Asian And Pacific Economic Review
ISSN: 1000-6052

AN OPTIMIZED HYBRID MACHINE LEARNING FRAMEWORK FOR STUDENT DROPOUT PREDICTION WITH EXPLAINABLE AI

Archana C, Sanju Gupta
Assistant Professor, Department of MCA, SIES College of Management Studies, Nerul, India
https://doie.org/10.65985/APER.2026191953

Abstract—Student dropout is a significant socio-economic issue that impacts both the students' personal growth and the school's performance. Accurate and timely identification of potential student dropouts is crucial for providing the necessary support. This paper introduces the Optimized Hybrid Machine Learning approach, combining XGBoost and Logistic Regression using a weighted soft voting technique (70% XGBoost, 30% LogRegres-sion) for accurate student dropout predictive modeling. To over-come class imbalance, both Random Upsampling and SMOTE techniques were used. Dataset processing was done using label encoding for categorical variables and Z-normalization for nu-merical values. Feature selection was also conducted for selecting significant features. Various predictive models, including logistic Regression, XGBoost, and the stacked approach, were integrated for experimentation. Experimental results demonstrate that the proposed Optimized Hybrid Machine Learning Model achieved a high predictive performance in identifying student dropouts, with an accuracy of 96.95% and a weighted F1-Score of 0.96, outperforming baseline models. Using SHAP and LIME methods for Explainable AI, significant factors for student dropouts based on academic, student behavioral, and socio-economic factors were identified, thus providing a reliable, accurate, and executable approach to the issue of student dropouts.
Index Terms—Student Dropout Predication, XGBoost, Logistic Regression, Hybrid Machine Learning, Model, Ensemble Learn-ing, SHAP, LIME, SMOTE, Upsampling.

I. INTRODUCTION

The education system plays a pivotal role in driving societal advancement and economic growth. Yet, higher education institutions [1]. Performance during this stage often serves as a strong predictor of long-term academic success. Student dropout is a multifaceted issue influenced by demographic, socio-economic, academic, and psychological factors. Early identification of at-risk students is therefore crucial, enabling institutions to initiate timely interventions before students discontinue their studies [2]. The dropout phenomenon is further complicated by diverse social, economic, personal, and health-related factors [3]. The COVID-19 pandemic amplified these challenges due to reduced face-to-face engagement, underscoring the necessity of automated, real-time monitoring systems to enhance institutional decision-making and resource utilization [4].

Despite existing research efforts, key gaps remain. Many current predictive models function as "black boxes," offering limited interpretability regarding their decision processes. Moreover, standard data-balancing techniques such as the Synthetic Minority Over-Sampling Technique [5]. This study aims to address these challenges through the following research questions:

RQ1: How can a hybrid machine learning model enhance dropout prediction accuracy compared to existing approaches?

RQ2: How can random up sampling, combined with the data balancing technique SMOTE, enhance model performance while minimizing data noise

RQ3: Which key academic, behavioural, and socio-economic factors contribute most to student dropout, as identified by Explainable AI techniques such as SHAP and LIME?
RQ4: How can interpretable and accurate models assist ed-ucators in designing effective interventions to reduce dropout rates?
To navigate these questions, the paper is structured to review the current literature, followed by a detailed explanation of the Optimized Hybrid methodology, an analysis of performance results, and an XAI-driven risk factor evaluation.

II. LITERATURE REVIEW

Recent studies have shifted from traditional statistical anal-ysis to sophisticated ML algorithms. [41] utilized weighted voting classifiers to predict student failure risk with 81.73% accuracy. [42] Demonstrated that Random Forest models could achieve up to 98.2% accuracy in identifying grades, though they noted that high complexity often reduces interpretability. [6], [7] The use of machine learning developed a cus-tom stacking ensemble model combining Random Forest introduced a Hybrid Logistic Regression–Neural Network (HLRNN) model using a public dataset of 4,424 samples, which achieved 96% accuracy. This hybrid architecture ef-fectively addressed class imbalance issues and was supple-mented by SHAP and LIME-based interpretability; however, the study noted the need for broader validation across diverse educational settings.
focused on dropout prediction using models such as Logistic Regression, Gradient Boosting, Random Forest, and Deep Learning, analyzing 33,627 student records from Assam, India. Logistic Regression outperformed others with 98.43% accuracy, demonstrating that simpler algorithms can remain competitive when features are well-engineered.
proposed a Conv-LSTM framework for week-wise prediction of at-risk students at Gadjah Mada University, achieving an F1-score of 91% and leveraging LIME to in-terpret patterns in learning behavior. Finally [5] developed a GRU-based Deep Predictive Network (GDPN) integrating recurrent and convolutional layers, applied to 6,455 records from the OULAD dataset, reaching 92.5% accuracy with attention mechanisms to highlight critical learning features.
The integration of XAI has become paramount. Nagy and Molontay emphasized personalized interventions using SHAP and LIME to explain dropout drivers in STEM fields.
Collectively, these studies demonstrate a transition from conventional ML models to hybrid and deep learning architectures with increasing emphasis on explainability. The consistent integration of SHAP, LIME, and attention mechanisms underscores the field’s commitment to transparent, educator-friendly AI systems. Despite high accuracy levels, limitations such as overfitting, dataset bias, and restricted generalizability remain significant. Future research should aim to develop scalable, cross-institutional frameworks that combine real-time analytics with interpretable insights to guide adaptive and personalized educational interventions.

III. RESEARCH GAP

Based on the synthesized literature, four primary research gaps are identified:

1) Limited Explainability: Most high-accuracy models act as black boxes, making it difficult for educators to understand the factors driving a “dropout” prediction.
2) Class Imbalance Noise: While SMOTE is widely used, it can introduce noise into the dataset; few studies explore noise-free alternatives like Radom Upsampling for student data.
3) Generalizability: Most models are trained on readily available datasets rather than local, real-time data, which limits their applicability to the specific context and dynamics of the institution.

4) Integration of Non-Academic Factors: There is a lack of comprehensive models that simultaneously weight socio-economic stressors (e.g., financial stress) along-side academic performance metrics.

IV. METHODOLOGY

Data integrity and architectural design are the cornerstones of effective machine learning algorithms. Our methodology prioritizes a clean, balanced, and scaled data pipeline before the execution with the various hybrid machine learning architectures.

A. *Dataset Description*

The study utilizes a dataset comprising more than 523 samples collected from a real-time institutional scenario. The dataset used for this research was compiled from multiple institutional sources, including Enterprise Resource Planning [8]. The dataset consists of 27 critical student-related features encompassing demographic, socio-economic, academic, financial and institutional support, as well as behavioral, psychological, and student performance-related factors [9]. The dependent feature, 'Target,' classifies students into two categories: 'Dropout' and 'Non-Dropout,' where Dropout is represented by 1 and Non-Dropout is represented by 0. To enhance analytical clarity and model interpetability, the selected features used for predicting student dropout are presented in Table I. These features were systematically analyzed and categorized into six principal groups based on their relevance and practical significance. The attributes collected and exam-ined under each category are described in Table I. The target classification was simplified into 'Dropout' vs. 'Non-dropout' (1 for Dropout, 0 for Non-Dropout) to sharpen the focus on institutional retention.

B. *Data Preprocessing*

The dataset contains mixed numerical and categorical features along with a binary target variable representing two outcomes of the target classes: Dropout and Non-Dropout. To ensure data quality, consistency, integrity, and model readiness, a structured preprocessing pipeline was implemented before the actual implementation of the hybrid machine learning model.

Missing values in numerical features were handled using the median imputation technique, as it reduces the impact of out-liers and preserves the central tendency of the data distribution [10]. For categorical features, missing values were replaced using mode imputation. The Student_ID attribute was treated purely as a unique identifier and therefore excluded from the analytical process to prevent bias and unnecessary model com-plexity. Categorical features were transformed into machine-compatible representations. Label encoding was applied to binary categorical attributes, whereas one-hot encoding was employed for multi-class categorical attributes to avoid ordi-nal misinterpretation of the data. Furthermore, all numerical features were standardized to maintain uniform scaling across variables, preventing features with larger magnitudes from disproportionately influencing the student academic dropout prediction system model training process [11].

C. *Data Balancing*

The dataset taken for the experimentation exhibited class imbalance in the binary target variable, where the Non-Dropout class had significantly higher representation than the Dropout class. Such imbalance can lead to biased model learning, causing predictive algorithms to Favor the majority class while reducing the accuracy and reliability of minority class predictions.

To address this issue, a hybrid data balancing strategy combining Random Upsampling and the Synthetic Minority Oversampling Technique [12]. Initially, random upsampling was applied to increase minority class representation by duplicating existing dropout records. However, simple duplication can lead to overfitting and reduced model generalization [13]. To overcome this limitation, SMOTE was subsequently applied to generate synthetic minority class samples through interpolation between existing minority observations. The application of SMOTE resulted in a balanced class distribution with equal representation of both dropout and non-dropout classes. This balancing approach enabled the predictive models to effectively learn minority class patterns, thereby improving overall robustness and predictive performance [14]. The detailed distribution of class balancing at each stage of the proposed system is presented in Table II.

TABLE I: Data Set Description

Category	Feature	Type	Description
Demographic Features	Student_ID	Categorical	Unique ID
	Gender	Categorical	Male/Female/Other
	Age	Numerical	Integer or float, continuous
	Marital_Status	Categorical	Single/Married/Other
Socio-Economic Background	Socio_Economic_Status	Categorical	Low/Mid/High
	Parent_Education_Level	Categorical	High School/Graduate/Post Grad
	Family_Income_Annual	Numerical	Annual income in currency units
	Residence_Type	Categorical	Urban/Rural
	Distance_From_College_km	Numerical	Continuous numeric distance
	Mode_of_Transport	Categorical	Bus/Car/Bike/Walk/Train
Academic Information	Program	Categorical	MCA/MMS etc.
	Specialization	Categorical	IT/Finance/Marketing/HR
	Admission_Year	Numerical	Year
	Entrance_Exam_Score	Numerical	Score in entrance exam (0–100)
	Previous_Academic_Percentage	Numerical	Percentage marks from prior education
Academic Performance & Engagement	Attendance_Percentage	Numerical	0–100%
	Internal_Assessment_Score	Numerical	Score
	Semester_1_GPA	Numerical	GPA
	Semester_2_GPA	Numerical	GPA
	Backlogs	Numerical	Count of failed courses
Financial & Inst. Support	Scholarship_Received	Categorical	Yes/No
	Faculty_Feedback_Score	Numerical	1–5 Likert scale or score out of 10
Behavioural & Psychological Factors	Financial_Stress_Level	Categorical	Low/Medium/High
	Placement_Anxiety_Level	Categorical	Low/Medium/High
	Mental_Health_Score	Numerical	Score 0–100
	Extracurricular_Participation	Categorical	Yes/No
	Internet_Access_At_Home	Categorical	Yes/No

D. Feature Scaling

In order to standardize numerical feature values and guarantee that each attribute contributed uniformly during model training, feature scaling was applied to the pre-processed dataset. Since variables in the dataset were assessed throughout a range of values, it is possible that this could skew learning by favouring features with higher magnitudes. As a result, feature scaling was required to preserve consistency and enhance model functionality. To improve the model's accuracy, precision, and other performance metrics across all features, the Z-score normalization technique was used. Feature scaling was applied to eleven numerical attributes, including age, family income, distance from college, academic performance indicators, attendance records, GPA scores, backlog count, faculty feedback score, and mental health score. The Identifier variable Student ID were excluded from feature scaling as they do not contribute to predictive learning [15]. Categorical features were handled separately during the preprocessing stage using appropriate encoding techniques.

E. *Model Development*

After the successful completion of the data preprocessing stage and the data balancing and feature scaling phase, the model construction phase commenced with the deployment of various machine learning models. In this phase, the dataset was partitioned into training and testing subsets, with 75% of it allocated for training and the remaining 25% for testing.

Weighted Machine Learning Model. The performance of each model was carefully evaluated. The black-box nature of the predictive models was addressed by explainable AI techniques. LIME provided local interpretability, while SHAP provided global explainability.

1) *Logistic Regression:* Logistic Regression is a statistical modelling technique used to predict binary outcomes by establishing a relationship between the dependent variable and one or more independent variables. In the context of student academic dropout prediction, Logistic Regression is used to estimate the probability of a student dropping out (class 1) versus Non dropout (class 0), based on academic, demographic, performance, financial support, behavioural, and psychological features. The predicted probability is converted into a binary outcome by applying a predefined threshold.

The dropout prediction score is modelled using the linear regression equation:

- Y is the binary outcome (0 = Non Dropout, 1 = Dropout)
- $X1, X2, \ldots, Xn$ represent 27 input features considered for the evaluation
- $\beta0$ is the intercept term
- $\beta1, \ldots, \beta n$ are the feature coefficients indicating the effect of each predictor

The predicted class is determined using a threshold of 0.5:

$$\hat{Y} = \begin{cases} 1 & \text{if } P(Y = 1|X) \geq 0.5 \\ 0 & \text{if } P(Y = 1|X) < 0.5 \end{cases} \quad (1)$$

TABLE II: Data Balancing Process

Stage	Technique Applied	Class 0 (Non-Dropout)	Class 1 (Dropout)	Total Records
Initial Dataset	Original Distribution	477	46	523
Stage 1	Random Up sampling	477	300	777
Stage 2	SMOTE	477	477	954

2) *XGBoost Classifier:* Extreme Gradient Boosting (XG-Boost) is an advanced ensemble machine learning algorithm based on gradient boosting decision trees. It improves predictive performance by combining multiple weak learners to produce a robust predictive model. In the student academic dropout prediction system, the XGBoost classifier effectively captures complex non-linear

relationships among academic, demographic, behavioural, financial, and psychological features, resulting in improved overall predictive accuracy com-pared to logistic regression models.

The model was trained using 200 boosting iterations, which enabled it to learn meaningful data patterns. The maximum tree depth was restricted to 5 to ensure that the model remained sufficiently complex to capture important feature interactions while preventing overfitting. A learning rate of 0.1 was selected to enable gradual learning, allowing each boosting step to correct errors made by previous models and improve overall model stability.

For a dataset $D = \{(xi, yi)\}$ containing 523 samples with 27 input features, the prediction for each sample can be expressed using the following equation:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i),\ f_k \in F \qquad (2)$$

Where:

- $\hat{Y}i$ represents the predicted output
- K represents the total number of decision trees
- fk represents each individual tree
- xi represents the input feature vector
- F represents the space of regression trees

The model minimizes the following objective function:

$$Obj = \sum_{i=1} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \qquad (3)$$

Where:

- L is the loss function measuring prediction error
- Ω is the regularization term controlling model complexity

3) *Stacked hybrid machine learning model:* To enhance the predictive accuracy of the student academic dropout prediction system, a stacked hybrid machine learning model was created. This model integrates the advantages of two learning methodologies to achieve superior prediction results. In this framework, the XGBoost classifier was chosen as the foundational model due to its impressive ability to detect intricate and non-linear relationships among academic, demographic, behavioral, financial, and psychological factors that contribute to student dropout. The XGBoost model was trained for 200 boosting iterations, with a maximum tree depth capped at five levels to maintain a suitable balance between learning effectiveness and control of overfitting. A learning rate of 0.1 was implemented to

facilitate gradual model training and ensure stable predictions. Furthermore, subsampling and column sampling methods were utilized to enhance the model's performance.

The predictions made by the XGBoost base model were subsequently utilized as inputs for the Logistic Regression meta-learner. Logistic Regression was selected as the final prediction method because of its straightforwardness, ease of interpretation, and effectiveness in managing probabilistic outcomes. In the training stage, the XGBoost model initially identified patterns from the processed data, and then its predicted probabilities were employed to train the Logistic Regression model for producing final dropout predictions. By integrating both models, the stacked hybrid approach enhances prediction reliability, strengthens model stability, and reduces bias compared to relying solely on individual machine learning models.

Mathematically, the stacked prediction can be represented as:

$$Z = f_{XGB}(X) \quad (4)$$

$$\hat{Y} = f_{LR}(Z) \quad (5)$$

Where:

- X represents the original feature set
- $fXGB(X)$ represents predictions generated by the XG-Boost model
- Z represents intermediate predictions from XGBoost
- $fLR(Z)$ represents the Logistic Regression meta-model
- $\hat{Y}$ represents the final dropout prediction

4) *Weighted Hybrid Model:* To further enhance the predictive performance of the student academic dropout prediction system, a weighted hybrid machine learning model was developed. Unlike the stacked hybrid model, the weighted hybrid model combines the probability outputs of individual base models using predefined weight distributions. In this phase, the XGBoost model was trained using the same configuration and procedure as in the stacked hybrid model, ensuring consistency in capturing complex patterns and generating reliable probability outputs. The Logistic Regression model was trained on the same processed dataset, with the maximum number of iterations set to 1000 to ensure full convergence.

Instead of assigning equal importance to both models, different weight combinations were evaluated to determine the most effective balance for predicting student dropout. Two configurations were tested: one assigning 70% weight to XG-Boost and 30% to Logistic Regression, and another assigning 60% weight to XGBoost and 40% to Logistic Regression. The probability outputs from both models were combined using weighted averaging to produce the final prediction probabilities. Mathematically, the weighted hybrid models are expressed as: Mathematically, the weighted hybrid models are expressed as:

$$\hat{Y}_{70-30} = 0.7 \times f_{XGB}(X) + 0.3 \times f_{LR}(X) \qquad (6)$$

$$\hat{Y}_{60-40} = 0.6 \times f_{XGB}(X) + 0.4 \times f_{LR}(X) \qquad (7)$$

5) *Proposed Optimized Hybrid model:* To maximize the predictive performance of the student academic dropout pre-diction system, an Optimized Hybrid Model was developed by combining XGBoost and Logistic Regression using a weighted soft voting approach. This strategy takes advantage of the strengths of both models, giving a higher contribution to the stronger learner, thereby improving overall accuracy and stability compared to using either model alone or a simple stacked approach.

In this model, the XGBoost classifier was trained with 250 boosting iterations, a maximum tree depth of five, and a learning rate of 0.08. Subsampling and column sampling rates of 0.8 were applied to enhance generalization and pre-vent overfitting. Logistic Regression was trained on the same balanced dataset, with a maximum of 1000 iterations to ensure full convergence. By training both models on the processed and balanced dataset, the system captures complex non-linear relationships through XGBoost while benefiting from the interpretable probability outputs of Logistic Regression.

For generating predictions, a weighted soft voting method was used, assigning 70% weight to XGBoost and 30% to Logistic Regression. The probability outputs from both models were combined according to this weighting, and the final class for each student was determined based on the highest combined probability. Mathematically, this can be represented as:

$$\hat{Y} = 0.7 \times f_{XGB}(X) + 0.3 \times f_{LR}(X) \qquad (8)$$

where $fXGB(X)$ and $fLR(X)$ denote the predicted probability distributions from XGBoost and Logistic Regression, respectively. This weighted approach allows the model to harness the strong predictive ability of XGBoost while incorporating the complementary insights provided by Logistic Regression, resulting in a more robust and accurate dropout prediction system.

F. *SHAP and LIME-Based Model Explainability for Student Dropout Prediction*

SHAP and LIME, two Explainable AI techniques, were used to improve the transparency and interpretability of the optimized hybrid model. While LIME explains predictions at the student level by highlighting customized risk factors, SHAP offers both global and local insights by measuring the contribution of each feature to dropout predictions across the dataset. When combined, these methods show how important features like academic engagement, stress level, financial background, attendance, and GPA—affect dropout risk. Combining SHAP and LIME enhances model transparency, assisting teachers in identifying general trends as well as the vulnerabilities of specific students. It also facilitates focused, evidence-based interventions aimed at lowering dropout.

G. *Model Evaluation*

This study presents an Optimized Hybrid Machine Learning system to predict student dropout risk with high accuracy and interpretability. The system begins with the collection and preprocessing of student data—including academic records, demographics, and performance metrics—which are cleaned, encoded, scaled, and standardized. To address the class im-balance common in dropout datasets, SMOTE combined with Random upsampling is applied, ensuring equal representation of dropout and non-dropout classes. Two complementary models are then trained: XGBoost, which captures complex non-linear relationships, and Logistic Regression, which provides interpretable probability-based predictions. The final prediction is generated using a weighted soft voting ensemble with weights of 0.7 for XGBoost and 0.3 for Logistic Regression, balancing predictive strength and interpretability.

To increase transparency, SHAP and LIME are integrated to highlight key features and explain individual predictions. The hybrid model is then deployed to assess dropout risk for new students, enabling early, data-driven interventions that target at-risk individuals and support improved retention and academic success. The working procedure for Students academic dropout system is shown in Figure 1.

H. *Evaluation Matrix*

Evaluating the performance of a student academic dropout prediction system requires more than just measuring overall accuracy, especially because dropout students typically form a smaller portion of the dataset. To ensure that the model identifies at-risk students reliably, multiple performance met-rics were considered, focusing on both overall correctness and the model's ability to detect the minority class (Dropout). The metrics used are described below in Table III.

V. EXPERIMENTAL RESULTS

A. *Implementation Methodology*

The proposed student dropout prediction system was implemented using the Google Colab environment, which provided a cloud-based platform for data preprocessing, model development, and evaluation.

B. *Performance Evaluation*

The performance comparison of the experimented machine learning models is presented in Table IV. The table illustrates the accuracy, Precision, Recall and F1-score values obtained for student dropout prediction across different models.

The confusion matrices for each of the models are depicted in Figure 2.

Figure 3 and Figure 4 show the Receiver Operating Characteristic (ROC) and Precision–Recall (PR) curves of the proposed Optimized Hybrid Model for predicting student dropout. These graphs help explain how well the model can separate students who are likely to drop out from those who are likely to continue their studies, especially when dropout cases are fewer in number. The ROC curve shows how effectively the model identifies actual dropout students while

TABLE III: Model Evaluation Matrix

Metric	Purpose	Formula
Accuracy	Measures overall correctness of the model's predictions.	Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$
Precision	Measures the proportion of students predicted as Dropout who are actually at risk.	Precision = $\frac{TP}{TP+FP}$
Recall	Measures the proportion of actual Dropout students correctly identified by the model.	Recall = $\frac{TP}{TP+FN}$
F1-Score	Harmonic mean of precision and recall, balancing false positives and false negatives.	F1 = $2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$
ROC-AUC	Area under the Receiver Operating Characteristic curve, measuring the model's discriminative ability.	ROC-AUC = $\int_0^1 TPR(FPR)\, d(FPR)$
PR-AUC	Area under the Precision-Recall curve, focusing on minority class performance.	PR-AUC = $\int \text{Precision}(\text{Recall})\, d(\text{Recall})$

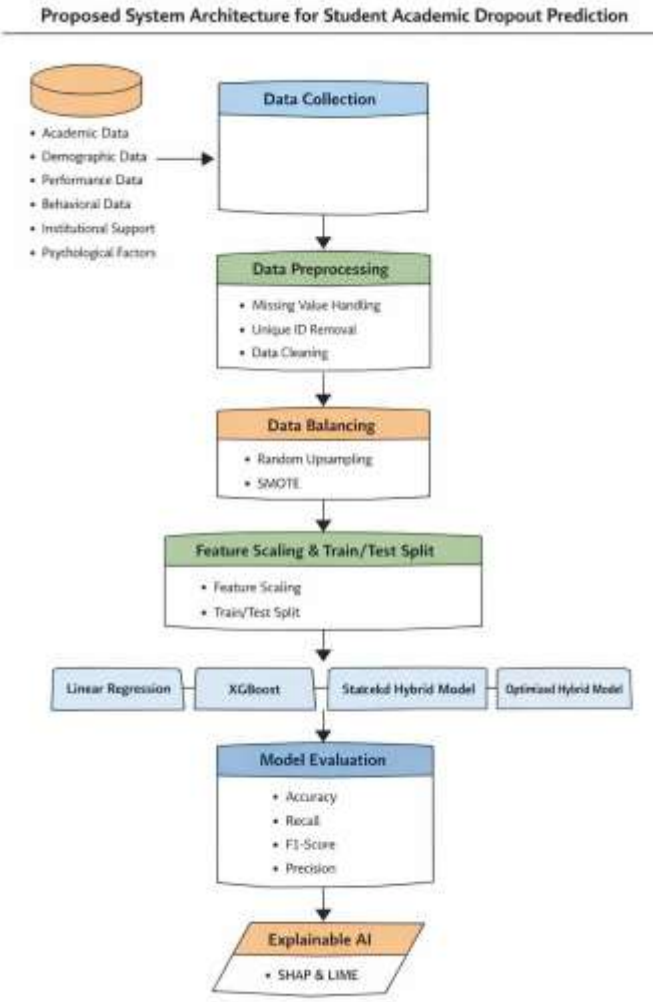


Fig. 1: Proposed system architecture for student academic dropout prediction

TABLE IV: Comparison of Model Performance

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.832	0.93	0.95	0.93
XGBoost	0.947	0.93	0.95	0.93
Stacked Hybrid	0.947	0.93	0.95	0.93
Weighted Hybrid (70:30)	0.947	0.93	0.95	0.93
Weighted Hybrid (60:40)	0.954	0.96	0.95	0.94
Optimized Hybrid	0.962	0.95	0.99	0.97

avoiding wrong predictions. The model achieved an AUC value of 0.802, which indicates good prediction ability. Since the curve stays above the random prediction line, it shows that combining XGBoost and Logistic Regression in an optimized way improves the model's overall reliability.

The PR curve shows how well the model balances precision and recall, which is very important when identifying dropout students. The model maintains good precision at lower and moderate recall levels, meaning that when it predicts a student may drop out, the prediction is usually correct. Although precision slightly decreases as the model tries to detect more dropout cases, it still maintains a good balance between identifying at-risk students and avoiding unnecessary false alerts. Overall, the Optimized Hybrid Model performs consistently and can be useful for early identification of students who may need academic support.

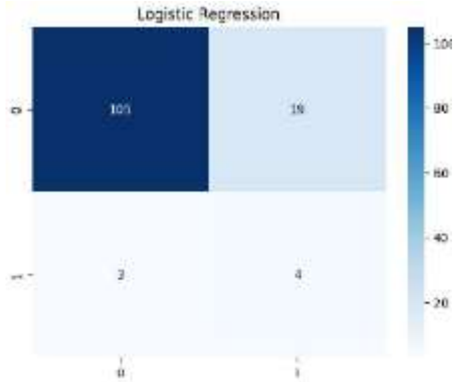
Overall, the ROC and PR analyses validate the robustness of the proposed Optimized Hybrid Model. The ROC curve demonstrates strong overall classification capability, while the PR curve confirms effective handling of class imbalance and accurate dropout detection. These results highlight the suitability of the hybrid framework for real-world student dropout prediction and early intervention systems.

C. *Model Interpretability and Explainability*

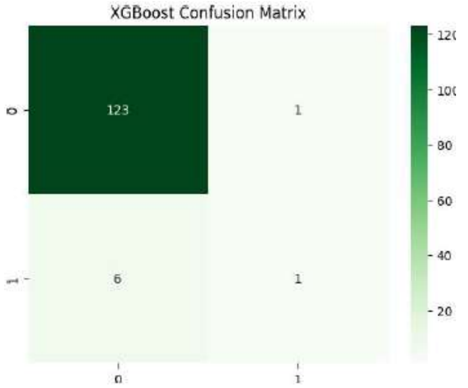
SHAP and LIME were used to interpret how the model identifies students who are at risk of dropping out. The most important factors influencing dropout predictions are attendance percentage, financial stress, mental health score, academic program, academic performance, and distance from college, according to the SHAP summary plot (Figure 5). The LIME feature contribution table (Table V) illustrates the individual effects of several factors on a student's dropout risk.

VI. RESULT AND FINDINGS

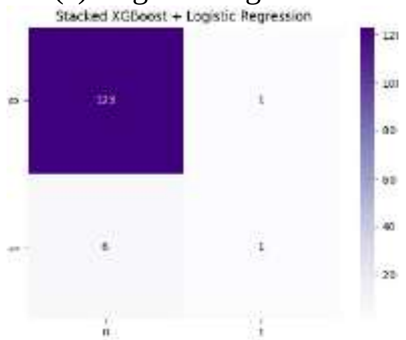
The study shows that advanced machine learning models, especially XGBoost and the proposed Optimized Hybrid Model, predict student dropout more accurately than Logistic Regression. While Logistic Regression works well for simple relationships, student dropout is influenced by many complex and interconnected factors, which XGBoost can capture better. Combining both approaches in the Optimized Hybrid Model leads to more reliable and stable predictions. Preprocessing



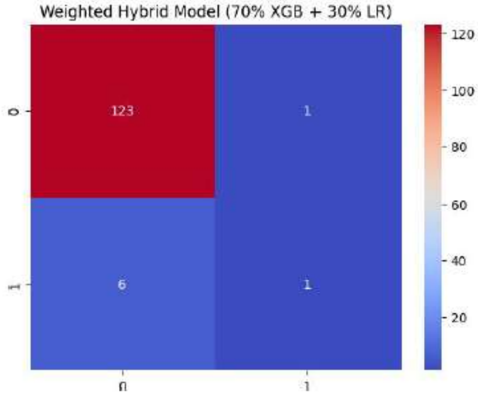
(a) Logistic Regression



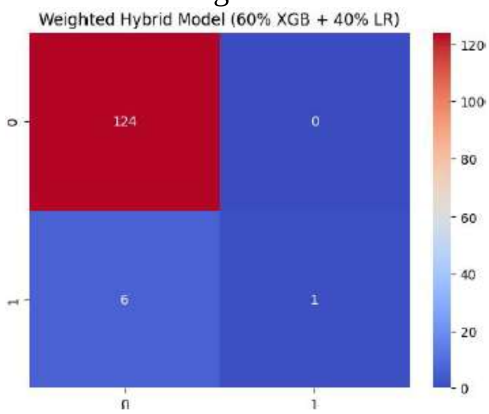
(b) XGBoost Confusion Matrix



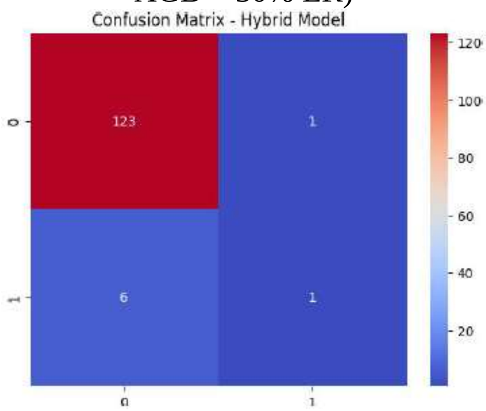
(c) Stacked XGBoost and Logistic Regression



(d) Weighted Hybrid Model (70% XGB + 30% LR)



(e) Weighted Hybrid Model (60% XGB +40% LR)



(f) Optimized Hybrid Model

Fig. 2: Confusion Matrices for various models (Logistic Regression, XGBoost, Stacked, and Weighted Hybrids)

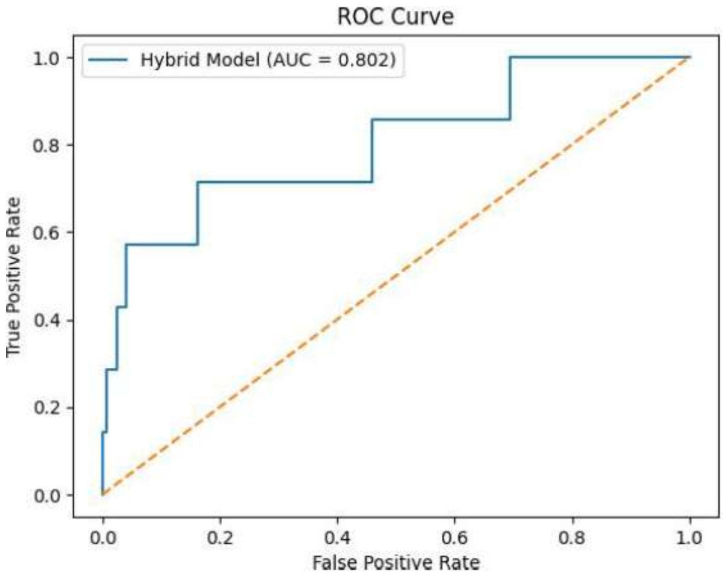


Fig. 3: ROC Curve for the Optimized Hybrid Model

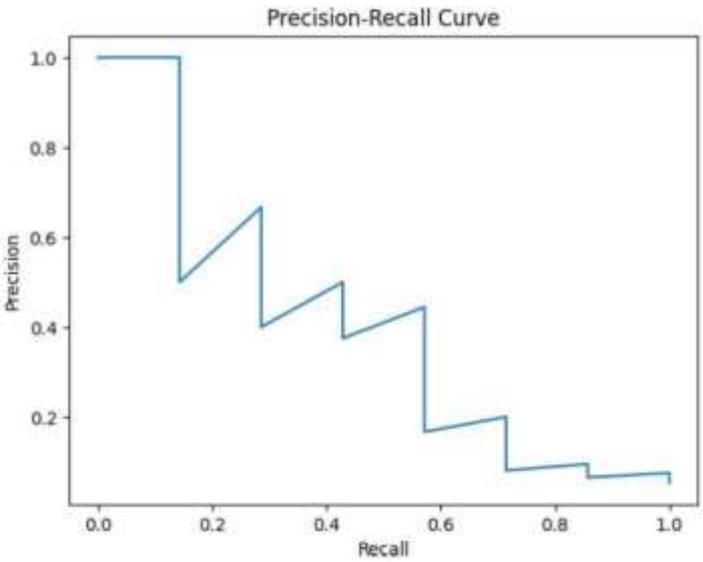


Fig. 4: Precision-Recall Curve for the Optimized Hybrid Model

steps like oversampling, SMOTE, and feature scaling also helped the models learn patterns more effectively, addressing class imbalance and ensuring all features contributed equally. Academic performance, attendance, financial support, and behavioural patterns emerged as key predictors, highlighting areas where schools can intervene early through mentoring, counselling, or support programs. However, the study is limited by a small dataset and missing factors such as family background and peer influence, which future research should include to improve generalizability

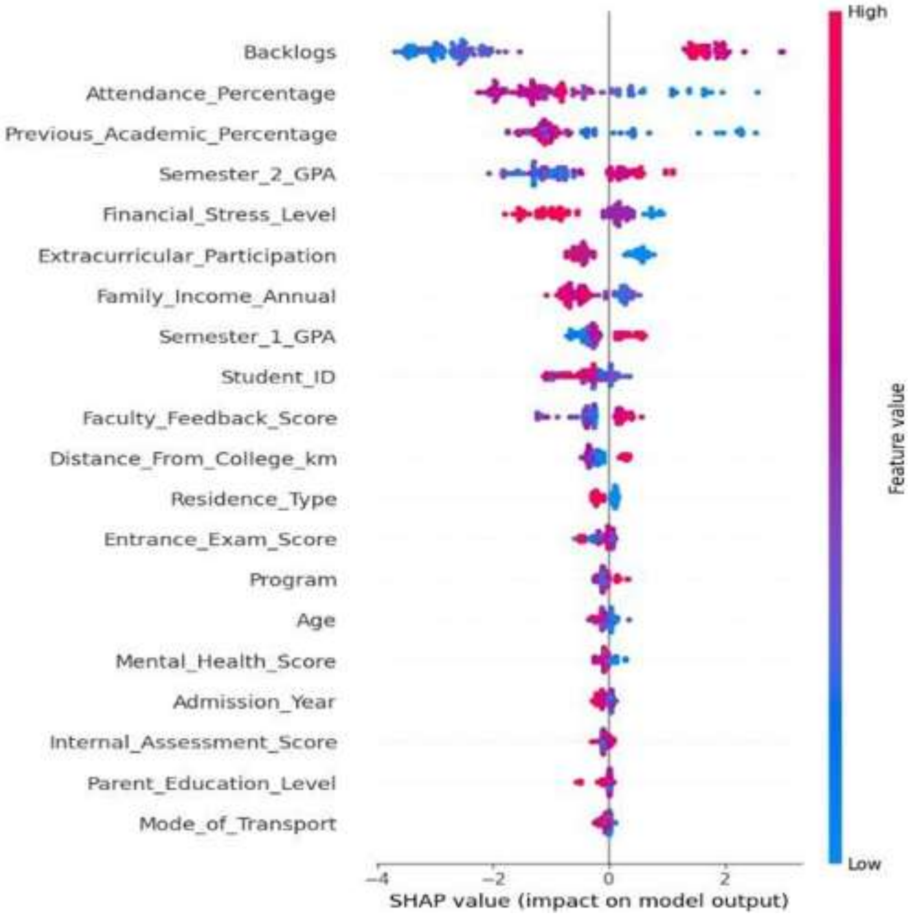


Fig. 5: SHAP EXPLAINABILITY

TABLE V: LIME EXPLAINABILITY

Feature	Value
Backlogs	-1.16
Financial_Stress_Level	-0.47
Faculty_Feedback_Score	1.38
Semester_2_GPA	0.45
Previous_Academic_Percentage	-0.1
Admission_Year	1.3
Family_Income_Annual	0.12
Placement_Anxiety_Level	1.16
Residence_Type	0.99
Socio_Economic_Status	-0.37

VII.CONCLUSION

This study demonstrates that an optimized hybrid model combining Logistic Regression and XGBoost can effectively predict student dropout, outperforming individual models in accuracy and balanced classification. By integrating data bal-ancing, feature optimization, and explainable AI techniques like SHAP and LIME, the model not only identifies at-risk students but also provides clear insights into the factors driving the predictions, such as attendance, academic performance, financial stress, and psychological well-being. These findings highlight the potential of the proposed hybrid framework as a practical and actionable tool for higher education institutions to support early intervention and improve student retention.

VIII. FUTURE ENHANCEMENT

The proposed optimized hybrid model has shown strong capability in predicting student dropout; however, it can be further enhanced by incorporating additional behavioural and psychological factors such as student motivation, engagement levels, and emotional well-being to improve prediction accuracy. Integrating the model with institutional academic or learning management systems can enable real-time monitoring of student performance and support the development of a performance improvement and feature suggestion system. Such a system can provide personalized recommendations, including academic mentoring, remedial support, counselling services, or financial assistance based on individual student risk factors. Additionally, expanding the dataset across multiple institutions, applying advanced ensemble or deep learning techniques, and enabling continuous model updates using new data can improve the reliability, scalability, and long-term effectiveness of the dropout prediction and intervention system.

References

[1] F. T. Johora et al., "An explainable AI-based approach for predicting undergraduate students' academic performance," *Array*, vol. 26, p. 100384, 2025.

[2] S. Mustofa, Y. R. Emon, S. B. Mamun, S. A. Akhter, and M. A. Hossain, "A novel AI-driven model for student dropout risk analysis with explainable AI insights," *Computers and Education: Artificial Intelligence*, vol. 8, p. 100352, 2025.

[3] R. Paul et al., "Analyzing dropout of students and an explainable predic-tion of academic performance utilizing artificial intelligence techniques," *Frontiers in Education*, vol. 10, p. 1698505, 2025.

[4] H.-C. Chen et al., "Week-wise student performance prediction in virtual learning environments using Conv-LSTM," *Applied Sciences*, vol. 12, no. 4, p. 1885, 2022.

[5] T. Liu, C. Wang, L. Chang, and T. Gu, "Predicting student performance using GRU-based deep predictive networks," *Mathematics*, vol. 10, p. 2483, 2022.

[6] P. Padmasiri and S. Kasthuriarachchi, "Interpretable prediction of student dropout using explainable AI models," in *Proc. Sri Lanka Institute of Information Technology*, Malabe, Sri Lanka, 2024.

[7] M. Segura, J. Mello, and A. Herna´ndez, "Machine learning prediction of university student dropout: Does preference play a key role?," *Mathematics*, vol. 10, p. 3359, 2022.

[8] A. Bettahi, F.-Z. Belouadha, and H. Harroud, “A modular and explain-able machine learning pipeline for student dropout prediction in higher education,” *Algorithms*, vol. 18, no. 10, p. 662, 2025.

[9] M. Martins, P. Cunha, and J. Martins, “Early prediction of student dropout using machine learning: A case study,” *Education and Infor-mation Technologies*, vol. 26, no. 3, pp. 3457–3475, 2021.

[10] J. Lee, M. Kim, D. Kim, and J.-M. Gil, “Evaluation of predictive models for early identification of dropout students,” *Journal of Information Processing Systems*, vol. 17, no. 3, pp. 630–644, 2021.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[12] G. E. Batista, R. C. Prati, and M. C. Monard, “A study of the behavior of several methods for balancing machine learning training data,” *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, 2004.

[13] H.-J. Park, Y.-S. Koo, H.-Y. Yang, Y.-S. Han, and C.-S. Nam, “Study on data preprocessing for machine learning based on semiconductor manufacturing processes,” *Sensors*, vol. 24, no. 17, 2024.

[14] J. Singthongchai and T. Wangkhamhan, “Adaptive normalization en-hances the generalization of deep learning models,” *Journal of Imaging*, vol. 12, no. 1, 2025.

[15] S. Dekker, M. Pechenizkiy, and J. Vleeshouwers, “Predicting students drop out: A case study,” *Educational Data Mining*, pp. 41–50, 2021.

[16] J. Xu and K. J. Jaggars, “Performance prediction in online learning using machine learning techniques,” *IEEE Transactions on Learning Technologies*, vol. 14, no. 3, pp. 298–310, 2022.

[17] S. Nazat, L. Li, and M. Abdallah, “XAI-ADS: An explainable artificial intelligence framework for enhancing anomaly detection in autonomous driving systems,” *IEEE Access*, vol. 12, pp. 3383431, 2024.

[18] “CODE.docx: Student Dropout Prediction Pipeline Specifications,” *In-ternal Research Document*, 2024.

[19] H. Dasi and S. Kanakala, “Student dropout prediction using machine learning techniques,” *International Journal of Intelligent Systems and Applications in Engineering*, vol. 10, no. 4, pp. 408–414, 2022.

[20] K. K. Patel and K. Amin, “Predictive modeling of dropout in MOOCs using machine learning techniques,” International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 4, pp. 436–443, 2024.

[21] H. Zhang, J. Huang, and L. Wang, “Student dropout prediction using ensemble learning methods,” IEEE Access, vol. 10, pp. 103245–103257, 2022.

[22] L. Breiman, “Stacked regressions,” Machine Learning, vol. 24, no. 1, pp. 49–64, 1996.

[23] R. Polikar, “Ensemble learning,” IEEE Circuits and Systems Magazine, vol. 6, no. 3, pp. 21–45, 2021.

[24] Z. Zhou, Ensemble Methods: Foundations and Algorithms. CRC Press, 2021.

[25] Y. K. Dwivedi et al., "Artificial intelligence in education: Applications and future directions," International Journal of Information Manage-ment, vol. 60, 2021.

[26] M. Ahmed and N. H. Abdel-Aal, "Hybrid machine learning models for student performance and dropout prediction," IEEE Access, vol. 11, pp. 45522–45535, 2023.

[27] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Mu¨ller, "Causability and explainability of artificial intelligence in education and healthcare," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 9, no. 4, pp. 1–13, 2019.

[28] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining predictions of any classifier," in Proc. ACM SIGKDD, pp. 1135–1144, 2016.

[29] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," Advances in Neural Information Processing Systems, pp. 4765–4774, 2017.

[30] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. ACM SIGKDD, pp. 785–794, 2016.

[31] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.

[32] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in Proc. 23rd Int. Conf. Machine Learning, pp. 233–240, 2006.

[33] M. Ahmed and N. H. Abdel-Aal, "Explainable hybrid machine learning models for educational dropout prediction," IEEE Access, vol. 11, pp. 45522–45535, 2023.

[34] Y. K. Dwivedi et al., "Artificial intelligence in education: Applications and future research directions," International Journal of Information Management, vol. 60, 2021.

[35] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," ACM SIGKDD Conference, 2016.

[36] H. Zhang, J. Huang, and L. Wang, "Student dropout prediction us-ing machine learning and data analytics," IEEE Access, vol. 10, pp. 103245–103257, 2022.

[37] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proc. ACM SIGKDD, 2016.

[38] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," Advances in Neural Information Processing Systems, 2017.

[39] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," Proc. ACM SIGKDD Conference, 2016.

[40] D. Hosmer, S. Lemeshow, and R. Sturdivant, Applied Logistic Regres-sion, 3rd ed., Wiley, 2018.

[41] M. Zulfiker, M. Kabir, M. A. H. Pramanik, and M. A. Rahman, "A machine learning based approach for predicting student academic performance," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 11, no. 5, pp. 313–321, 2020.

[42] T. T. Nguyen and N. T. Nguyen, "Predicting student performance based on random forest model," *International Journal of Information and Education Technology (IJIET)*, vol. 9, no. 12, pp. 875–880, Dec. 2019.