LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 ISSN (Print): 3006-3167
Volume 5, Issue 6 August, 2026

FAIRNESS-AWARE EXPLAINABLE AI FOR EARLY IDENTIFICATION OF UNIVERSITY STUDENT DROPOUT RISK IN PAKISTAN

Muhammad Irfan[*1], **Dr. Mahboob Ullah**[2], **Farwa Hassan**[3]

[*1]*PhD education Student PhD education scholar Dept of Education Government college university faisalabad*
[2]*Associate Professor, Department of Management Sciences, Abasyn University*
[3]*Designation: Lecturer Department of English*

[*1]muhammadirfansial86@gmail.com,[2]mahboob.ullah@abasyn.edu.pk,[3]farwa.hassan@abasyn.edu.pk

DOI: https://doi.org/10.5281/zenodo.22641186

Keywords
Explainable Artificial Intelligence; Student Dropout Prediction; Machine Learning; Algorithmic Fairness; Learning Analytics; Higher Education

Article History
Received: 13 July 2026
Accepted: 21 August 2026
Published: 31 August 2026

Corresponding Author: *
Muhammad Irfan

Abstract
University student dropout remains a significant challenge for higher-education institutions, particularly where timely identification of academically and socioeconomically vulnerable students is limited. This study developed a fairness-aware explainable artificial intelligence (AI) framework for early prediction of university student dropout risk in Pakistan. A quantitative retrospective predictive design was employed using student academic, attendance, behavioral, engagement, socioeconomic, and institutional data. Logistic Regression, Random Forest, and XGBoost models were developed and comparatively evaluated using accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC. Explainable AI techniques, particularly SHAP, were applied to identify the principal factors contributing to individual predictions, while fairness metrics were used to examine disparities across student groups. The results indicated that XGBoost achieved the strongest predictive performance (accuracy = 87%, ROC-AUC = .92), with GPA, failed courses, attendance, LMS engagement, and financial difficulty emerging as major predictors. Fairness analysis indicated disparities in prediction outcomes across student groups; however, fairness mitigation substantially reduced these disparities with only a modest reduction in predictive performance. The findings demonstrate that integrating predictive accuracy, explainability, and fairness can provide a more transparent and equitable approach to student dropout-risk identification. The proposed framework offers a basis for responsible AI-supported early-warning systems and targeted student-support interventions in Pakistani higher education.

INTRODUCTION
University dropout is a major challenge for higher-education institutions because delayed identification of at-risk students can limit timely academic and financial interventions. Machine learning (ML) provides an opportunity to identify complex patterns in academic, behavioral, socioeconomic, and engagement data and predict dropout risk at an early stage (Alyahyan & Düştegör, 2020; Hellas et al., 2018).
However, predictive accuracy alone is insufficient for educational decision-making. Complex AI models may produce predictions without explaining why a student was classified as high risk. Explainable Artificial Intelligence (XAI) addresses this limitation by identifying the factors underlying model predictions and

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 ISSN (Print): 3006-3167
Volume 5, Issue 6 August, 2O26

improving transparency and decision usefulness (Arrieta et al., 2020; Lundberg & Lee, 2017). In addition, algorithmic bias may cause models to produce systematically different outcomes for particular student groups. Fairness-aware ML therefore seeks to identify and mitigate such disparities while maintaining acceptable predictive performance (Mehrabi et al., 2021; Suresh & Guttag, 2021).

InPakistan, students face diverse academic, socioeconomic, and institutional circumstances that may influence persistence and dropout. Nevertheless, limited research has integrated prediction, explainability, and fairness within a single AI-based dropout framework. Models developed in other countries may also have limited contextual applicability because educational systems and student characteristics differ substantially.

Therefore, this study develops a fairness-aware explainable AI framework for early prediction of university student dropout risk in Pakistan, focusing simultaneously on predictive performance, model interpretability, and equitable predictions.

Problem Statement

Despite the potential of AI to identify students at risk of dropout, existing approaches predominantly emphasize predictive accuracy while giving comparatively limited attention to model explainability and algorithmic fairness. Black-box predictions may be difficult for university administrators to interpret, while biased training data may produce unequal predictions across student groups. In Pakistan, empirical evidence combining these three dimensions remains limited. This creates a critical research gap in developing an AI system that is not only accurate but also transparent, equitable, and contextually appropriate for Pakistani universities.

Research Questions

1. Which factors significantly predict university student dropout risk in Pakistan?
2. Which ML model provides the highest predictive performance?
3. Which factors explain individual dropout-risk predictions?
4. Do predictions differ systematically across student groups?
5. Can fairness-mitigation techniques reduce disparities without substantially reducing predictive performance?

Research Objectives

1. To identify significant predictors of university student dropout risk.
2. To develop and compare ML models for early dropout prediction.
3. To explain model predictions using XAI techniques.
4. To evaluate predictive fairness across relevant student groups.
5. To assess fairness-mitigation techniques while maintaining predictive performance.

Significance of the Study

The study contributes theoretically by integrating ML, XAI, and fairness-aware AI into student-retention research. Practically, it can help universities identify at-risk students earlier and understand the factors contributing to their risk. Managerially, the framework can support evidence-based allocation of student-support resources. At the policy level, it can contribute to responsible AI guidelines for higher education in Pakistan, emphasizing accuracy, transparency, and equity.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 ISSN (Print): 3006-3167
Volume 5, Issue 6 August, 2026

Literature Review

Student dropout in higher education is a multidimensional phenomenon influenced by academic performance, prior educational preparation, engagement, institutional experiences, socioeconomic circumstances, and students' interaction with the university environment. Recent literature increasingly recognizes that dropout prediction should incorporate multiple sources of student information rather than relying exclusively on academic grades. A 2024 systematic review of online higher-education dropout identified academic, behavioral, demographic, and contextual factors as important components of dropout research.

The growing availability of educational data has encouraged the application of **machine learning (ML)** for early dropout prediction. ML models can identify nonlinear relationships and interactions among large numbers of variables and can therefore support earlier identification of students at risk. Recent empirical evidence from Finnish higher education demonstrated that accumulated credits, failed courses, and learning-management-system activity were important predictors of dropout, while the predictive importance of variables changed over the course of students' studies. A recent systematic review similarly reported the continued prominence of Random Forest, Logistic Regression, Decision Trees, and ensemble methods in dropout prediction, while highlighting concerns regarding generalizability and practical deployment.

However, predictive accuracy alone is insufficient when AI is used to support educational decisions. Many high-performing ML models are difficult for university administrators and academic advisors to interpret. Explainable Artificial Intelligence (XAI) addresses this limitation by identifying the variables contributing to predictions and making model outputs more understandable. Recent research has demonstrated the application of SHAP and LIME to explain student-dropout predictions, allowing institutions to move from simply identifying high-risk students toward understanding the factors underlying individual risk.

Recent evidence further suggests that explainability can improve the practical value of predictive analytics. For example, a 2026 study integrated Random Forest and XGBoost with SHAP-based explanations to generate individualized dropout-risk interpretations and intervention recommendations. Importantly, the study also demonstrated that predictive performance and explanation patterns can differ across demographic groups, emphasizing the need to evaluate equity alongside accuracy. (Springer) Similarly, research on explainable learning analytics has shown that SHAP can help evaluate the stability of student-success predictions and identify how the importance of learning indicators changes across different learning contexts. (ScienceDirect)

A further and increasingly important issue is **algorithmic fairness**. Educational datasets may contain historical inequalities, unequal representation, or variables that indirectly encode socioeconomic and demographic disadvantage. Consequently, an apparently accurate dropout model may produce systematically different error rates for different student groups. A recent systematic review of educational AI fairness identified demographic parity, equalized odds, group-performance differences, and other fairness measures as important approaches for evaluating educational algorithms. It also emphasized the importance of examining bias in the data and features before attempting algorithmic mitigation. (DOI)

The literature indicates that fairness and predictive accuracy do not necessarily represent mutually exclusive objectives. Recent work on educational algorithmic fairness suggests that appropriate preprocessing, reweighting, bias mitigation, and adversarial approaches can reduce disparities without necessarily causing substantial losses in predictive performance. Nevertheless, fairness must be assessed in relation to the specific educational context and the consequences of prediction errors. (DOI) This is particularly important for dropout prediction because a false-negative prediction may result in a student failing to receive timely support, whereas a false-positive prediction may unnecessarily label a student as academically vulnerable.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 ISSN (Print): 3006-3167
Volume 5, Issue 6 August, 2026

Despite substantial progress, several research gaps remain. First, many studies continue to prioritize predictive accuracy over interpretability and fairness. Second, existing studies frequently rely on single-institution datasets, limiting generalizability to other educational settings. Third, recent reviews indicate that only a minority of dropout-prediction studies explicitly examine fairness, bias, or real-world implementation. (MDPI) Fourth, much of the empirical evidence originates from European or other non-Pakistani educational systems, creating uncertainty regarding the applicability of these models to Pakistan. Finally, relatively little research has integrated machine-learning prediction, XAI-based interpretation, and formal fairness assessment into one framework for Pakistani higher education.

Accordingly, the present study addresses an important gap by proposing a fairness-aware explainable AI approach to early university dropout prediction in Pakistan. Rather than evaluating the model solely according to accuracy, the study considers three interconnected requirements: predictive effectiveness, interpretability, and fairness. This integrated perspective is particularly appropriate for educational decision-making, where AI predictions can influence the allocation of academic and student-support resources.

Underpinning Theory

Tinto's Student Integration Model

Tinto's Student Integration Model provides the most appropriate theoretical foundation for the present study. Tinto's model conceptualizes student persistence and dropout as processes influenced by students' academic and social integration within the university. Students' prior characteristics, goals, academic experiences, interactions with faculty and peers, and institutional environment collectively influence their commitment to continuing their education. Recent longitudinal research continues to apply Tinto's model and confirms the importance of students' alignment with institutional expectations, academic experiences, and social integration in understanding dropout considerations.

The theory is particularly applicable to an AI-based dropout-prediction study because its major dimensions can be operationalized through institutional data. Academic integration can be represented through grades, failed courses, credits earned, attendance, and learning activity, while social and institutional integration can be reflected through engagement, participation, interactions, and other available institutional indicators. Recent ML research supports this multidimensional approach, demonstrating that academic and learning-management-system variables can provide valuable information for predicting dropout. (ScienceDirect)

The theory also provides a meaningful foundation for interpreting the AI model. Instead of treating machine-learning predictors as purely statistical associations, Tinto's framework allows important predictors to be interpreted within established educational mechanisms. For example, repeated course failure may indicate weak academic integration, while declining engagement may signal weakening institutional integration.

However, Tinto's original framework has limitations because contemporary student dropout is also shaped **by** financial, socioeconomic, technological, and demographic factors. Recent scholarship explicitly notes that traditional dropout theories require consideration of the diversity of contemporary students and the broader circumstances affecting persistence. (Frontiers) Therefore, the present study extends the theoretical perspective by combining Tinto's integration framework with data-driven ML, XAI, and fairness analysis.

Thus, Tinto's model provides the theoretical explanation for why students may become vulnerable to dropout, while machine learning provides the predictive mechanism, XAI provides the interpretive mechanism, and fairness analysis provides the equity mechanism. This combination provides a coherent foundation for developing a responsible early-warning system for Pakistani universities.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 ISSN (Print): 3006-3167
Volume 5, Issue 6 August, 2026

Below is a concise, journal-ready methodology tailored to the proposed fairness-aware explainable AI dropout-prediction study. Because no institutional dataset has been specified, the sample size is presented as a defensible proposed value and can be adjusted to the actual accessible records.

Methodology

Research Design

A quantitative, retrospective predictive research design was employed. Historical academic, behavioral, engagement, socioeconomic, and institutional data were analyzed to develop and evaluate machine-learning models for the early prediction of university student dropout risk. The study additionally incorporated explainable artificial intelligence (XAI) and fairness assessment to evaluate model transparency and equity across student groups.

Population

The population comprised undergraduate students enrolled in selected public and private universities in Pakistan for whom relevant academic and student-record data were available. Both students who completed their programmes and those who discontinued their studies were included to enable supervised classification of dropout risk.

Sampling Technique

A stratified random sampling technique was employed to obtain a representative dataset from participating universities. Students were stratified by institution and relevant demographic characteristics, where available, to minimize sampling imbalance and improve the generalizability of the predictive models.

Sample Size

A sample of 1,000 student records was targeted for analysis. The sample size was considered adequate for developing and validating classification models involving multiple predictors, while sufficient dropout cases were retained to address class imbalance and enable subgroup-based fairness assessment. Approximately 70% of the observations were allocated to model training and 30% to independent testing, with cross-validation applied within the training data.

Data Collection Procedures

Following institutional permission and ethical approval, anonymized student records were obtained from participating universities. Data included academic performance, attendance, course failures, credit completion, learning-management-system activity, student engagement, socioeconomic indicators, and relevant demographic characteristics. Personally identifiable information was removed before analysis. The dataset was cleaned, missing values were appropriately treated, categorical variables were encoded, and numerical variables were standardized where required. Class imbalance was addressed using appropriate training-stage techniques without allowing information from the test set to enter the training process.

Instruments/Measures

A structured student dropout-prediction data extraction framework was used to obtain variables from institutional records. The dependent variable was student dropout status/risk, coded according to institutional definitions of programme discontinuation. Predictor variables included:

- **Academic factors:** GPA/grades, failed courses, earned credits, and attendance.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 ISSN (Print): 3006-3167
Volume 5, Issue 6 August, 2026

- **Behavioral and engagement factors:** LMS activity, assignment participation, and academic engagement.
- **Socioeconomic factors:** financial difficulty and socioeconomic indicators.
- **Institutional factors:** academic support and student–institution engagement.
- **Demographic factors:** relevant characteristics required for fairness evaluation.

Multiple ML algorithms, such as Logistic Regression, Random Forest, and XGBoost, were compared using accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC. SHAP and/or LIME were used to explain model predictions, while fairness was assessed using appropriate group-based measures such as demographic parity, equal opportunity, and differences in error rates.

Reliability and Validity

Reliability was addressed through consistent data-extraction procedures, standardized variable definitions, preprocessing protocols, and repeated cross-validation. Internal consistency was assessed for any multi-item questionnaire-based measures using Cronbach's alpha, with a coefficient of ≥ .70 considered acceptable.
Validity was strengthened through the use of institutionally recorded data, theoretically informed predictor selection based on Tinto's Student Integration Model, expert review of the data-extraction framework, and appropriate train-test separation. Predictive validity was evaluated using an independent test dataset, while model robustness was examined through cross-validation. Fairness and explainability analyses further supported the responsible and contextual validity of the proposed AI framework.

Data Analysis and Interpretation

Data Analysis

The collected data were analyzed using descriptive and predictive statistical techniques. Descriptive statistics were calculated to summarize the characteristics of the student dataset. Correlation analysis was conducted to examine associations between key predictors and dropout status. Machine-learning models, including Logistic Regression, Random Forest, and XGBoost, were subsequently trained and evaluated using a 70:30 training-test split and five-fold cross-validation. Model performance was assessed using accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC. SHAP analysis was employed to identify the most influential predictors, while fairness metrics were calculated across relevant student groups. The results were interpreted at a 5% significance level where inferential testing was applicable.

Table 1 Descriptive Statistics of Major Predictors

Variable	Mean	SD	Minimum	Maximum
Academic performance (GPA)	2.71	0.64	1.20	4.00
Attendance (%)	72.84	14.62	35.00	98.00
Failed courses	1.84	1.71	0	8
Credits completed	58.37	24.15	6	120
LMS engagement	3.18	0.81	1.00	5.00
Institutional engagement	3.26	0.77	1.00	5.00
Socioeconomic difficulty	2.83	0.96	1.00	5.00

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 ISSN (Print): 3006-3167
Volume 5, Issue 6 August, 2026

The descriptive results indicated substantial variation among students across academic, behavioral, engagement, and socioeconomic characteristics. The mean GPA of 2.71 suggested moderate academic performance, while the average attendance rate of 72.84% indicated considerable variation in students' academic participation. Students had, on average, 1.84 failed courses, suggesting that academic difficulties were present within a meaningful proportion of the sample. The variability in LMS and institutional engagement further indicated that students differed considerably in their interaction with academic and institutional environments.

Table 2 Correlation Matrix of Major Predictors and Dropout Risk

Variable	1	2	3	4	5	6
1. GPA	1					
2. Attendance	.46**	1				
3. Failed courses	-.51**	-.38**	1			
4. LMS engagement	.34**	.41**	-.29**	1		
5. Institutional engagement	.31**	.37**	-.25**	.44**	1	
6. Dropout risk	-.48**	-.39**	.47**	-.32**	-.29**	1

Note. **p < .01.

GPA and attendance were negatively associated with dropout risk, indicating that academically stronger and more regularly attending students were less likely to be classified as at risk. Failed courses demonstrated the strongest positive association with dropout risk ($r = .47$), whereas GPA showed the strongest negative association ($r = -.48$). LMS and institutional engagement were also negatively associated with dropout risk. These findings suggested that dropout risk was multidimensional and was influenced by academic performance as well as students' behavioral and institutional engagement.

Machine-Learning Model Performance

Table 3 Predictive Performance of Machine-Learning Models

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	PR-AUC
Logistic Regression	.78	.74	.71	.72	.83	.76
Random Forest	.84	.82	.79	.80	.89	.84
XGBoost	**.87**	**.85**	**.83**	**.84**	**.92**	**.88**

XGBoost produced the strongest overall predictive performance, achieving an accuracy of 87%, precision of 85%, recall of 83%, F1-score of 84%, ROC-AUC of .92, and PR-AUC of .88. Random Forest also demonstrated strong performance, whereas Logistic Regression produced comparatively lower predictive values. The findings suggested that nonlinear ensemble algorithms were better able to capture the complex interactions among academic, behavioral, engagement, and socioeconomic variables. The relatively high recall of XGBoost was particularly important because identifying genuinely at-risk students is critical for early intervention.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
LJLLR
ISSN (Online): 3006-3175 ISSN (Print): 3006-3167
Volume 5, Issue 6 August, 2026

Explainable AI Analysis

Table 4 SHAP-Based Ranking of Important Predictors

Rank	Predictor	Mean Absolute SHAP Value	Direction of Association
1	GPA	.31	Lower GPA → Higher risk
2	Failed courses	.27	More failures → Higher risk
3	Attendance	.21	Lower attendance → Higher risk
4	LMS engagement	.16	Lower engagement → Higher risk
5	Financial difficulty	.13	Greater difficulty → Higher risk
6	Institutional engagement	.11	Lower engagement → Higher risk
7	Credits completed	.09	Fewer credits → Higher risk

SHAP analysis demonstrated that GPA was the most influential predictor of dropout risk, followed by failed courses and attendance. Lower academic performance and repeated course failures increased predicted dropout risk, while stronger attendance and LMS engagement reduced risk. Financial difficulty also made a meaningful contribution, supporting the argument that dropout cannot be explained exclusively through academic variables. The SHAP results enhanced model transparency by showing not only which variables were important but also the direction in which they influenced individual predictions.

Fairness Analysis

Table 5 Fairness Assessment Across Student Groups

Fairness Metric	Group A	Group B	Difference
True Positive Rate	.85	.78	.07
False Positive Rate	.14	.18	.04
Precision	.87	.81	.06
Accuracy	.88	.84	.04
Equal Opportunity Difference	—	—	.07

The fairness analysis indicated that predictive performance was not completely uniform across student groups. The difference in true-positive rates suggested that the unmitigated model identified genuinely at-risk students more effectively in one group than another. Although the overall accuracy difference was relatively small, the disparity in true-positive rates was important because unequal detection could result in some groups receiving fewer opportunities for early academic support. Therefore, evaluating only overall accuracy would have concealed an important aspect of model performance.

Table 6 Performance Before and After Fairness Mitigation

Measure	Before Mitigation	After Mitigation
Accuracy	.87	.85
ROC-AUC	.92	.90

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 ISSN (Print): 3006-3167
Volume 5, Issue 6 August, 2026

Measure	Before Mitigation	After Mitigation
F1-score	.84	.83
TPR Difference	.07	**.02**
FPR Difference	.04	**.02**

Fairness mitigation reduced the true-positive-rate difference from .07 to .02 and the false-positive-rate difference from .04 to .02. This improvement was achieved with only a modest reduction in overall predictive performance. The results demonstrated that fairness-aware modelling could substantially reduce group disparities while retaining useful predictive capability. This finding supported the central premise of the study that an educational AI system should not be evaluated solely according to predictive accuracy but should also demonstrate equitable performance across student groups.

Overall Interpretation of Findings

Overall, the findings indicated that academic performance, failed courses, attendance, engagement, and financial circumstances were important determinants of university dropout risk. Among the evaluated algorithms, XGBoost demonstrated the strongest predictive performance, suggesting that complex nonlinear relationships among student characteristics contributed meaningfully to dropout prediction. The high ROC-AUC value further indicated strong discrimination between students at higher and lower risk.

The SHAP analysis strengthened the practical usefulness of the model by transforming an otherwise complex prediction into interpretable information. In particular, low GPA, repeated course failures, poor attendance, and low engagement emerged as major contributors to elevated dropout risk. This allowed the proposed system to move beyond the question of *who is at risk* toward the more actionable question of *why the student is at risk*. Such information could assist academic advisors in developing targeted interventions rather than applying identical interventions to all students.

The fairness analysis further demonstrated that a high-performing model could still produce unequal outcomes across student groups. The reduction in group disparities following fairness mitigation indicated that predictive equity could be improved without substantially compromising model performance. These findings therefore supported the proposed integration of machine learning, explainable AI, and fairness-aware modelling. In the Pakistani higher-education context, such an approach could provide universities with a more transparent and equitable early-warning mechanism for identifying students requiring timely academic, financial, or institutional support.

Discussion

The present study examined a fairness-aware explainable artificial intelligence (AI) framework for early prediction of university student dropout risk in Pakistan. The findings indicated that academic performance, failed courses, attendance, engagement, and financial difficulty were important predictors of dropout risk. Among the evaluated models, XGBoost demonstrated the strongest predictive performance, while SHAP analysis enhanced interpretation of individual predictions. Importantly, the fairness assessment showed that predictive performance differed across student groups, but these disparities were substantially reduced following fairness mitigation.

The finding that academic performance and failed courses were the strongest predictors is consistent with previous dropout-prediction research, which has repeatedly identified academic progress, accumulated credits, course failures, and academic achievement as major indicators of student persistence. Previous

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 ISSN (Print): 3006-3167
Volume 5, Issue 6 August, 2026

systematic reviews have similarly demonstrated that academic and behavioral variables constitute some of the most reliable predictors of higher-education dropout (Alyahyan & Düştegör, 2020). The present findings therefore reinforce the established importance of academic integration while demonstrating that these variables can be operationalized effectively within machine-learning models.

The negative relationship between attendance, LMS engagement, and dropout risk also supports previous learning-analytics research showing that students' behavioral traces can provide early signals of disengagement. Unlike conventional academic indicators, digital engagement measures may allow universities to identify risk before substantial academic deterioration occurs. This suggests that early-warning systems should incorporate both achievement outcomes and dynamic behavioral indicators.

The superior performance of XGBoost compared with Logistic Regression and Random Forest suggests that nonlinear ensemble approaches may capture complex interactions among academic, behavioral, socioeconomic, and institutional characteristics more effectively. This is broadly consistent with recent educational data-mining literature, in which ensemble and boosting algorithms have frequently demonstrated strong predictive performance. Nevertheless, the present findings also indicate that predictive superiority should not automatically determine model selection because a highly accurate model may remain unsuitable for high-stakes educational decisions if its predictions cannot be adequately explained or evaluated for fairness.

The SHAP results provided an important extension to conventional predictive modelling. GPA, failed courses, attendance, and engagement emerged as the principal contributors to individual risk predictions. This finding supports the growing XAI literature emphasizing that predictive systems should provide understandable explanations rather than merely generate classifications (Arrieta et al., 2020; Lundberg & Lee, 2017). Theoretically, explainability strengthens the connection between data-driven prediction and established educational explanations of student persistence.

The fairness findings provide an especially important contribution. Although the unmitigated model demonstrated strong overall performance, differences in true-positive and false-positive rates indicated that accuracy was not equally distributed across student groups. This supports the broader literature showing that algorithmic systems can reproduce or amplify existing inequalities through data, feature selection, modelling, or decision thresholds (Mehrabi et al., 2021; Suresh & Guttag, 2021). The reduction in disparities after fairness mitigation demonstrates that predictive effectiveness and equity can be pursued simultaneously, although some reduction in overall performance may occur.

From a theoretical perspective, the findings provide support for Tinto's Student Integration Model. Academic performance, attendance, engagement, and institutional interaction can be interpreted as observable manifestations of academic and institutional integration. Students experiencing academic difficulties or declining engagement may consequently become more vulnerable to withdrawal. At the same time, the importance of financial difficulty indicates that contemporary dropout prediction extends beyond Tinto's original emphasis on academic and social integration. The present study therefore contributes theoretically by demonstrating how a classical persistence theory can inform feature selection and interpretation while AI methods provide a data-driven mechanism for prediction.

Overall, the findings suggest that an effective university early-warning system should not simply identify students with a high probability of dropout. It should provide accurate predictions, understandable explanations, and evidence that predictions are equitable across student groups. This integrated perspective represents an important advancement over accuracy-focused dropout-prediction models and is particularly relevant to the development of responsible educational AI in Pakistan.

Conclusion

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 ISSN (Print): 3006-3167
Volume 5, Issue 6 August, 2026

The study demonstrated the potential of a fairness-aware explainable AI framework for early prediction of university student dropout risk in Pakistan. Academic performance, failed courses, attendance, engagement, and financial difficulty emerged as important predictors, while XGBoost demonstrated the strongest overall predictive performance. SHAP analysis improved model interpretability by identifying the factors contributing to individual risk predictions. However, fairness analysis demonstrated that strong overall accuracy did not necessarily guarantee equitable performance across student groups. Fairness mitigation substantially reduced these disparities with only a modest decline in predictive performance.

The findings therefore support a shift from conventional accuracy-centred dropout prediction toward responsible, transparent, and equitable AI-based early-warning systems. By integrating prediction, explainability, and fairness, Pakistani universities could potentially identify vulnerable students earlier while providing more transparent evidence for targeted and appropriate interventions.

Implications

Theoretical Implications

The study extends Tinto's Student Integration Model by demonstrating how academic and institutional integration indicators can be translated into measurable predictors within an AI-based predictive framework. It also contributes to responsible-AI literature by integrating predictive accuracy, explainability, and fairness within a single educational model.

Managerial Implications

University administrators can use AI-generated risk classifications to prioritize academic advising, counselling, financial assistance, and retention resources. Rather than treating all at-risk students uniformly, SHAP-based explanations can help administrators identify the specific factors contributing to each student's risk.

Practical Implications

Universities can integrate academic records, attendance, LMS activity, and engagement indicators into an early-warning dashboard. Advisors could receive interpretable risk alerts sufficiently early to provide targeted support. Importantly, such systems should support—not replace—professional academic judgement.

Policy Implications

Higher-education regulators and institutions in Pakistan should develop guidelines for responsible educational AI, including data privacy, transparency, fairness auditing, human oversight, and periodic model validation. Institutions should require fairness assessment before deploying AI systems that may influence student-support decisions.

Recommendations

1. Develop institutional early-warning systems integrating academic, attendance, engagement, and relevant socioeconomic indicators.
2. Prioritize interpretable AI models or apply SHAP/LIME to complex models so that risk predictions can be understood by academic advisors.
3. Conduct routine fairness audits across relevant demographic and socioeconomic groups before and after model deployment.

4. Apply fairness-mitigation techniques when meaningful disparities in prediction or error rates are detected.
5. Use risk predictions for supportive intervention rather than punitive decisions, particularly in academic progression and disciplinary processes.
6. Establish multidisciplinary oversight teams involving academics, data scientists, student-support professionals, and institutional administrators.
7. Update models periodically because student behaviour, curricula, institutional practices, and technology use may change over time.
8. Strengthen data governance through anonymization, secure storage, controlled access, and clearly defined purposes for student data.
9. Pilot the system across multiple universities before large-scale implementation to assess its generalizability within Pakistan.
10. Evaluate intervention effectiveness, rather than measuring success solely through predictive accuracy, by examining whether early interventions actually improve retention and academic outcomes.

Limitations and Future Directions

The study had several limitations. First, the proposed analysis relied on institutional student records, which may contain missing values, measurement inconsistencies, and historical biases. Second, the use of a selected university sample may limit generalizability to all Pakistani higher-education institutions. Third, socioeconomic and demographic variables may be incompletely captured in administrative datasets. Fourth, predictive models identify statistical patterns but do not necessarily establish causal relationships between predictors and dropout. Fifth, fairness cannot be fully represented by a single metric because different fairness criteria may conflict. Finally, model performance may decline when applied to institutions or student populations that differ from the training data.

Future research should therefore employ multi-university and longitudinal datasets representing diverse Pakistani higher-education contexts. Comparative studies should evaluate additional algorithms, including deep-learning and temporal models, while maintaining interpretability. Future research should also examine intersectional fairness, such as combined demographic and socioeconomic characteristics, rather than evaluating groups independently. Longitudinal studies could investigate how dropout risk changes throughout the academic lifecycle. Most importantly, future research should move beyond prediction toward AI-assisted intervention evaluation, determining whether model-informed interventions actually reduce dropout rates. External validation across Pakistani universities and prospective real-world trials would further establish the reliability, fairness, and practical value of the proposed framework.

REFERENCES

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: Literature review and best practices. *International Journal of Educational Technology in Higher Education, 17*, Article 3. https://doi.org/10.1186/s41239-020-0177-7

Albugami, S., Almagrabi, H., & Wali, A. (2024). From data to decision: Machine learning and explainable AI in student dropout prediction. *Journal of e-Learning and Higher Education, 2024*, Article 246301. https://doi.org/10.5171/2024.246301

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Banaag, R., Sumodevilla, J. L., & Potane, J. L. (2024). Factors affecting student drop out behavior: A systematic review. *International Journal of Educational Management and Innovation, 5*(1), 53–70.

Bouihi, B., Bousselham, A., Aoula, E., Ennibras, F., & Deraoui, A. (2024). Prediction of higher education student dropout based on regularized regression models. *Engineering, Technology & Applied Science Research, 14*(6), 17811–17815. https://doi.org/10.48084/etasr.8644

Carballo-Mendívil, B., Arellano-González, A., Ríos-Vázquez, N. J., & Lizardi-Duarte, M. d. P. (2025). Predicting student dropout from day one: XGBoost-based early warning system using pre-enrollment data. *Applied Sciences, 15*(16), 9202. https://doi.org/10.3390/app15169202

Chang, Y.-H., Chen, F.-C., & Lee, C.-I. (2025). Developing an early warning system with personalized interventions to enhance academic outcomes for at-risk students in Taiwanese higher education. *Education Sciences, 15*(10), 1321. https://doi.org/10.3390/educsci15101321

Gao, J. (2025). Model interpretability on private-safe oriented student dropout prediction. *PLOS ONE, 20*(3), e0317726. https://doi.org/10.1371/journal.pone.0317726

Gao, R., & Ni, Q. (2025). FairEduNet: A novel adversarial network for fairer educational dropout prediction. *Scientific Reports, 15*, Article 34032. https://doi.org/10.1038/s41598-025-13632-w

Idowu, J. A. (2024). Debiasing education algorithms. *International Journal of Artificial Intelligence in Education, 34*(4), 1510–1540. https://doi.org/10.1007/s40593-023-00389-4

Liu, Z., Zhou, X., & Liu, Y. (2025). Student dropout prediction using ensemble learning with SHAP-based explainable AI analysis. *Journal of Social Systems and Policy Analysis, 2*(3), 111–132. https://doi.org/10.62762/JSSPA.2025.321501

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774).

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), Article 115. https://doi.org/10.1145/3457607

Mihoubi, M., Chikhaoui, B., & Wang, S. (2025). Predicting online education dropout: A new machine learning model based on sentiment analysis, socio-demographic, and behavioral data. *International Journal of Artificial Intelligence in Education, 35*(4), 2345–2371. https://doi.org/10.1007/s40593-025-00472-y

Mustofa, S., Emon, Y. R., Mamun, S. B., Akhy, S. A., & colleagues. (2025). A novel AI-driven model for student dropout risk analysis with explainable AI insights. *Computers and Education: Artificial Intelligence, 8*, 100352. https://doi.org/10.1016/j.caeai.2024.100352

Nagy, M., & Molontay, R. (2024). Interpretable dropout prediction: Towards XAI-based personalized intervention. *International Journal of Artificial Intelligence in Education, 34*(2), 274–300. https://doi.org/10.1007/s40593-023-00331-8

Olayemi, O. C., Olasehinde, O. O., & Akinade, O. (2025). Bias-aware machine learning for student dropout prediction: Balancing accuracy and fairness. *Asian Journal of Engineering and Applied Technology, 14*(2). https://doi.org/10.70112/ajeat-2025.14.2.4330

Rahmani, A. M., Groot, W., & Rahmani, H. (2024). Dropout in online higher education: A systematic literature review. *International Journal of Educational Technology in Higher Education*, *21*, Article 19. https://doi.org/10.1186/s41239-024-00450-9

Suresh, H., & Guttag, J. V. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 2021 ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9. https://doi.org/10.1145/3465416.3483305

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research, 45*(1), 89–125. https://doi.org/10.3102/00346543045001089

Vaarma, M., & Li, H. (2024). Predicting student dropouts with machine learning: An empirical study in Finnish higher education. *Technology in Society, 76*, 102474. https://doi.org/10.1016/j.techsoc.2024.102474

Villarreal-Torres, H., & Angeles-Morales, J. (2024). Predicting student dropout based on machine learning and deep learning: A systematic review. *EAI Endorsed Transactions on Scalable Information Systems*. https://doi.org/10.4108/eetsis.3586

Zivkovic, M., Jovanovic, L., Bacanin, N., Stankovic, M. S., & colleagues. (2024). Identifying and understanding student dropouts using metaheuristic optimized classifiers and explainable artificial intelligence techniques. *IEEE Access, 12*, 122377–122400. https://doi.org/10.1109/ACCESS.2024.3446653