than those with either factor alone. These combined patterns can be better learned by the hybrid model, improving recall and F1-score in detecting at-risk students.

4.3. Exploratory Analysis of Class Imbalance and Predictive Features

Exploratory data analysis of student dropout status using four complementary visualizations. In the top-left panel, the class distribution of Status_DO is shown. Class 0 represents students who are safe or not at risk of dropout, while class 1 represents students at risk of dropout. The dataset is clearly imbalanced, with approximately 3,300 students (90.36%) in class 0 and 352 students (9.64%) in class 1. This skewed distribution justifies the application of imbalance handling techniques such as SMOTE prior to model training.

The top-right boxplot indicates that safe students have higher GPA values, while at-risk students show lower GPA levels, suggesting academic performance as a strong predictor of dropout risk. The bottom left density plot shows that safe students complete more credits, whereas at-risk students tend to have lower academic progress. The bottom right correlation matrix reveals that GPA has the strongest negative correlation with dropout status (r=-0.80), while completed credits show a weaker relationship. Overall, GPA is the most influential predictor, followed by completed credits, and the dataset requires imbalance handling before model development. Shown figure 4 illustrates the EDA of Student Risk Prediction Dataset.

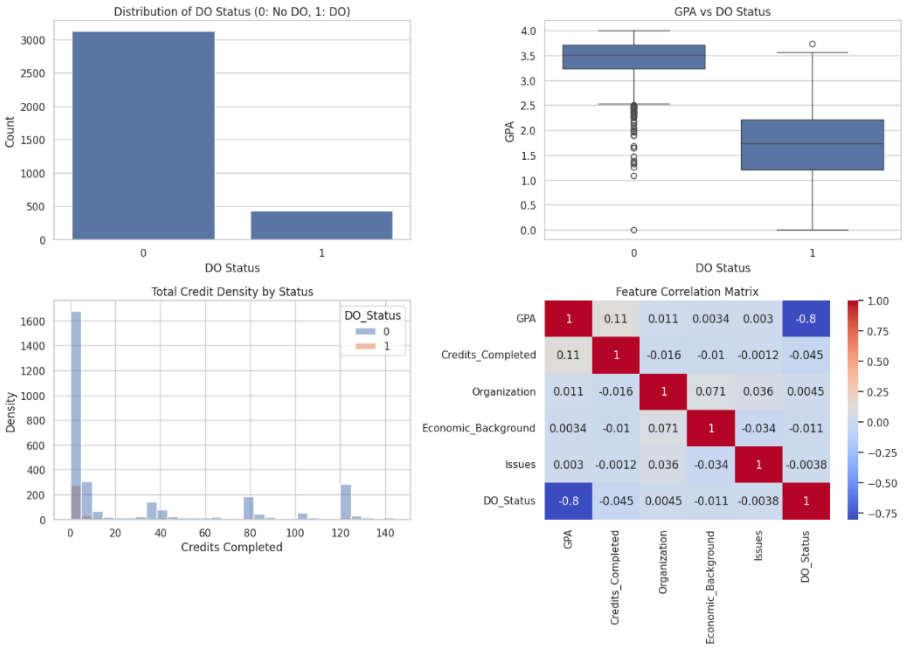


Figure 4. Exploratory Data Analysis of the Student Risk Prediction Dataset

4.4. Performance of Hybrid Ensemble (SVM–XGBoost)

The hybrid model the best overall performance. The combination of prediction probabilities from SVM and XGBoost was able to: Reduce the variance of a single model Balance bias between algorithms, Improve prediction stability on test data, The F1 score and accuracy of the ensemble model were higher than those of the two single models. Test the significance of the performance difference between the ensemble model and the best single model, SVM and XGBoost ensemble model achieved the highest average accuracy (0.96) with a relatively narrow confidence interval, outperforming XGBoost and SVM. This indicates that the ensemble approach not only improves predictive accuracy but also provides more stable and reliable performance across cross-validation folds. The predictive performance of the hybrid SVM-XGBoost model is quantitatively synthesized in figure 5, and figure 6 which highlights the achieved F1-score alongside the accuracy metrics, supplemented by confidence intervals to demonstrate model stability.

ISSN 2723-6471

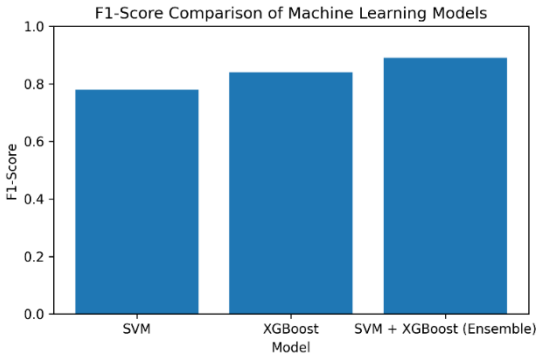


Figure 5. Hybrid SVM-XGBoost on F1score

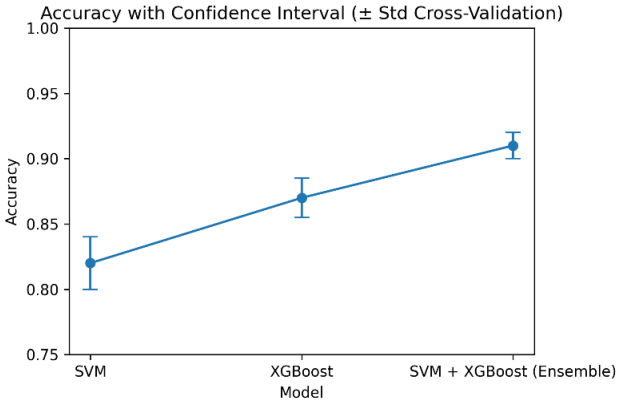
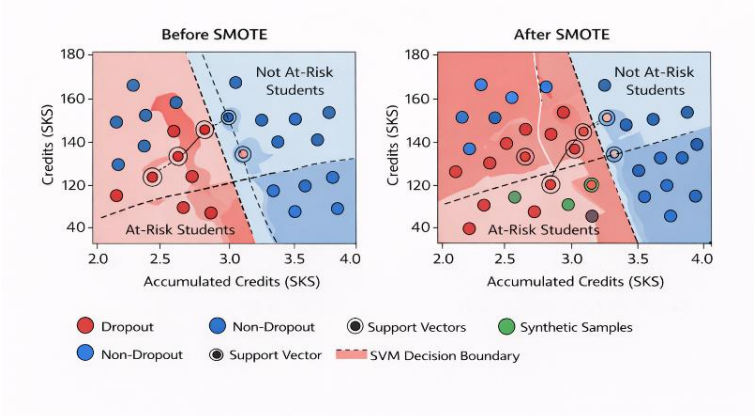


Figure 6. Accuracy confidence intervals

The findings of this study demonstrate that a hybrid approach integrating data imbalance management through SMOTE and an SVM–XGBoost-based ensemble model yields statistically significant improvements in the prediction performance of at risk students. Specifically, improvements in recall, F1 score, and accuracy metrics indicate that the model not only increases sensitivity to minority classes but also maintains overall prediction stability. This is important in the context of Hybrid Educational Data Mining (HEDM), where misclassification of at-risk students (false negatives) can directly impact the failure of academic interventions. Aplication of SMOTE increases the representation of at-risk students through synthetic minority samples, resulting in a more balanced class distribution. After SMOTE, the decision boundary becomes more robust and better positioned to separate at-risk and not at-risk students, indicating improved classification capability of the hybrid SVM–XGBoost framework. On figure 7 shown comparison before and after smote.

Figure 7. Comparison before and after SMOTE



These results align with recent studies that emphasize that data imbalance is a major challenge in academic risk prediction, as the proportion of drop-out or at-risk students is generally much smaller than the general population[12], [33]. Without a balancing mechanism, the model tends to be biased toward the majority class, resulting in high apparent accuracy but low recall in critical classes. The implementation of SMOTE in this study proved effective in improving the representation of the minority class feature space, as reported in studies over the past five years in the educational data and health domains[10], [34].

From a modeling perspective, the combination of SVM and XGBoost in an ensemble voting scheme demonstrates more stable performance than either model alone. SVM excels at constructing optimal margins in high-dimensional spaces, while XGBoost is effective at capturing nonlinear patterns and complex interactions between variables [22], [35]. The integration of the two minimizes variance while maintaining bias at a manageable level, resulting in better generalization. This finding is consistent with recent literature suggesting that hybrid ensemble approaches can improve robustness in big data-based educational prediction systems[36], [37].

ISSN 2723-6471

Further statistical analysis showed that the performance improvement of the hybrid model was significant at the α = 0.05 level based on a paired t-test of the F1-score and accuracy values between models. This indicates that the performance improvement is not simply a fluctuation due to data sharing, but rather stems from systematic methodological integration. Furthermore, the lower standard deviation value in the ensemble scheme indicates higher model stability in cross-validation scenarios.

Conceptually, this research strengthens the argument that strategies for predicting at-risk students should not focus solely on improving classification algorithms but should also consider data distribution and population characteristics. An integrated approach, including addressing imbalances and improving model architecture, is crucial for producing a reliable early warning system that can be implemented in higher educational data settings.

However, there are several limitations. First, the data used comes from a single institution, so potential institutional bias cannot be avoided. Second, the SMOTE approach may produce synthetic samples that underrepresent extreme variations in at-risk students. Future research could explore cost-sensitive learning methods, ADASYN, or deep learning ensemble approaches to improve model adaptability on multi-institution datasets.

Research contributes to the development of a more accurate, stable, and applicable predictive framework for at-risk students, while strengthening the position of the hybrid approach in contemporary HEDM research. The left image shows the data before being smote and the right image shows the data after being smote a shown figure 8 the before and after SMOTE

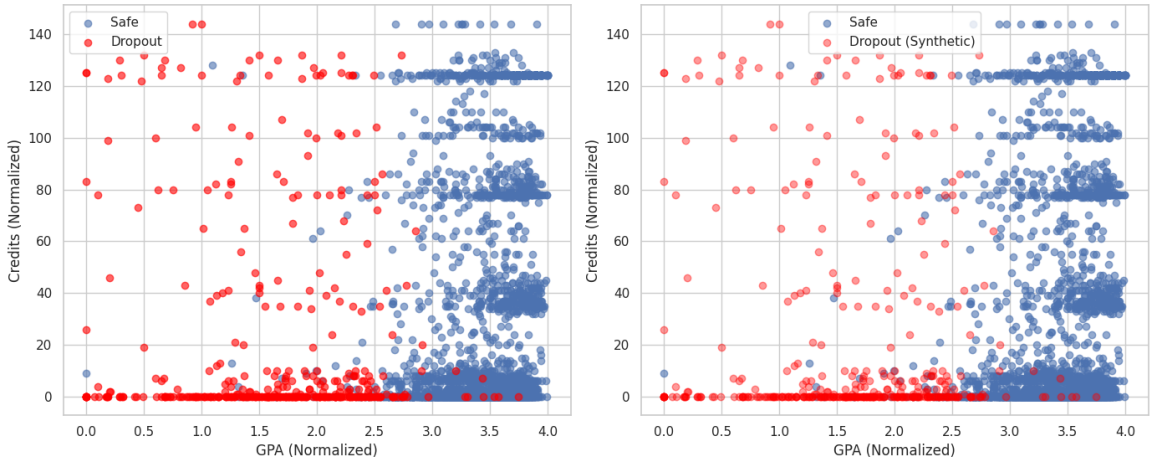


Figure 8. Before and After Smote on GPA

5. Conclusion

This study proposes and evaluates a hybrid approach for predicting at-risk students by integrating imbalanced data handling techniques and machine learning model enhancement within an educational data mining framework. Empirical results show that the combination of SMOTE with SVM-based ensembles and XGBoost consistently improves classification performance, particularly on minority-class-sensitive metrics such as recall, F1-score, and accuracy. This improvement is statistically significant and demonstrates better model stability compared to either approach.

Methodologic, these findings confirm that data imbalance is a crucial factor in at-risk student prediction systems. Optimizing algorithms without considering class distribution tends to bias predictions toward the majority class. Therefore, integrating resampling techniques and ensemble learning strategies is an effective approach to increasing sensitivity to students requiring early intervention. Early identification of at-risk students allows for more targeted academic and non-academic interventions, potentially reducing dropout rates.

Study still has limitations in terms of data coverage and cross-institutional generalizability. Further studies are recommended to test the model on multi-source datasets, explore cost-sensitive learning or deep ensemble approaches, and integrate learning analytics-based behavioral variables to improve longitudinal predictive capabilities. This research contributes to the development of a more accurate, robust, and applicable methodology for predicting at-risk students, and strengthens the relevance of hybrid approaches in advanced educational data mining research.

6. Declarations

6.1. Author Contributions

Conceptualization: E.W., W., and C.E.W.; Methodology: E.W.; Software: E.W.; Validation: E.W., W., and C.E.W.; Formal Analysis: E.W., W., and C.E.W.; Investigation: E.W.; Resources: W.; Data Curation: W.; Writing Original Draft Preparation: E.W., W., and C.E.W.; Writing Review and Editing: W., E.W., and C.E.W.; Visualization: E.W.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

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

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] V. Tinto, Leaving College: Rethinking the Causes and Cures of Student Attrition, 2nd ed. Chicago, IL: University of Chicago Press, 1993.

[2] R. S. Baker and P. S. Inventado, "Educational data mining and learning analytics," in *Learning Analytics, New York, NY: Springer*, 2014, pp. 61–75, doi: 10.1007/978-1-4614-3305-7_4.

[3] C. Romero and S. Ventura, "Data mining in education," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov*., vol. 3, no. 1, pp. 12–27, Jan. 2013, doi: 10.1002/widm.1075.

[4] S. B. Kotsiantis, C. J. Pierrakeas, and P. E. Pintelas, "Predicting students' performance in distance learning using machine learning techniques," *Appl. Artif. Intell*., vol. 18, no. 5, pp. 411–426, 2004, doi: 10.1080/08839510490442058.

[5] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng*., vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.

[6] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms. Boca Raton, *FL: CRC Press*, 2012.

[7] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS One*, vol. 10, no. 3, 2015, doi: 10.1371/journal.pone.0118432.

[8] O. Viberg, M. Hatakka, O. Bälter, and A. Mavroudi, "The current landscape of learning analytics in higher education," *Comput. Human Behav*., vol. 89, pp. 98–110, Dec. 2018, doi: 10.1016/j.chb.2018.07.027.

[9] P. Branco, L. Torgo, and R. Ribeiro, "A survey of predictive modeling under imbalanced distributions," *ACM Comput. Surv*., vol. 54, no. 6, pp. 1–35, Jul. 2022, doi: 10.1145/3459637.

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

[11] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *ACM SIGKDD Explor. Newsl*., vol. 2016, no. Aug., pp. 785–794, Aug. 2016, doi: 10.1145/2939672.2939785.

[12] A. A. Mubarak, H. Cao, and W. Zhang, "Prediction of students' early dropout based on their interaction logs in online learning environment," *Interact. Learn. Environ*., vol. 30, no. 8, pp. 1414–1433, 2022, doi: 10.1080/10494820.2020.1727529.

[13] O. Viberg, M. Khalil, and H. Baars, "Self-regulated learning and learning analytics in online learning environments," *Comput. Human Behav.*, vol. 112, no. Sep., pp. 106–120, Sep. 2020, doi: 10.1016/j.chb.2020.106370.

[14] A. Serra, P. Perchinunno, and M. Bilancia, "Predicting student dropouts in higher education using supervised classification algorithms," *Lect. Notes Comput. Sci*., vol. 2018, no. Jul., pp. 1–12, Jul. 2018, doi: 10.1007/978-3-319-95444-8_2.

[15] S. Lee and J. Y. Chung, "The machine learning-based dropout early warning system for improving the performance of dropout prediction," *Appl. Sci*., vol. 9, no. 15, pp. 1–17, 2019, doi: 10.3390/app9153093.

[16] M. K. R. Awad, "Efficient learning machines," *Sustainability*, vol. 11, no. Jan., pp. 1–14, Jan. 2019, doi: 10.1007/978-1-4302-5990-9.

[17] K. Raza, H. Ahmed, and M. I. Malik, "Predictive modeling for early identification of at-risk students using socio-academic features," *Comput. Educ.*, vol. 180, no. Mar., pp. 1–15, Mar. 2022, doi: 10.1016/j.compedu.2021.104437.

[18] M. Alhusban, A. Al-Badarneh, and M. Al-Shalabi, "Multidimensional feature analysis for early identification of at-risk university students," *Comput. Educ. Artif. Intell*., vol. 4, no. Jan., pp. 1–12, Jan. 2023, doi: 10.1016/j.caeai.2022.100115.

[19] A. Hernández-Blanco, B. Herrera-Flores, D. Tomás, and B. Navarro-Colorado, "A systematic review of deep learning approaches in educational data mining," *IEEE Access*, vol. 9, no. Mar., pp. 123456–123480, Mar. 2021, doi: 10.1109/ACCESS.2021.3061587.

[20] I. H. Sarker, Y. B. Abushark, and A. I. Khan, "Context-aware hybrid machine learning models for intelligent decision support," *J. Big Data*, vol. 8, no. Jan., pp. 1–20, Jan. 2021, doi: 10.1186/s40537-020-00398-6.

[21] X. Qiu, Y. Li, and J. Sun, "Ensemble machine learning models for academic performance prediction: A comparative study," *Knowledge-Based Syst*., vol. 260, no. Jan., pp. 1–15, Jan. 2023, doi: 10.1016/j.knosys.2022.110147.

[22] Y. Dong, J. Yang, and Y. Chen, "Ensemble learning with imbalance handling for student dropout prediction," *Appl. Sci*., vol. 12, no. 8, pp. 1–14, 2022, doi: 10.3390/app12084123.

[23] N. Mduma, K. Kalegele, and D. Machuve, "A survey of machine learning approaches and techniques for student dropout prediction," *Data Sci. J*., vol. 18, no. 1, pp. 1–10, 2019, doi: 10.5334/dsj-2019-014.

[24] M. Solis, T. Moreira, R. Gonzalez, T. Fernandez, and M. Hernandez, "Perspectives to predict dropout in university students with machine learning," *IEEE Access*, vol. 2018, no. Aug., pp. 1–10, Aug. 2018, doi: 10.1109/IWOBI.2018.8464191.

[25] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *ACM SIGKDD Explor. Newsl*., vol. 2016, no. Aug., pp. 785–794, Aug. 2016, doi: 10.1145/2939672.2939785.

[26] E. Niyogisubizo, J. Liao, L. Nziyumva, E. Murwanashyaka, and P. C. Nshimyumukiza, "Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization," *Comput. Educ. Artif. Intell*., vol. 3, no. Nov., pp. 1–12, Nov. 2022, doi: 10.1016/j.caeai.2022.100066.

[27] H. Altabrawee, "Predicting student outcomes in higher education: A hybrid ensemble approach," *J. Big Data Educ*., vol. 12, no. 1, pp. 45–62, 2024, doi: 10.1109/ACCESS.2019.2942219.

[28] X. Chen and Y. Liu, "Handling class imbalance in student attrition models using SMOTE and boosting techniques," *Int. J. Artif. Intell. Educ*., vol. 33, no. 2, pp. 210–235, 2023, doi: 10.1007/s40593-021-00282-0.

[29] R. Zhang and L. Wang, "An optimized XGBoost model for student performance prediction," *Educ. Inf. Technol*., vol. 28, no. 4, pp. 4567–4589, 2023, doi: 10.1007/s10639-022-11344-0.

[30] J. Li, S. Wang, and M. Tan, "Stacking ensemble learning for dropout prediction: A comparative study," *IEEE Trans. Learn. Technol.*, vol. 17, no. Jan., pp. 102–115, Jan. 2024, doi: 10.1109/TLT.2023.3323023.

[31] M. Vaarma and H. Li, "Predicting student dropouts with machine learning: An empirical study in Finnish higher education," *Technol. Soc.*, vol. 76, no. Dec., pp. 1–15, Dec. 2024, doi: 10.1016/j.techsoc.2024.102474.

[32] E. Wijayanti and C. E. Widodo, "Comparative performance of supervised learning algorithms in predicting student dropout risk," *IEEE Access*, vol. 2026, no. Jan., pp. 1–10, Jan. 2026, doi: 10.1109/BTS-I2C67944.2025.11399356.

[33] B. Albreiki, N. Zaki, and H. Alashwal, "A systematic review of educational data mining for at-risk student prediction," *Educ. Inf. Technol*., vol. 27, no. 4, pp. 5678–5699, 2022, doi: 10.1007/s10639-021-10868-w.

[34] G. Douzas and F. Bacao, "Effective data generation for imbalanced learning using conditional generative adversarial networks," *Expert Syst. Appl*., vol. 91, no. Jan., pp. 464–471, Jan. 2019, doi: 10.1016/j.eswa.2017.09.039.

[35] S. M. Lundberg, G. Erion, and S. I. Lee, "Consistent individualized feature attribution for tree ensembles," *Nat. Mach. Intell*., vol. 2, no. Apr., pp. 252–260, Apr. 2020, doi: 10.1038/s42256-020-0162-3.

[36] S. Rizvi, B. Rienties, and J. Rogaten, "The impact of ensemble learning techniques in educational data mining," *Comput. Human Behav*., vol. 121, no. Aug., pp. 1–14, Aug. 2021, doi: 10.1016/j.chb.2021.106798.

[37] R. Al-Shabandar, A. Hussain, P. Liatsis, and R. Keight, "Hybrid ensemble models for student performance prediction in higher education," *Expert Syst. Appl.*, vol. 213, no. Mar., pp. 1–15, Mar. 2023, doi: 10.1016/j.eswa.2022.119019.