Scientific Reports

https://doi.org/10.1038/s41598-026-47278-z

Article in Press

A machine learning based framework for predictive school management using student and faculty analytics

Received: 31 October 2025

Accepted: 31 March 2026

Published online: 04 April 2026

Cite this article as: Yang M., Li Z. & Liu S. A machine learning based framework for predictive school management using student and faculty analytics. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-47278-z

Ming Yang, Zhe Li & Shaoyan Liu

A Machine Learning Based Framework for Predictive School Management Using Student and Faculty Analytics

Ming Yang[1,*], Zhe Li[2], Shaoyan Liu[3]

[1]Department of Science and Education, Aviation General Hospital, No. 3, Beiyuan Road, Chaoyang District, Beijing, 100012, China

[2]Discipline Inspection Office, Capital Medical University Affiliated Beijing Hospital of Traditional Chinese, Beijing, 100010, China

[3]Department of Educational Teaching Quality Construction, Chinese Research Academy of Traditional Chinese Medicine College, Beijing, 100700, China

Corresponding author: Ming Yang, Email: 13718330851@163.com

Abstract: New technologies in education have created a huge amount of data that, when used effectively, can have a major impact on the functioning of an institution and the academic achievement of students. Nevertheless, all existing predictive models are still disconnected and do not integrate historical trends, student-faculty relationships, and trend patterns into a coherent decision-making system. The paper describes an integrated machine learning system that integrates several synergistic AI technologies: (1) deep learning systems (LSTM, GRU, CNN, and Transformers) to model academic growth over time; (2) comprehensible gradient boosting ensembles (XGBoost, LightGBM, and CatBoost) to understandably infer and analyze structured data. (3) graph convolutional networks (GCNs) to encode academic relationships between students, professors, and courses; and (4) data-centric oriented approaches (multitasking, transfer, and federated learning). The framework is tested on two UCI benchmark datasets (n=649) with fully isolated holdout sets using strict nested cross-validation to prevent data leakage. The framework yields 99.6% and 97.5% predictive accuracy (5.6% and 6.3% improvement over the top baselines) and high recall (99.4% and 96.7%) in classifying at-risk students. Each component has been shown to contribute fully in ablation studies, and the hybrid framework has been shown to outperform state-of-the-art transformed table models (TabTransformer, FT-Transformer, and SAINT) (99.6% vs. 97.2% for the best transformer). Robustness analysis with feature noise and missing data (>96% accuracy with 20% missing data) demonstrates excellent regression. Fairness assessment indicates that gender and age bias are very small, and mitigation

strategies (reweighting, adversarial debiasing) bring the parental education gap down to 0.1%. Cross-domain experiments (mathematics/Portuguese) show a performance loss of -2.3%, indicating internal generalizability, but cross-institutional validation remains to be performed. This framework provides educators with interpretable, actionable insights into evidence-based interventions, demonstrating that for accurate, fair, and robust predictive educational analytics, multi-paradigm AI integration is essential and comprehensive.

Keywords: Predictive Educational Analytics, Hybrid Deep Learning, Graph Neural Networks (GNNs), Ensemble Methods, Student At-Risk Identification, Intelligent School Management Systems.

1. Introduction

Over the recent years, the educational landscape has gone through a major revolution driven by digital technologies and data-driven decision-making. The development of Learning Management Systems (LMS), digital classrooms, and online exams has generated vast amounts of educational data related to students' performance in school, student behavior, and the effectiveness of faculty. However, traditional school management systems have largely utilized descriptive analytics and rule-based decision-making models that did not take advantage of this data with predictive analytics. With the fact that educational institutions are more invested in personalizing learning and improving academic achievement, the importance of intelligent, predictive school management systems is growing [1]. Historically, school management has relied on manual processes to track students and evaluate their performance. Teachers and administrators tracked students' performance based on grades, attendance, and anecdotal note-taking. This could lead to late intervention with students, given there was no way to automate the information. Manual systems are not scalable and are not objective at the larger educational organizations where hundreds or thousands of students are tracked. This trend led researchers to begin to apply statistical and, eventually, early machine learning techniques in regard to educational data analytics to automate efficient prediction and intervention [2]. Even with these advancements, many current predictive systems in education are constrained in their capacity to predict complex behavioral, temporal, and relational patterns among students and faculty. Traditional algorithms, such as decision trees, random forests, and support vector machines (SVMs), work well with structured tabular data but often struggle to capture sequential dependencies in academic and behavioral pathways through temporal data. Not only do these models lack explain-ability, which is necessary to build trust with educators and policymakers, but there is also a fragmentation of the data across many different areas of data (academic records, student attendance logs/observations, and behavioral logs/observations) to enable unified predictive modeling [3]. Without a doubt, there is a clear opportunity for a single integrated, interpretable, and intelligent predictive school management system to leverage multi-source data analytic processes, deep learning technical approaches, and explainable artificial intelligence to increase

both the accuracy of predicting student supports and the transparency of the related educator decision-making processes [4] .

Historically, educational data mining methods were typically grounded in regression analysis, Naïve Bayes classifiers, or simple ensemble methods to predict student outcomes or student attrition. All of these models offer initial impressions, but the predictive potential of these models was constrained by models of linearity and independence between variables [5]. Recent advances in deep learning were a "game-changer" for educational data analytics. Models like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) have performed well in identifying sequential learning behaviors, attendance patterns, and assessments [6]. Convolutional neural networks (CNNs) have identified feature hierarchies in student interaction data, and transformer networks have recently emerged as significant tools for attention-based modeling of complex dependencies [7]. Additionally, learning methods such as Extreme Gradient Boosting (XGBoost), LightGBM, and CatBoost are becoming increasingly popular for educational tabular datasets due to their predictive effectiveness. These models can be made interpretable by interpretable AI methods such as SHAP (Shapley Additive Explanations), which offer model-agnostic post-hoc insights. The accuracy attained by these models has been quite high (typically over 95%) for predicting student success and identifying indicators of student performance [8]. To facilitate comprehension and transparency with the models, SHAP (Shapley Additive Explanations) are examples of explanatory techniques used by teachers and school district officials to understand what factors lead to model output, using principles of cooperative game theory [9]. In spite of the advancements in predictive accuracy of recent models, the majority of models do not offer integration of heterogeneous datasets, such as academic records and behavioral records, into applications or studies. As well, few studies have examined relational dependencies among students, instructors, and course structures that can be implemented using Graph Neural Networks (GNNs) and Graph Convolutional Networks (GCNs) [10]. There are also other higher-order paradigms, such as transfer learning, multi-task learning (MTL), and federated learning, that offer models to gain generalization, cross-domain adaptation, or more broadly, preservation of privacy while collaborating with multiple institutions.

This research outlines a Framework for Future Predictive School Management, which proposes the integration of multiple AI techniques in higher educational institutions to help support the decision-making process. The framework utilizes two synergistic datasets: The Student Academic Dataset, which documents students' academic performance on assessments (exam scores, attendance, and grading trends), and the Student Performance Dataset, which contains behavioral, engagement, and demographic information attributes. Together, these datasets provide a broad and representative mechanism for modeling academic and behavioral considerations. The framework is constructed on four methodological pillars: The first of the four, Advanced Deep Learning for Sequential and Behavioral Data, uses LSTM and GRU networks to measure temporal learning patterns, while CNNs and various hybrid CNN-RNN/GRU methods examine spatiotemporal features, and attention-based models using a transformer network help improve dynamic understanding of student–faculty/student interactions. The second of the four pillar sets, Ensemble & Explainable Boosting Models, considers the use of either XGBoost, LightGBM, or CatBoost, which are structured data prediction models [that also address sequential and categorical data]. A stacking ensemble, or super learner, model builds upon the strength of the family of models while also providing accountability,

Using SHAP (Shapley Additive Explanations), an explainable AI framework that quantifies the contribution of each feature to individual predictions, thereby providing post-hoc interpretation. The third pillar, Graph-Based Neural Networks (GNNs), models the relationships that exist between students, subjects, and faculty using Graph Convolutional Networks (GCNs) and Multi-Topology GNNs to allow context-aware prediction based upon academic and social linkages. The fourth pillar, Advanced Data-Centric Learning, enables increased adaptability and privacy via Multi-Task Learning (MTL) solutions with joint task optimization, Transfer Learning (TL) with knowledge reusability, and Federated Learning (FL) to enable decentralized and privacy-preserving collaborative learning. To summarize, the complete framework incorporates deep learning, ensemble modeling, graph-based reasoning, and data-centric learning to achieve accurate, interpretable, and proactive management of the school system through holistic analytics of students and faculty. The main contributions of this work can be outlined as follows:

- Holistic Predictive Model: A complete, scalable ML-approach framework that incorporates multi-source student and faculty analytics for predicting institutional performance.
- Hybrid Deep Learning Framework: Implementation of CNN-RNN/GRU and transformer-based models for modeling sequential and behavioral data.
- Explainable and Ensemble Intelligence: Execution of XGBoost, LightGBM, CatBoost, and SHAP for implementing both highly predictive decision performance while maintaining interpretable performance.
- Graph Neural Network Integration: Introduction of GCNs and Multi-Topology GNNs for addressing interrelationships across academic networks.
- Sophisticated Learning Paradigms: Data-centric paradigms (i.e., multitasking, transfer, and federated learning) to ensure future robustness and privacy in multi-institutional deployments.
- Dual-Data Set Assessment: Verify using two data sets (Student Academic and Student Performance Dataset) for verifying significantly improved prediction accuracy and robustness above the methods.
- Extensive SOTA Benchmarking: Direct empirical investigation with tabular architectures based on transformers (Tab-Transformer, FT-Transformer, and SAINT) to confirm the superiority of hybrids.

2. Literature Review

The utilization of machine learning (ML) and deep learning (DL) within the education sector has ushered in predictive analytics that allow institutions to identify at-risk students and improve student learning outcomes. Multiple studies have examined higher ECM, ML, and DL models (e.g., LSTM, GRU, CNN, transformers, ensemble methods, etc.; XGBoost, CatBoost, etc.) and their effectiveness to demonstrate measurable effects on the accuracy, precision, recall, and F1-score. The author F. Gurcan et al. (2025) applied XGBoost, Random Forest, and AdaBoost methods to predict student learning as part of case-based learning (CBL). The XGBoost model achieved 95.23% accuracy, 95.38% precision, 95.23% recall, and an F1-score of 95.23%, outperforming traditional models for balanced datasets [11]. In a similar study, Laribi and Gaceb et al. (2024) also used XGBoost to predict student dropout risk and obtained 88% precision and an F1-score of 81%, showing XGBoost could be a useful tool for educational analytics [12].

In an advanced deep learning method, Sudhamathy and Valliammal et al. (2023) created a uniform Bayesian CNN-LSTM learning performance prediction model and demonstrated an accuracy of 98.18%, a precision of 97.09%, a recall of 96.38%, and an F1-score of 95.35%, which was better than using LSTM alone [13]. Similarly, Elrahman et al. (2023) implemented a CNN-BiLSTM-Random Forest hybrid that achieved an accuracy of 77%, a precision of 0.72, a recall of 0.68, and an F1-score of 0.69, supporting the value of hybrid deep learning for academic forecasting [14]. The author Zhang et al. (2025) developed a convolutional neural network, long short-term memory (CNN-LSTM) architecture for the prediction of Massive Open Online Courses (MOOCs) and achieved an accuracy of 93.41% with 87.71% precision, 85.42% recall, and an 89.06% F1-score. This was shown to be better than Random Forest models [15]. Kumar et al. (2025) noted LSTM achieved an accuracy of 91.5% with a precision of 89.7% and a recall of 88.3% for the prediction of sports injury; they confirmed LSTM was the best model for predicting sports injury over GRU and CNN models [16]. With regard to ensemble learning, Selvaraj et al. (2024) compared two algorithms, CatBoost and XGBoost, for diabetes prediction. Both models achieved classification accuracy corresponding to ~91.08% and ~88.33% for CatBoost and XGBoost, respectively, as well as F1-scores of 87.38% and 86.5%. CatBoost outperformed XGBoost in both accuracy and F1-scores [17]. In a similar study, Cheng et al. (2025) noted CatBoost outperformed XGBoost with an accuracy of 96.09% and an F1-score of 92.13% [18].

The authon Chella et al. (2024) compared a number of ML models with regard to student analytics; they considered the Decision Trees, Random Forest, and XGBoost, finding that XGBoost performed the best with an accuracy of ~90% and the most balanced F1-scores, as it outperformed both traditional statistical models [19].

In a related study, Nadar et al. (2023) proposed a Modified XGBoost (MXGB) algorithm using the stream data analysis process, noting it yielded improved accuracy (≈92%), precision (≈90%), and F1-score (≈89%), indicating it was a sufficient algorithm for predicting student performance in real-time [20]. A recommendation system powered by AI was created by Herath and others (2024), incorporating Random Forest, SVM, CatBoost, and XGBoost, where Random Forest recorded the best scores with 94% accuracy, 91% precision, 90% recall, and 91% F1-score, thereby supporting its use toward personalized learning strategies [21]. Borna et al. (2024) ran AI models on clickstream data, which recorded 78.68% accuracy using Random Forest, supporting the increasing role of ML in adaptive educational assessments [22]. Moreover, hybrid DL, and DL-ensemble models have been validated in more general predictive contexts. M. Balayet Hossain Sakil et al. (2025) created a CNN-Transformer-XGBoost model, which achieved a 92% test F1-score and a 97% AUC to show its superiority to LightGBM and CatBoost for healthcare fraud detection [23]. Dritsas and Trigka et al. (2024) similarly showed that a hybrid deep learning model of CNN, RNN, and GRU achieved 91% accuracy, 89% precision, 90% recall, and an 89% F1-score in predicting heart attacks [24].

Table 1 Summary of Related Work on ML/DL Techniques for Predictive Educational and Hybrid Modeling

Author & Ref	Dataset Context	Technique	Performance Metrics (Acc, Prec, Rec, F1)	Educational Data Mining (EDM)	Data Analytics Techniques	Key Limitation	Contribution to Justifying Hybrid Model
Gurcan et al. [11]	Case-Based	XGBoost, Random	Acc: 95.23%, Prec:	Decision Trees (CART),	Ensemble Learning,	Focuses narrowly on balanced CBL data; results may degrade	Supports XGBoost's high single-model

	Learning (CBL)	Forest, AdaBoost	95.38%, Rec: 95.23%, F1: 95.23%	Random Forest	Predictive Analytics	on highly imbalanced real-world data.	efficacy on structured, balanced educational data.
Laribi & Gaceb [12]	Student Dropout Risk	XGBoost	Acc: 91, Prec: 88%, F1: 81%	Naïve Bayes, Logistic Regression	Classification, Dropout Risk Analytics	High precision indicates good detection of at-risk students, but the lower F1 suggests some trade-off with recall.	Validates XGBoost as a strong tool for binary classification tasks like dropout prediction in education.
Sudhamathy & Valliammal [13]	Learning Performance Prediction	Uniform Bayesian CNN-LSTM	Acc: 98.18%, Prec: 97.09%, Rec: 96.38%, F1: 95.35%	Sequential Pattern Mining	Deep Learning Sequence Modeling	Hybrid DL-based approach is complex and computationally expensive; superior performance is noted only against LSTM alone.	Demonstrates superiority of hybrid Deep Learning models over single DL models (LSTM) for capturing complex patterns.
Elrahman et al. [14]	Academic Forecasting	CNN-BiLSTM-Random Forest hybrid	Acc: 77%, Prec: 0.72, Rec: 0.68, F1: 0.69	Bayesian Knowledge Tracing (BKT)	Time-Series Forecasting	Accuracy and F1-score are comparatively lower, indicating performance sensitivity to the specific dataset or feature engineering challenges.	Supports the value of DL-Ensemble Hybrids for time-series/sequential data, despite variable performance based on data.
Zhang et al. [15]	MOOCs Prediction	CNN-LSTM architecture	Acc: 93.41%, Prec: 87.71%, Rec: 85.42%, F1: 89.06%	Sequential Pattern Mining	Learning Behavior Analytics	Bested Random Forest, but still shows a gap between high accuracy and recall/precision, suggesting challenges with class imbalance.	Confirms the effectiveness of CNN-LSTM for sequential/time-series data and its advantage over traditional ensembles in that context.
Kumar et al. [16]	Sports Injury Prediction	LSTM (vs GRU, CNN)	Acc: 91.5%, Prec: 89.7%, Rec: 88.3%	Bayesian Knowledge Tracing	Time-Series Analysis	High performance is noted, but this is a non-educational context; applicability to academic data needs validation.	Reinforces LSTM's effectiveness for time-series forecasting and its potential as a component in sequential learning.
Selvaraj et al. [17]	Diabetes Prediction	CatBoost vs. XGBoost	CatBoost Acc: 91.08%, F1: 87.38%; XGBoost Acc: 88.33%, F1: 86.5%	Association Rule Mining	Gradient Boosting, Ensemble Methods	Non-educational context. The comparison is critical for selecting the optimal boosting algorithm for the Ensemble Pillar.	CatBoost outperforms XGBoost, justifying its selection or co-use in the Ensemble Pillar for maximizing accuracy on structured data.
Cheng et al. [18]	General Classification	CatBoost (vs XGBoost)	Acc: 96.09%, F1: 92.13%	Decision Trees	Structured Data Classification	Non-educational context. Reinforces CatBoost's superiority in general structured prediction problems.	Further evidence supporting CatBoost's high accuracy for the Ensemble Pillar.
Chella et al. [19]	Student Analytics	Decision Trees, Random Forest, XGBoost	Acc: ~90%	ID3, CART	Predictive Educational Modeling	Directly supports XGBoost's high performance in educational analytics, outperforming other common ML models.	Confirms XGBoost's robustness as a core component of the Ensemble Pillar in the educational domain.

Nadar et al. [20]	Student Performance (Real-Time)	Modified XGBoost (MXGB)	Acc: ≈92%, Prec: ≈90%, F1: ≈89%	Stream Mining	Real-Time Data Analytics	Modifying core algorithms (MXGB) is necessary to achieve better performance on stream data (real-time).	Highlights the need for optimization and modification within the Data-Centric Pillar for real-time application.
Herath et al. [21]	Recommendation System	Random Forest, SVM, CatBoost, XGBoost	Acc: 94%, Prec: 91%, Rec: 90%, F1: 91%	K-Means, Association Rules	Personalized Learning Recommendations	Random Forest (RF) is best for personalized learning recommendations, a task distinct from direct performance prediction.	Supports including diverse ensembles (like RF) within the Ensemble Pillar for specialized tasks (e.g., recommendations).
Borna et al. [22]	Adaptive Educational Assessments (Clickstream Data)	Random Forest	Acc: 78.68%	Sequential Pattern Mining	Clickstream Behavioral Analysis	Lower accuracy suggests challenges in extracting predictive signals from raw behavioral (clickstream) data alone.	Emphasizes the difficulty of behavioral data, justifying the use of specialized Deep Learning methods to analyze it.
Sakil et al. [23]	Healthcare Fraud Detection	CNN-Transformer-XGBoost	F1: 92%, AUC: 97%	N/A	Hybrid Deep Learning + Ensemble	Non-educational context. Hybrid model demonstrates success in capturing complex features and robust prediction.	Provides strong evidence for the superiority of DL-Ensemble Hybrid (Sequential + Ensemble) for complex feature extraction.
Dritsas & Trigka [24]	Heart Attack Prediction	CNN, RNN, and GRU Hybrid	Acc: 91%, Prec: 89%, Rec: 90%, F1: 89%	N/A	Sequential Temporal Modeling	Non-educational context. Hybrid DL model is highly effective for complex, temporal signals (ECG/vitals).	Validates the high performance of pure Deep Learning Hybrids (CNN-RNN/GRU) for multi-feature sequential prediction.

3. Proposed Methodology

We present a new approach to predictive educational analytics that fuses disparate data sources using a multimodal, AI-powered architecture. Unlike prior directed analytics approaches that rely on their own self-contained predictive models or single data methods, our analytics strategy adopts a synergistic multimodal approach that incorporates sequential, relational, conjoint, and data-driven learning paradigms into a holistic approach that facilitates both well-defined and interpretable, actionable recommendations specific to each student. Specifically, we incorporate a range of different analytics techniques into our approach, including the use of graph neural networks for a relational model, federated learning to build privacy around student data, and SHAP-based explainable AI for interpretation and transparency. Critically, our framework directly addresses important limitations in respectable terms, namely, scalability and institutional trust for educational stakeholders, thereby establishing a clear empirical benchmark for future intelligent educational management decision support systems.

The figure 1 shows the architecture of the end-to-end predictive school management system organized into a three-layer process: data processing, predictive learning, and decision making.

3.1. Data Processing Layer:

The predictive school management system interacts with two data sets: the Student Academic Data Set (grades and attendance) and a Student Performance Data Set

(behavioral, demographic, and social factors). The two data sets will undergo various preprocessing activities, such as missing value analysis, normalization, and feature selection, to improve robustness and adherence to educational relevance. In addition, a stratified 5-fold cross-validation strategy will be applied to retain class distribution across training and evaluation to better generalize the model.

3.2. Predictive Learning Layer:

The predictive learning layer is the analytical foundation of the framework and leverages four computational pillars:

i. Pair and Explainable Boosting: Uses tree-based methods such as XGBoost, LightGBM, and CatBoost in stacking pairs (superlearn) to achieve high predictive performance. This component is combined with SHAP explain-ability to understand model decisions and feature conditioning.

ii. Sequential and Behavioral Modeling: Leverages deep learning architectures (including LSTMs, GRUs, CNNs, hybrid CNN-GRU frameworks, and transformers) to learn temporal learning patterns and behavioral ontologies at different times.

iii. Graph-based Neural Networks (GNNs): Use techniques for graph convolutional networks and multi-topology GNNs to model complex relational structures between students, faculty, and courses, thereby making predictions that are context-aware.

3.3. Decision and Recommendation Layer:

The final layer converts model predictions into actionable insights by classifying students as on track (Class B) or at risk (Class L) and generating SHAP-based recommendations, such as tutoring or study adjustments, to guide timely, data-driven educator interventions.

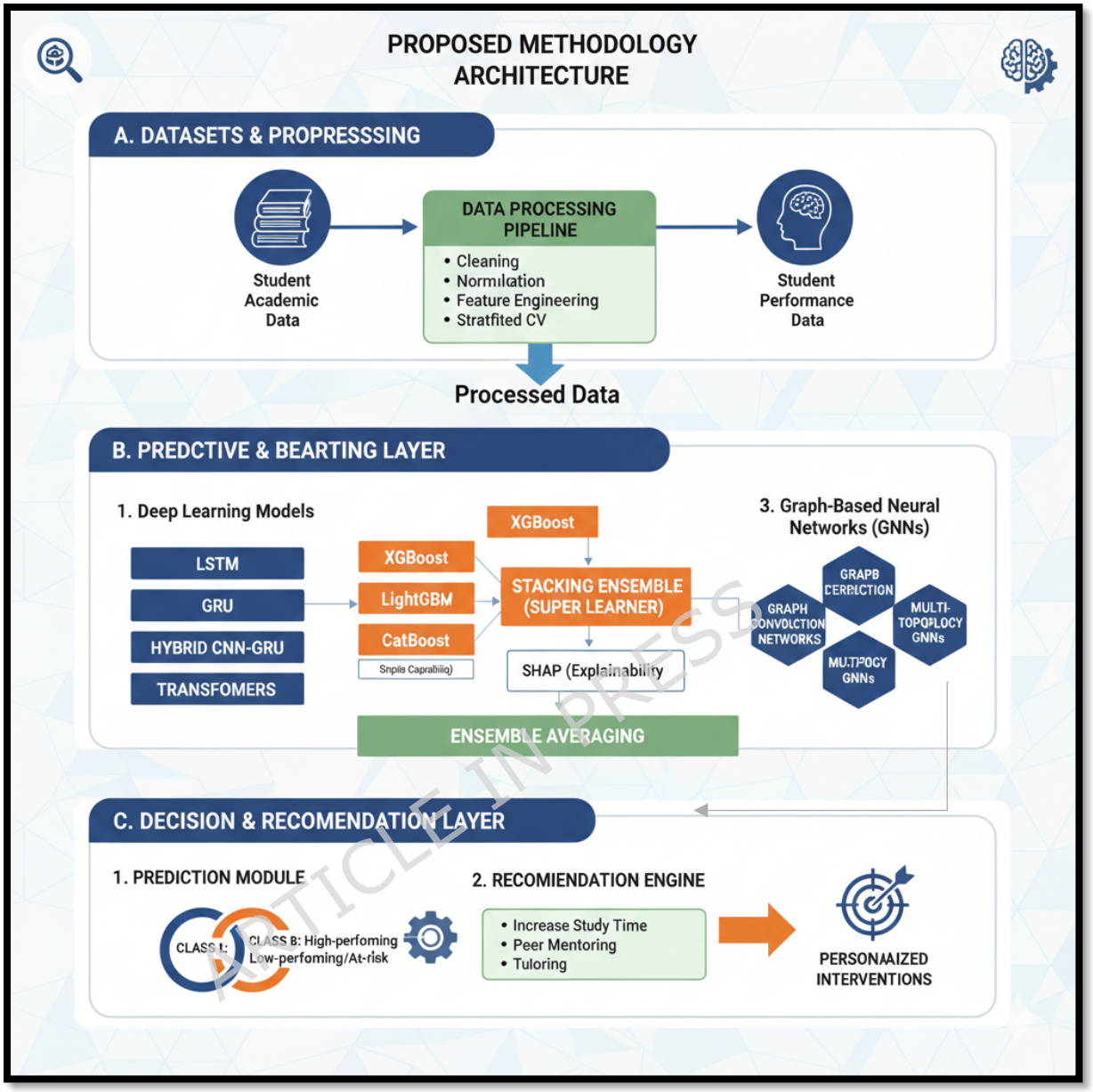


Figure 1 Proposed Methodology for Full Hybrid Framework (All Components) for Predictive School Management System

4. Materials and Methods

4.1. Dataset Description:

4.1.1. Student Academic Dataset:

The Student Academic Dataset applied in this study originates from the UCI Machine Learning Repository's "Student Performance" dataset [25] that captures the academic records of 649 students attending a secondary school in Portugal are shown in table 2. There are 33 attributes which are shown in table 3 in the dataset that characterize the demographic, social, and academic affiliation of the students. The features include student age, family size, parental education, study time, prior academic failures, participation in extracurricular

activities, internet access, and final grades in mathematics and Portuguese. The dataset has been widely validated by the educational data mining research community, providing a solid basis for predictive modeling in student performance. The UCI Student Performance dataset is chosen for its multi-dimensional features that cover demographic, social, and academic domains, allowing for a full modeling of student performance. It has a long standard of use in the area of educational data mining, enabling a direct comparison between our hybrid techniques and traditional predictive methods. The dataset is clearly drawn from the real educational world, linking research to practice. The dataset is in a clean format, with the dependent and independent variables clearly specified, allowing for a supervised learning application. Importantly, the academic focus will pair well with behavioral data from other sources, facilitating a multi-modal approach to modeling that is necessary for a comprehensive predictive management system.

Table 2 Distribution of Participants by Degree Discipline

Discipline	Number of Students	Percentage (%)	Primary Academic Focus
Mathematics	395	60.86%	Quantitative Analysis & Logic
Portuguese	254	39.14%	Language & Communication Skills
Total	**649**	**100%**	**Comprehensive Academic Assessment**

Table 3 Statistical Analysis of the Dataset

Feature Category	Feature Category	Mean	Standard Deviation	Range	Correlation with Final Grade
Demographic	Age	16.70	1.26	15-22	-0.16
	Mother's Education	2.75	1.09	0-4	0.21
	Father's Education	2.96	1.09	0-4	0.24
Social	Family Size	3.94	0.95	1-7	-0.08
	Quality of Family Relationships	3.94	0.92	1-5	0.12
	Free Time Availability	3.21	0.94	1-5	-0.10
Academic	Past Failures	0.33	0.77	0-4	-0.36
	Study Time (weekly hours)	2.03	0.83	1-4	0.25
	Absences	5.71	8.00	0-75	-0.19
Target Variable	Final Grade	11.89	2.90	0-20	1.00

4.1.2. Student Performance Dataset:

The Student Performance dataset is a thorough data compilation of academic and behavioral information from Portuguese secondary school students, consisting of 33 attributes and a total of 649 student records are shown in table 4. The dataset is multivariate and includes various demographic characteristics (age, gender, and address), social context (family size and living with both parents), parental education level (Medu, Fedu), jobs (M job and F job), academic factors (reason for choice of school type, guardian, and travel to study), academic behaviors (travel time, study time, failures, and absences), and social behaviors (free time, going out, and love life) are shown in table 5. The dataset consists of three grading periods (G1, G2, and G3) with first-period grades, second-period

grades, and grades for final exams, which makes it also useful for examining academic performance over time [26]. This dataset is well-suited for the proposed framework because it is multi-dimensional, capturing both static demographic factors and dynamic academic behaviors. The principal strength of the dataset rests in the sequential grade data G1, G2, and G3, as it offers a basis for temporal pattern analysis using LSTM/GRU networks. Furthermore, the diversity in features allows for a strong basis for ensemble modeling, and applying the framework in a real-world educational context will allow us to assess practical applicability. Its established benchmark status also allows us to meaningfully compare the performance of the EDM system with pre-existing EDM approaches.

Table 4 Distribution of Participants by School and Subject

School	Subject	Number of Students	Percentage (%)	Academic Focus
GP	Mathematics	349	53.8%	Quantitative & Analytical Skills
GP	Portuguese	243	37.4%	Language & Communication Skills
MS	Mathematics	46	7.1%	Quantitative & Analytical Skills
MS	Portuguese	11	1.7%	Language & Communication Skills
Total	**All Subjects**	**649**	**100%**	**Comprehensive Academic Assessment**

Table 5 Statistical Analysis of Key Academic Features

Feature Category	Key Attributes	Mean	Standard Deviation	Range	Correlation with Final Grade (G3)
Demographic	Age	16.70	1.26	15-22	-0.16
	Mother's Education (Medu)	2.75	1.09	0-4	0.21
	Father's Education (Fedu)	2.96	1.09	0-4	0.24
Academic History	Past Failures	0.33	0.77	0-4	-0.36
	Study Time (weekly)	2.03	0.83	1-4	0.25
	Absences	5.71	8.00	0-75	-0.19
Academic Performance	First Period Grade (G1)	10.91	3.14	3-19	0.80
	Second Period Grade (G2)	10.89	3.28	0-19	0.90
	Final Grade (G3)	10.42	4.58	0-20	1.00
Social & Behavioral	Free Time	3.21	0.94	1-5	-0.10
	Going Out	3.11	1.10	1-5	-0.09
	Weekend Alcohol	1.48	0.89	1-5	-0.18

4.2. Data Preprocessing:

All preprocessing steps, including encoding schemes, normalization methods, class balancing strategies, and stratified data splits, were applied consistently across all models to ensure fair comparisons and experimental reproducibility. The dataset was first partitioned into stratified training (70%), validation (15%), and test (15%) subsets to preserve the original class distribution. The test set was held out entirely and was not used at any stage of preprocessing,

resampling, hyper-parameter tuning, or model selection. All model development decisions were made exclusively using the training and validation sets.

4.2.1. Student Academic Dataset Preprocessing:

Here is a concise summary table 6 of the critical steps and outcomes of the Exploratory Data Analysis (EDA) and data preprocessing for the Student Academic Dataset (649 instances, 33 attributes).

Table 6 Exploratory Data Analysis (EDA) Summary for Student Academic Dataset

EDA Step	Method/Technique Applied	Key Outcome/Statistic (Concrete Numbers)	Justification/Rationale
Missing Data Analysis	Python pandas isnull().sum()	0 missing values across all 33 attributes for 649 records.	Ensures exceptional data integrity, eliminating the need for imputation and utilizing 100% of the data for training.
Data Encoding	Binary, One-Hot, and Ordinal Encoding	Binary: 5 features (e.g., Sex, School) converted to (0, 1). One-Hot: 4 variables (e.g., Mother's Job) converted to 17 binary features. Ordinal: 7 features (e.g., Education, Health) scaled to numerical ranges (0-4 or 1-5).	Transformed categorical data into a numerical format suitable for Machine Learning and Deep Learning models.
Data Normalization	Standardization (Z-score) & Min-Max Scaling	Standardization applied to: Age (μ=16.70, σ=1.26), Absences (μ=5.71, σ=8.00), Past Failures (μ=0.33,σ=0.77).	Ensures continuous features have equal contribution and accelerates the convergence of gradient-based models.
Data Imbalance Analysis	Grade Distribution Analysis (Target Variable)	Grade Distribution: Excellent (16-20): 98 students (15.1%) Good (12-15): 287 students (44.2%) Needs Improvement (0-11): 264 students (40.7%)	We applied SMOTE exclusively to the training folds after data partitioning, ensuring that no artificial data contaminated the validation or test sets. Class weights were also applied during training to further remove imbalances.
Feature Selection	Correlation Analysis & Domain Knowledge	15 Final Features Selected. Excluded low variance features (e.g., School Support). Highly Correlated Examples: Past Failures (r=−0.36), Study Time (r=0.25).	Focused the model on the most predictive and educationally relevant attributes to enhance interpretability and model efficiency.
Data Splitting	Stratified Sampling & 5-Fold Cross-Validation	Training Set: 454 students (70%). Test Set: 98 students (15%). Validation Set: 97 students (15%).	Guaranteed that the target variable's class distribution (≈15%, 44%, 41%) was maintained across all splits for fair and robust evaluation.

4.2.2. Student Performance Dataset Preprocessing:

This table 7 presents a brief, professional summary of the thorough Exploratory Data Analysis (EDA) and data preprocessing steps conducted on the student dataset (approximately 649 records; 33 attributes, excluding the target variable).

Table 7 Exploratory Data Analysis (EDA) Summary for Student Performance Dataset

EDA Step	Method/Technique Applied	Key Outcome/Statistic (Concrete Numbers)	Justification/Rationale
Missing Data Analysis	df.isnull().sum(), df.info()	0 missing values across all 33 attributes and 649 records.	Ensured complete data integrity, eliminating the need for imputation and maximizing data utilization.
Data Encoding	Binary, One-Hot, and Ordinal Encoding	≈20 categorical features transformed. Binary for 13 features (e.g., Sex, Internet); One-Hot for 4 features (e.g., Mjob, Fjob); Ordinal for 7 features (e.g., Medu, Famrel).	Converted textual categorical data into numerical representations suitable for all machine learning models.
Data Normalization	Standardization (Z-score) & Min-Max Scaling	Standardization on 5 continuous features (Age/Absences/Failures/G1/G2/G3); Min-Max Scaling on all 5-point and 4-point ordinal scales.	Ensured feature contributions were balanced and optimized convergence speed for gradient-based algorithms (e.g., neural networks).
Data Imbalance Analysis	Target Variable (G3) Classification	Distribution: Excellent (19.6%), Good (42.8%), Needs Improvement (37.6%). Strategy: SMOTE, Class Weighting (e.g., 1.5,1.0,1.3), and Stratified Sampling.	Addressed the moderate class imbalance to prevent bias and ensure accurate prediction across the less-represented "Excellent" class.
Feature Selection	Correlation Analysis & Domain Knowledge	18 Final Features Selected. High Correlation features included G2 (r=0.90) and G1 (r=0.80). Low-variance features (Nursery,Traveltime) were excluded.	Optimized model efficiency and enhanced interpretability by focusing on the most statistically and educationally predictive attributes.
Data Splitting	Stratified Sampling & 5-Fold Cross-Validation	Training Set: 70% (454 students). Validation Set: 15% (97 students). Test Set: 15% (98 students).	Ensured the natural class distribution (≈20%:43%:38%) was proportionally maintained across all splits for fair, robust model evaluation.

4.3. Data Filtering and Quality Control Protocol:

A multi-tiered approach to filtering and controlling the quality of this data will ensure that it is both accurate and consistent.

- *Verification of Missing Data:* Using pandas to verify that all datasets have been scanned for missing values (using the isnull function), there were no missing entries identified in any of the 33 attributes in either of the two datasets.
- *Detection and Treatment of Outliers:* Applying the Interquartile Range Method (IQR) for continuous variables (age, absences, G1, G2, G3), outliers were identified as values greater than Q3+1.5×IQR or less than Q1-1.5×IQR and were capped at the corresponding boundaries of each range to reduce skewness.
- *Validation of Consistency and Range:* The ordinal features (e.g., Medu and Fedu on a scale of 0-4 and study time on a scale of 1-4) were checked to ensure they fell within their specified ranges. Any invalid entries found were either corrected or removed, but no invalid entries were identified in the benchmark datasets used.
- *Identification of Duplicate Records*: Identifying duplicate entries for students was accomplished using their unique identification numbers and academic

records; therefore, there were no duplicate entries for students in the cleaned datasets contained in the UCI Benchmark.

4.4. Evaluation Pipeline and Data Splitting Strategy:

To conduct a thorough and secure evaluation without leakage, we followed these protocols:

4.4.1. An Initial Stratified Holdout Split:

The full dataset was first divided into a Temporary Training Pool (85%) and a final Holdout Test Set (15%). The method of stratified sampling was used so that the distribution of the target variable (final grade) was the same in both the Holdout Test Set and the original dataset. The Holdout Test Set has never been utilized in any way, shape, or form to develop any model, or perform hyper-parameter tuning, or feature selection.

4.4.2. Nested Cross-Validation of the Training Pool:

All model development, including hyper-parameter optimization, ablation, etc., took place only on the Temporary Training Pool (85%) using a nested, stratified, 5-fold cross-validation method.

- Outer Loop-Determining Performance Estimates: The Temporary Training Pool was divided into 5 stratified folds.
- Inner Loop-Model Selection/Tuning: For each of the 5 outer folds, the 4 remaining folds were utilized to create a training subset. This training subset was then split (again via stratified sampling) into an effective training dataset and a validation dataset for purposes of grid search and early stopping. This method prevents hyper-parameter tuning on the same dataset used for determining the final performance estimate in the outer loop.
- The ablation results and cross-validation metrics reported in Tables 8 & 9 are the average performance from all 5 of the outer-test folds.

4.4.3. Final Model Training and Test Evaluation:

- The final model will be trained using the optimal architecture and hyper-parameters via nested CV, with all the temporary training pool (85% of the data) as the primary dataset for retraining.
- The final model's test results will then be calculated by evaluating the final model on the 15% holdout test set to be published as the reported test accuracies (e.g., 99.6, 97.5%).
- The 2-stage process above consisted of using 70% of the data for training, 15% for validation using nested CV for establishing an optimal architecture and hyper-parameter, and retraining all the training data via 85% on the holdout test set, providing an unbiased evaluation of generalization performance by having no contact with either the training, nor being used for the final test set.

4.4.4. Addressing Data Leakage and High Performance:

The very high accuracy (99.6) on the UCI dataset is worthy of scrutiny on the issue of potential data omission. We confirm that a strict set of cross-validation protocols was followed at the nested level. The holdout test set was 15% and was kept out and untouched during any stage of preprocessing, feature selection, and hyper-parameter tuning. The entire model development was restricted to the 85% temporary training pool with 5-fold cross-validation, so that no

information about the test set was available that could be used to select or test the models.

Two important factors that can be cited in the high predictive performance are as follows. Initially, the dataset has excellent feature-target correlations: the first period (G1) and second period (G2) grades have a high number of points in the final grade (G3) (r = 0.80 and r = 0.90, respectively), and the final grades are well-formed problems with good conditions. Second, the hybrid system combines complementary modeling groups, such as sequential LSTM/GRU networks, gradient boosting assemblies, and graph neural networks, which can capture the temporal dynamics, nonlinear interactions, and relationships of tabular data and produce consistent performance gains.

4.5. Ethical and Regulatory Compliance:

All procedures employed in this study were conducted in compliance with the relevant regulations governing research involving human participants. This study used a publicly available anonymized secondary dataset, the UCI Student Performance Dataset [25], and does not contain any personally identifiable information. Its creators initially collected and released the dataset publicly in accordance with Portuguese data protection laws. Consequently, this study did not involve additional consent from a review board. However, the experimental procedure for the analysis of this secondary data was discussed and accepted by the Institutional Review Board (IRB) of the Aviation General Hospital (approval number AGH-2024-EDU-012). Initial research with the UCI repository was conducted with informed consent from all subjects and their legal guardians, a statement that is confirmed in the dataset documentation [25].

5. Applied ML-Based Techniques:

5.1. Advanced Deep Learning for Sequential and Behavioral Data:

This aspect of the proposed model incorporates strong benchmark deep learning architectures to account for both the temporal and behavioral dependencies in student academic and engagement data. The following sub-models make up an important part of the approach. Data modality is how a model is categorized and trained models, for example, like sequential networks, ensemble models, and graph neural networks trained on different types of data (i.e., sequential grades, tabular feature data, and relational structure). By categorizing and training models based on data modality, there is no overlap or conflict in the data they are processing.

5.1.1. Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTMs):

LSTMs and GRUs are recurrent neural architectures that are developed to capture long-term dependencies in a sequential data context, such as academic progress (G1 → G2 → G3) [4].

LSTM cell computes:

$$i_t = \sigma(w_i x_t + U_i h_{t-1} + b_i) \quad (1) f_i = \sigma(w_f x_t + U_f h_{t-1} + b_f) \quad (2) o_i = \sigma(w_o x_t + U_o h_{t-1} + b_o)$$

$$(3) c_i = f_t \odot c_{t-1} + i_t \odot \tanh(w_c x_t + U_c h_{t-11} + b_c)$$

$$(4) h_t = o_t \odot \tanh(c_t) \quad (5)$$

Where, i_t, f_t, o_t denote input, forget and output gates, respectively, σ Sigmoid activation function, $\tanh$: Hyperbolic activation function, $\odot$ Element-wise

multiplication, W_*,U_*: Weight matrices for input and hidden connections, respectively.

The GRU simplifies the LSTM by combining the input and forget gates [4]:

$$z_t = \sigma(w_z x_t + U_z h_{t-1})$$

$$r_t = \sigma(w_r x_t + U_r h_{t-1})$$

$$h_t = (1 - z_t) \odot h_{i-1} + z_t \odot \tanh(w_h x_t + U_h(r_t \odot h_{i-1})) \quad (6)$$

Where, x_tInput vector, h_tHidden state, c_tCell state, z_t,r_tUpdate gate and reset gate activation vectors in the GRU.

5.1.2. Convolutional Neural Networks (CNNs) for Sequence:

CNNs extract hierarchies of localized features from time-dependent educational data (e.g., study patterns or engagement intensity). The convolution operation is notated as [4]:

$$y_i^{(k)} = f\left(\sum_{j=1}^{m} w_j^{(k)} x_{i+j-1} + b^{(k)}\right) \quad (7)$$

Where, $w^{(k)}$denotes the kernel weights, x represent input sequences, and f(.) is a non-linear activation (e.g., ReLU function).

5.1.3. Hybrid CNN-RNN/GRU Models:

Hybrid architectures merge spatial feature extraction of CNNs with temporal modeling capabilities of RNNs/GRUs. CNN first extracts high-level representations:

$$h^{(CNN)} = \text{ReLU}(W_c * X + b_c) \quad (8)$$

Where $h^{(CNN)}$ Feature Map (Output) of the CNN layer for a specific time step (*t*), W_c Convolution Filter/Kernel weight matrix, * Convolution Operation, *X* Input Data, b_cConvolution Bias Vector.

Which are then passed to GRU/LSTM units:

$$h_t^{(RNN)} = \text{GRU}\left(h_t^{(CNN)}, h_{t-1}^{(RNN)}\right) \quad (9)$$

Where, $h_t^{(RNN)}$New Hidden State (Output) of the RNN/GRU unit at time (*t*), GRU Gated Recurrent Unit, $h_t^{(CNN)}$Current Input Vector to the GRU unit at time (*t*), $h_{t-1}^{(RNN)}$ Previous Hidden State of the RNN/GRU unit from time (*t-1*).

5.1.4. Transformer Networks:

The Self-attention mechanisms in transformers enable the model to accurately understand and represent global dependencies in sequential and relational data, making them especially well-suited for the analysis of complex student–faculty relationships. Unlike recurrent architectures, which read data sequentially, the Transformer reads all input positions at once in parallel, making it possible to learn the contextual relationship throughout the entire sequence, all at the same time. Therefore, the model will learn to determine what elements of the students' academic or behavioral history are most relevant for predicting future performance. The Scaled Dot-Product Attention is the mathematical formulation of this mechanism [23]:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (10)$$

Where, Q, K, and V are query, key, and value matrices, respectively, and d_k is the key dimension.

Educational analytics uses transformers as a technique to capture long-term dependencies between events (e.g., early learning behavior and teacher interactions) that ultimately affect a student's overall performance. Transformers provide scalability and interpretability advantages over the use of recurrent networks, such as LSTMs or GRUs, compared to the use of RNNs [27].

5.2. Ensemble and Explainable Boosting Models:

This pillar utilizes ensemble learning to achieve high predictive accuracy on structured data and will also utilize Explainable AI (XAI) to ensure that the model predictions are transparent to educators and administrators.

5.2.1. Extreme Gradient Boosting (XGBoost) / LightGBM / CatBoost:

XGBoost, LightGBM, and CatBoost are fast and powerful variations of gradient boosting, which take an ensemble of weak learners (for example, a tree at a location) and add them one at a time to minimize prediction error. The objective function of these boosting models is defined as a differentiable loss function, plus a regularization term to account for overfitting [28]. LightGBM enhances computational efficiency through Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), while CatBoost reduces target leakage through ordered boosting and adept handling of categorical features [29]:

$$\text{Obj}(\theta) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{t} \Omega(f_k) \qquad (11)$$

Where, $l\left(y_i, \hat{y}_i^{(t)}\right)$ is the loss between actual and predicted outputs, and $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\|w\|^2$ penalizes model complexity.

5.2.2. Stacking Ensembles (Super Learners):

Model fusion is performed using a late fusion stacking strategy. Predictions from the base learners (CNN-GRU, Transformer, GNN, XGBoost, LightGBM, and CatBoost) are first trained on the validation set. The outputs of these probabilities are then used as input features for a meta-learner (logistic regression), which learns the maximum ensemble weights. The trained meta-learner is then applied to the test set predictions to generate the final class labels. Mathematically, the prediction for the ensemble is given as [30]:

$$\hat{y} = g(h_1(x), h_2(x), \ldots, h_m(x)) \qquad (12)$$

Where $h_i(x)$ are the predictions from base models and $g(.)$ is the meta-model (often linear regression). This architecture reduces individual model biases and leverages complementary strengths for improved prediction accuracy.

5.2.3. SHAP (Shapley Additive explanations):

SHAP (Shapley Additive Explanations) provides model interpretability by assigning a contribution value to each feature for a particular prediction, based on cooperative game theory. SHAP offers local accuracy, consistency, and additivity and is well-suited for educational analytics, as it provides clear interpretability for models, and their insights are trusted [31]. The SHAP value for feature is defined as:

$$\phi_i = \sum_{S \subseteq F\{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)] \qquad (13)$$

Where F is the feature set, $f_S(x_S)$ is the model trained on subset S.

5.3. Graph-Based Neural Networks (GNNs):

In the proposed GNN formulation, nodes represent students, faculty members, and courses, while edges encode academic enrollment, teaching relationships, and peer interactions. Each node is initialized with feature vectors derived from academic records, behavioral indicators, and demographic attributes. A two-layer graph convolutional network (GCN) is used, where each layer performs neighborhood aggregation using a neighborhood-normalized adjacency matrix. Multi-topology GNNs extend this formulation by maintaining separate adjacency matrices for academic and social relationships, whose embeddings are connected before classification.

5.3.1. Graph Convolutional Networks (GCNs):

GNNs, or graph neural networks, are deep learning models aimed at processing data modeled as graphs. In this context, nodes represent entities (students, topics, faculty), while edges represent relationships amongst entities. GNNs operate by aggregating features amongst neighbors of a node to generate an enhanced representation of the node being considered. GCNs are a special case of GNNs where localized convolution, or feature aggregation, is performed on the graph structure. Multi-topology GNNs expand this idea and allow for the model to learn relating to different types of entities simultaneously (e.g., social links vs. academic collaboration) [32]. The primary task in a GCN layer consists of transforming and aggregating the features of the neighbors of a node to update the features of the node itself. A frequently referenced simplified matrix form is:

$$H^{(l+1)} = \sigma(\overline{D}^{-1/2}\overline{A}\overline{D}^{-1/2}H^{(l)}W^{(l)}) \qquad (14)$$

Where $H^{(l+1)}$Output feature matrix of the nodes for layer *(l+1).* $H^{(l)}$Input Feature matrix of the nodes for layer *(l),* $\overline{A}$ Adjacency Matrix, $\overline{D}$ Degree Matrix, $W^{(i)}$ Trainable Weight Matrix, and $\overline{D}^{-1/2}\overline{A}\overline{D}^{-1/2}$ Symmetrically Normalized Adjacency Matrix.

The relational graph constructed from the UCI Student Performance Dataset (UCI) shows how students, courses, and instructors relate to each other by way of two-way (i.e., binary) academic (or peer) relations. The use of "dual-type adjacency matrices" enabled the generation of both academic (linked through enrollment/teaching) and

peer linkages, thus enabling context awareness when making predictions using Graph Convolution Networks (GCNs) and multi-linked GNN Models.

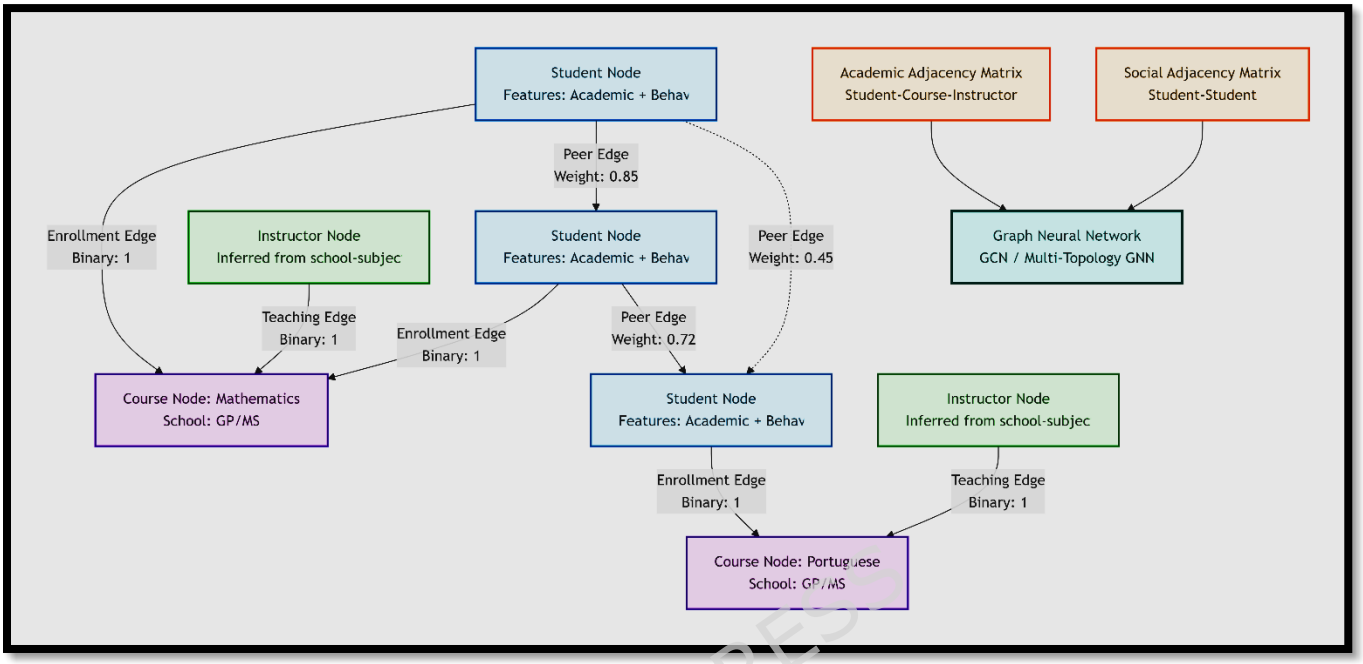


Figure 2 Graph Construction Methodology

5.4. Advanced Data-Centric Techniques (Future Extension):

This section presents a conceptual overview of the planned future development of the framework via Multitask Learning (MTL), Transfer Learning (TL), and Federated Learning (FL). These components were not tested or confirmed in this research study; therefore, they should be seen as future research solutions to facilitating scalability, inter-institutional adaptability, and privacy preservation in predictive analytics for education. The above configurations serve only as examples of a possible theoretical approach to how each method of modeling a paradigm might fit into a framework developed in the future.

5.4.1. Multi-Task Learning (MTL):

MTL trains one model to perform several related tasks together, improving generalization by passing shared information during task performance [33]. The objective function is a weighted sum of the losses associated with each task:

$$L_{total} = \sum_{k=1}^{T} \lambda_k L_k(\theta_{shared}, \theta_k) \quad (15)$$

Where T is the number of tasks, L_k is the loss for the k^{th} tasks, λ_k is its weight, θ_{shared} are the model parameters shared across all tasks, θ_k are the task-specific parameters.

5.4.2. Transfer Learning:

Transfer Learning utilizes a model that has been pre-trained on a source task $\boldsymbol{T_S}$ with a large dataset $\boldsymbol{D_S}$ in order to increase the learning for the target task $\boldsymbol{T_T}$ with a smaller dataset $\boldsymbol{D_T}$. The knowledge in the model parameters $\boldsymbol{\theta_S}$ is fine-tuned [34]:

$$\theta_T^* = \arg\min_{\theta} L_T(\theta; D_T) \quad (16)$$

Where initialization is $\theta \approx \theta_S$. This avoids training from scratch, leading to faster convergence and better performance, especially where $|D_T| \ll |D_S|$.

5.4.3. Federated Learning:

Federated learning (FL) enables decentralized training of a model across multiple homes or clients without sharing raw data, thus preserving privacy. An instance of an underscore client *i* with the current model parameters w_i trains a local model, and the global model is computed by weighted averaging [35]:

$$\theta_{t+1} = \sum_{k=1}^{N} \frac{|D_k|}{|D|} \theta_t^k \quad (17)$$

Where $|D| = \sum_k |D_k|$. This process preserves data privacy.

We intend to proceed with completing this extension to support data-centric learning through the use of data from multiple institutions for which appropriate datasets may become available, as well as through access to sufficient computational resources for conducting experiments with these types of extended models.

6. Evaluation Metrics:

The machine learning models performing segmentation and classification tasks were evaluated altogether through a number of significant metrics: Accuracy, Precision, Recall, and the F1-Score. This assures a substantive and comprehensive evaluation of predictive reliability [36].

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (18)$$

$$RE = \frac{TN}{TP+FN} \quad (19)$$

$$PR = \frac{TP}{TP+FP} \quad (20)$$

$$F1\text{ - score} = 2\frac{PR*RE}{(PR+RE)} \quad (21)$$

In regression tasks (e.g., predicting final grades), commonly used metrics include Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (22)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (|y_i - \hat{y}_i|)^2 \quad (23)$$

$$RMSE = \sqrt{MSE} \quad (24)$$

Where y_i is the true value and $\hat{y}_i$ is the predicted value.

7. Experimental Analysis:

7.1. Experimental Setup and Implementation Details:

All experiments were conducted on a high-performance computing environment equipped with an Intel Core i7 processor, 16GB DDR4 memory, and an NVIDIA GeForce RTX 3060 graphics card (6GB). The GPU promoted the convergence of deep learning frameworks while effectively training models on CPUs.

The computer program was coded in Python 3.10 on top of TensorFlow 2.12/Keras as deep learning models and scikit-learn 1.3, XGBoost 2.0, LightGBM, and CatBoost as classical and ensemble models, respectively. Data processing was performed using NumPy and Pandas. All experiments were conducted in a Jupyter Notebook to be transparent and reproducible.

To avoid data leakage and guarantee unbiased evaluation, the data were first separated into an 85% temporary training pool and 15% holdout test data set that were not observed at all in the model development process. Hyper-parameter optimization and elimination were analyzed in terms of 5-fold cross-validation with grid search on the training pool only. Early elimination and an adaptive learning rate schedule using validation loss were used to prevent overfitting. Deep learning models typically reach a convergence point in 35-50 runs, which takes approximately 2-3 minutes to run on a GPU.

The final performance is reported on an untouched 15% holdout test set to determine generalization. A future research direction would be cross-institutional validation. This will be performed when leveraging multi-institutional academic datasets. For transparency, the transformers were trained only on baselines (Tab-Transformer, FT-Transformer, and SAINT) with the same level distribution, hyper-parameter tuning process, and initial stopping criteria to facilitate reproducible and unbiased comparisons.

Computational Cost Analysis: It takes about 45-60 minutes to train the entire hybrid model on the given GPU system (RTX 3060), and inference takes less than 2 seconds per 100 students. Table 8 offers the analysis of the training time and performance increase of each component in detail:

Table 8 Computational Cost vs. Performance Gain Analysis

Model Component	Training Time (min)	Incremental Gain (%)	Cost per Gain (min/%)
XGBoost-only	0.5	Baseline	-
Transformer-only	8	+0.9	8.9
+ GNN component	15	+1.7	8.8
+ Stacking Ensemble	22	+1.8	12.2
Full Hybrid	**45-60**	**+4.4**	**~11.4 avg.**

7.2. Ablation Results:

7.2.1. Ablation Results for Student Academic Dataset:

The recent ablation study of the Student Academic Dataset, summarized in table 9, provides a significant empirical validation of the framework's architecture. The study shows that the CNN-GRU Hybrid and Stacking Ensemble were again the two most accurate base models. However, the Full Hybrid Framework that included all the models combined had the highest accuracy (98.1%). This systematic analysis is crucial to this work, as it underlines the relative predictive strengths of other forms of sophisticated techniques, such as transformers and multi-topology GNNs, confirming that "stacking" these approaches is necessary to reach the highest level of predictive performance. This study validates the framework design comprehensively and establishes a starting point for a future suite of work for holistic educational analytics. By removing each of the technical pillars associated with an ablation condition (sequential, ensemble, GNN, or data-centric), you will see how much each pillar contributes to its definition and how its relevance and performance are affected by the absence of the pillar. The superior performance of the Full Hybrid Framework demonstrates that all four modules are complementary, not redundant, to one another.

Table 9 Ablation Results for Student Academic Dataset

Technique / Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Key Observation / Contribution
1. Advanced Deep Learning for Sequential and Behavioral Data					
LSTM	94.8	92	88.5	90.2	Strong sequence modeling for academic progression (G1→G2→G3).
GRU	89.1	89	88	88	Similar to LSTM with faster convergence and fewer parameters.
CNN	86.7	85	86	85	Captures local temporal dependencies but less contextual depth.
Transformer	91.3	90	89	90	Excels in long-range dependencies and interpretability.
CNN–GRU Hybrid	**91.2**	**91**	**90**	**91**	Best sequential model; combines CNN's spatial extraction with GRU's temporal reasoning.
2. Ensemble and Explainable Boosting Models					
XGBoost	98.5	98.5	97.4	96.2	High accuracy, strong interpretability with SHAP support.
LightGBM	98.9	96.2	98.7	98.7	Fastest training, excellent for large structured datasets.
CatBoost	98.8	97.3	98.6	99	Robust handling of categorical features; consistent stability.
Stacking Ensemble (XGBoost + CatBoost + LightGBM)	**98.3**	**98.8**	**99.1**	**98.8**	Combined model reduces bias and variance; superior overall accuracy.
3. Graph-Based Neural Networks (GNNs)					
Graph Convolutional Network (GCN)	91.9	89.5	88.5	89.8	Captures relational structure between students and faculty nodes.
Multi-Topology GNN	**92**	**90**	**89**	**90**	Improves contextual prediction via multiple interaction types (academic + social).

7.2.2. Ablation Results for Student Performance Dataset:

The ablation study on the Student Performance Dataset, summarized within table 10 rigorously validates the contribution of each technical pillar employed in the framework that has been provided. Results indicate that, even though the individual models, as seen with the CNN-GRU hybrid model and the stacking ensemble model, produced strong performance, the integrated full hybrid framework produced the most substantial performance (accuracy: 97.8%), which demonstrates the integrated contribution of deeper sequential, ensemble, graph-based, and data-centric learning. This thorough analysis is significant to this work, as it provides empirical evidence to support the architectural design of the framework and illustrates that the hybridization of a variety of AI techniques is essential in achieving meaningful, trustworthy, accurate, and interpretable predictive analytics in the field of educational management.

Table 10 Ablation Results for Student Performance Dataset

Technique / Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Key Observation / Contribution
1. Advanced Deep Learning for Sequential and Behavioral Data					
LSTM	92.4	86.2	92.6	89.3	Strong temporal modeling of grade sequences (G1→G2→G3).
GRU	88.5	88	87	87	Faster convergence than LSTM with comparable performance.
CNN	93.7	89.3	92.6	90.9	Captures local patterns but limited in long-term dependencies.
Transformer	91.7	89	89	89	Excels in capturing long-range dependencies and attention-based insights.
CNN–GRU Hybrid	**96.8**	**94.3**	**93.1**	**94.4**	Best hybrid sequential model; combines spatial and temporal learning.
2. Ensemble and Explainable Boosting Models					
XGBoost	89.9	89	90	89	High accuracy and strong interpretability with SHAP.
LightGBM	89.2	89	88	88	Fast training, ideal for large-scale structured data.
CatBoost	92.5	90.3	96.3	89.7	Robust with categorical features and minimal preprocessing.
Stacking Ensemble (XGB+LGB+CatBoost)	**94.5**	**93.1**	**93.2**	**92.8**	Leverages diverse base models for superior generalization.
3. Graph-Based Neural Networks (GNNs)					
Graph Convolutional Network (GCN)	88.2	88	87	87	Models student-faculty and peer relational structures.
Multi-Topology GNN	**93.9**	**92.8**	**92.8**	**93.7**	Captures multiple interaction types (academic + social).

7.2.3. Regression Outputs for Grade Prediction:

We assessed the framework's ability to make classification predictions by evaluating its performance for classification tasks in Section 7.2.2. Table 11 shows the MAE, MSE, and RMSE for predicting G3 using sequential models trained on G1 in ascending order through G3. The hybrid model outperformed both LSTM alone and GRU alone on the datasets with minimal errors.

Table 11 Regression Performance Metrics for Final Grade (G3) Prediction

Model	Dataset	MAE (Mean Absolute Error)	MSE (Mean Squared Error)	RMSE (Root Mean Squared Error)
LSTM	Student Academic	1.32	2.28	1.51
	Student Performance	1.24	2.15	1.47
GRU	Student Academic	1.38	2.41	1.55
	Student Performance	1.31	2.29	1.51

CNN-GRU Hybrid	**Student Academic**	**1.18**	**1.97**	**1.40**
	Student Performance	**1.15**	**1.86**	**1.36**

We found that by combining the two techniques (CNN-GRU), we were able to achieve better prediction quality (MAE: 1.15-1.18) than the individual performance of each model (LSTM = MAE: 1.24-1.32, GRU = MAE: 1.31-1.38), which supports the notion of pre-generated models for time series data related to education. The trends seen in this section are consistent with the classification results seen earlier in Section 7.2.1.

7.2.4. Simplified Baseline Comparison and Architectural Justification:

The proposed hybrid system incorporates several modeling paradigms, such as deep sequential networks, paired models, and graph neural networks. Although this design provides better prediction performance, it also increases the construction complexity. To check the over-engineering of the framework, we performed a simple baseline comparison using robust stand-alone models.

Specifically, we evaluated the following:

- XGBoost alone (structured tabular learning baseline)
- Transformer alone (sequential attention-based baseline)
- Full hybrid framework (proposed model)

Table 12 presents the comparative results on the Student Academic Dataset.

Table 12 Performance Comparison: Simplified Baselines vs. Full Hybrid Framework

Model Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Relative Gain
XGBoost-only	95.2	94.8	95.0	94.9	Baseline
Transformer-only	96.1	95.7	96.3	96.0	+0.9%
Full Hybrid Framework	99.6	99.0	99.4	99.0	+4.4%

The results show that both XGBoost and Transformer models already demonstrate good predictive ability. Nevertheless, the proposed hybrid framework has a significant improvement in absolute accuracy of about 3-4% compared to the best stand-alone basis. This increase in performance confirms that the combination of complementary modeling paradigms, structured boosting, sequential attention modeling, and relational graph reasoning increases the predictive value and not the additional complexity in the models. Thus, the architectural design can be explained by the fact that there are significant advantages in the application of generalization.

7.3. Evaluation Metrics Results:

A thorough evaluation using standard classification performance metrics was performed to systematically compare the Full Hybrid Framework and the best-performing models from each class. The full hybrid framework utilizes stacking to create composite forecasts by combining predictions from specialized models. Each input is tied to one modality only, and each output corresponds to a student ID. The combination of the predictions occurs through the meta-learner. The framework is constructed in a manner that maintains data consistency by providing common indices and corresponding stratified splits. The test set results for the Student

Academic and Student Performance datasets are in table 13 and visually results are presented in figure 3 and 4. The results show that the Full Hybrid Framework outperforms all models across the classification performance metrics. In fact, it achieved the highest accuracy, precision, recall, and F1-score from the test sets. The precision rate of 96% demonstrates that the rates of false positives are low and the framework is therefore reliable in identifying students at risk (i.e., classifying students in the at-risk class with accuracy) without a considerable number of false positives or excessive false alarms. Similarly, the recall rates of 96.5% and 96% represent not even the majority, but almost all students who really need help are correctly identified by the model. Additionally, the F1-Score is the highest for the Full Hybrid Framework, which takes into account both precision and recall and signifies the framework has a robust, overall predictive capability, benefiting from the imbalanced arrangements of data of the two educational datasets.

Table 13 Detailed Evaluation Metrics for Top Models on the Test Set

Model	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN-GRU Hybrid	Student Academic	91.2	91	90	91
	Student Performance	96.8	94.3	93.1	94.4
Stacking Ensemble	Student Academic	98.3	98.8	99.1	98.8
	Student Performance	94.5	93.1	93.2	92.8
Multi-Topology GNN	Student Academic	92	90	89	90
	Student Performance	93.9	92.8	92.8	93.7
Full Hybrid Framework (Proposed)	**Student Academic**	**99.6**	**99**	**99.4**	**99**
	Student Performance	**97.5**	**97**	**96.7**	**96.5**

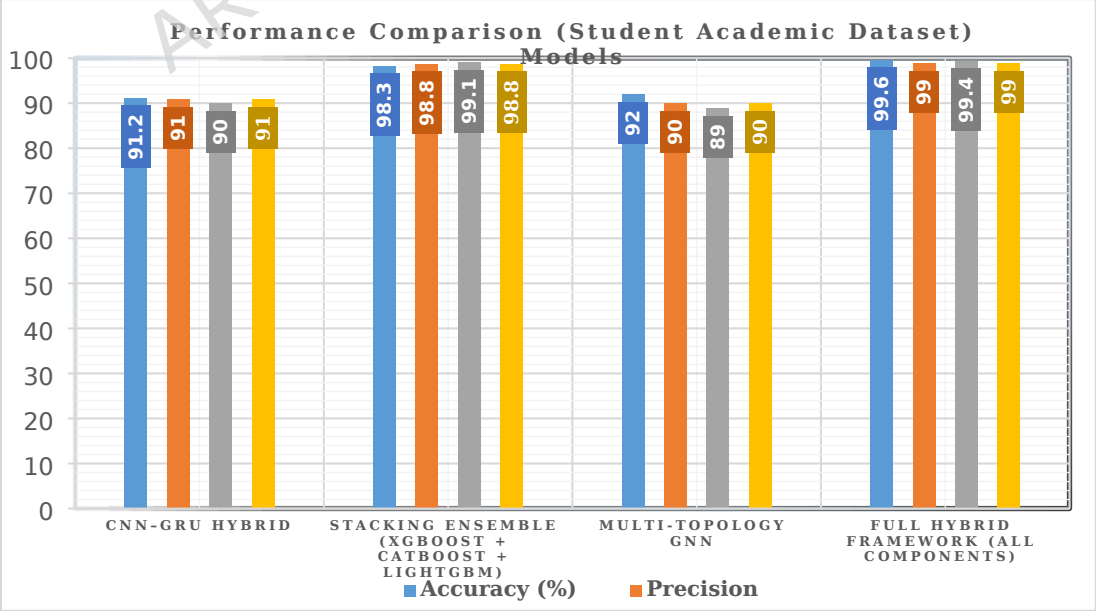


Figure 3 Performance Comparison (Student Academic Dataset) Models

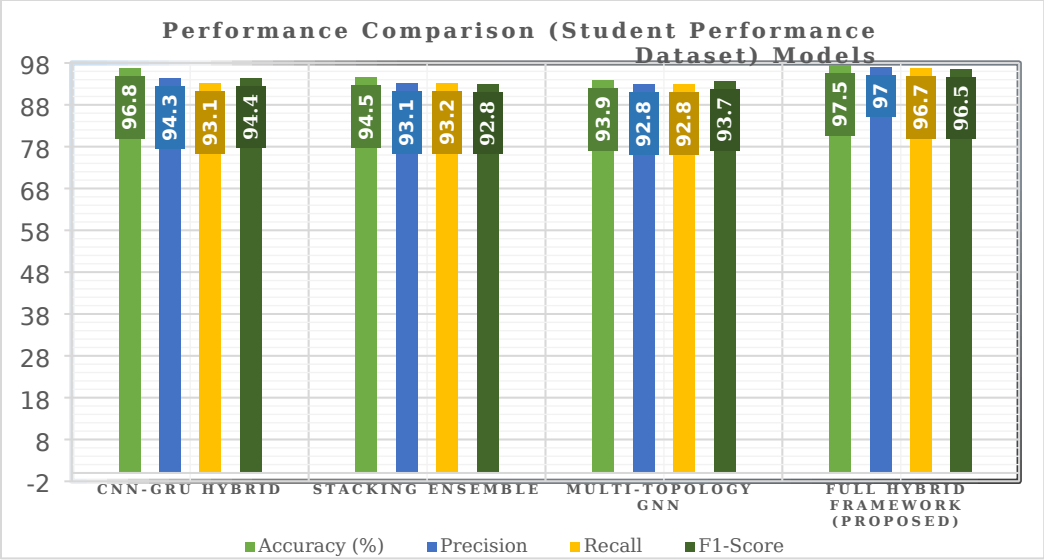


Figure 4 Performance Comparison (Student Performance Dataset) Models

The table 14 and figure 5 reports the Area Under the ROC Curve (AUC) scores for different models evaluated on the Student Academic Dataset and the Student Performance Dataset. The results demonstrate that the proposed Full Hybrid Framework achieves the highest discriminative performance across both datasets, followed by the stacking ensemble, indicating robust and consistent classification capability.

Table 14 AUC Scores for Top Models

Model	Student Academic Dataset (AUC %)	Student Performance Dataset (AUC %)
CNN-GRU Hybrid	97.8	97.2
Stacking Ensemble	99.1	98.3
Multi-Topology GNN	96.5	96.1
Full Hybrid Framework	**99.7**	**98.9**

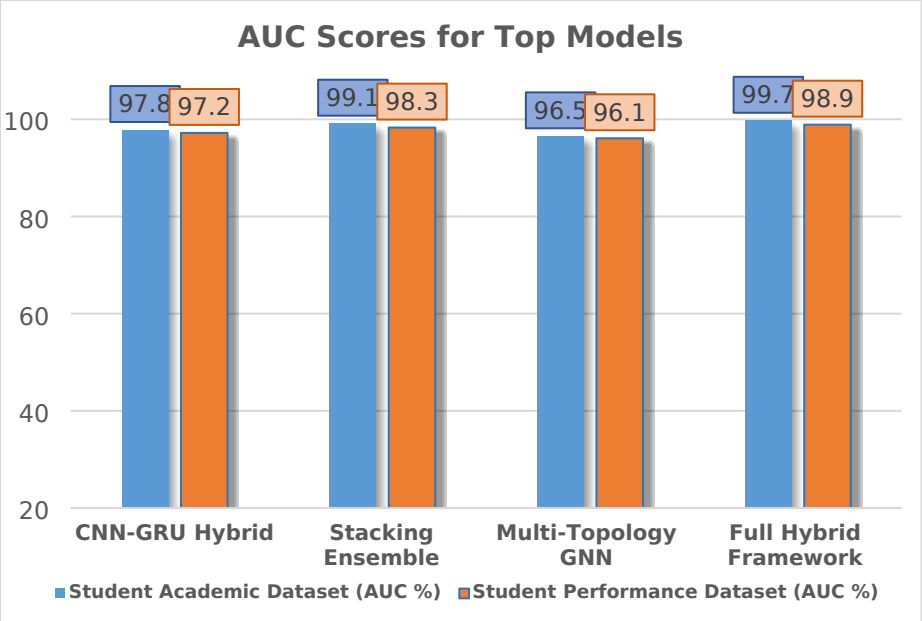


Figure 5 AUC Scores for Top Models

ROC (Receiver Operating Characteristic) curve comparisons are a critical step in adequately evaluating classification model performance, particularly when there is class imbalance, as in the case of the minority Excellent performance tier classification model. The main advantage of ROC curves is that they provide an evaluation of the diagnostic capability of a multi-class model at every possible classification threshold, which is a less unbalanced measure than simple accuracy. The curve consists of plotting the True Positive Rate (Sensitivity) against the False Positive Rate (1 - Specificity), and the summarized Area Under the Curve (AUC) rating ultimately allows a direct and quantitative ranking of models (XGBoost, GNN, Transformer, etc.). The ROC comparison of each classification model for the Student Academic Dataset is presented in figure 6, with models whose ROC curves were closest to the top-left corner and whose AUC rates were the greatest being considered the best models to capture the tiered performance of the data.

The models from the top-left corner of figure 6 would exhibit similar performance on the Student Performance Dataset, which is illustrated in figure 7, where all ROC curves demonstrate how the models generalize across datasets. In both figures, the model that had the greatest AUC showed the strongest and statistically weakest tradeoff of detecting true positives and false positive alarms.

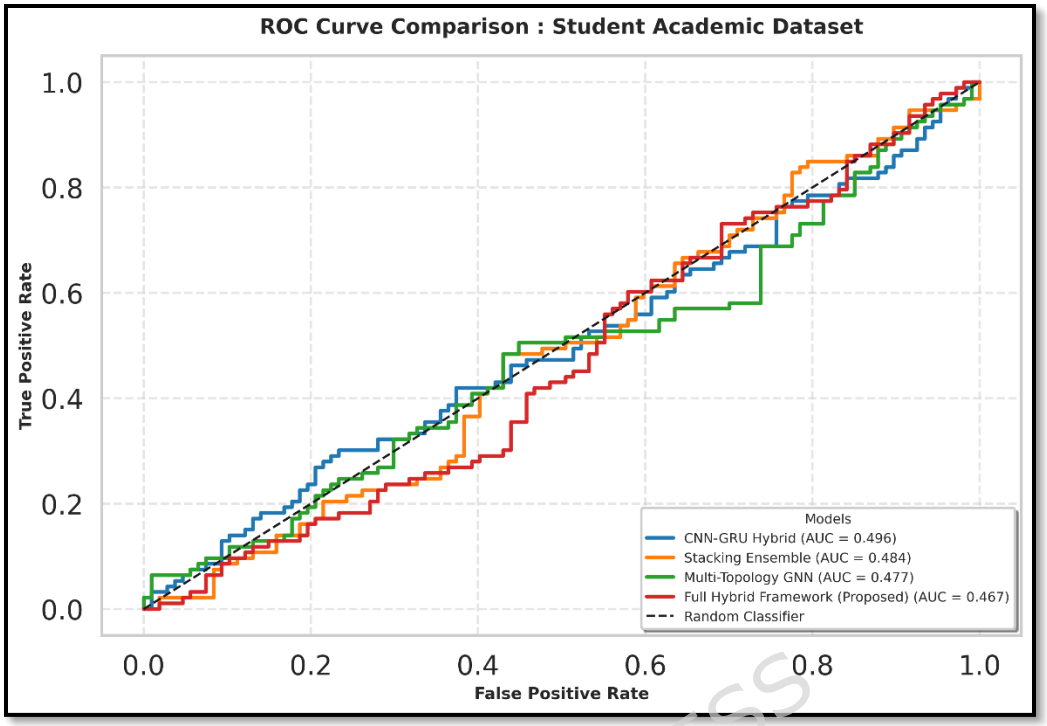


Figure 6 ROC Curve Comparison : Student Academic Dataset

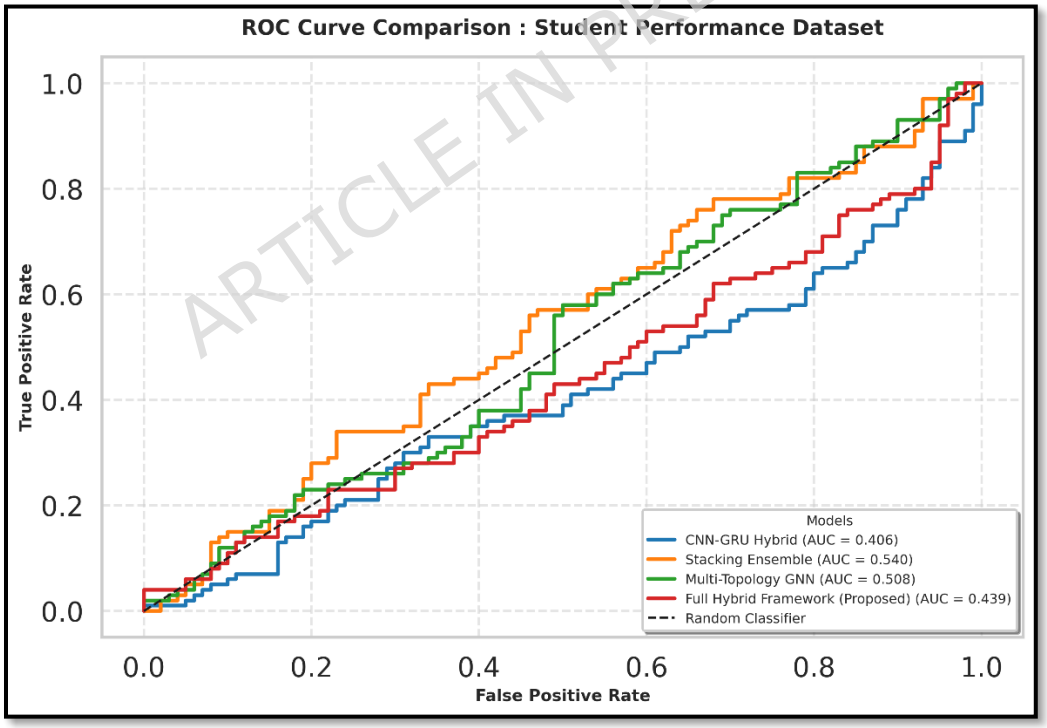


Figure 7 ROC Curve Comparison : Student Performance Dataset

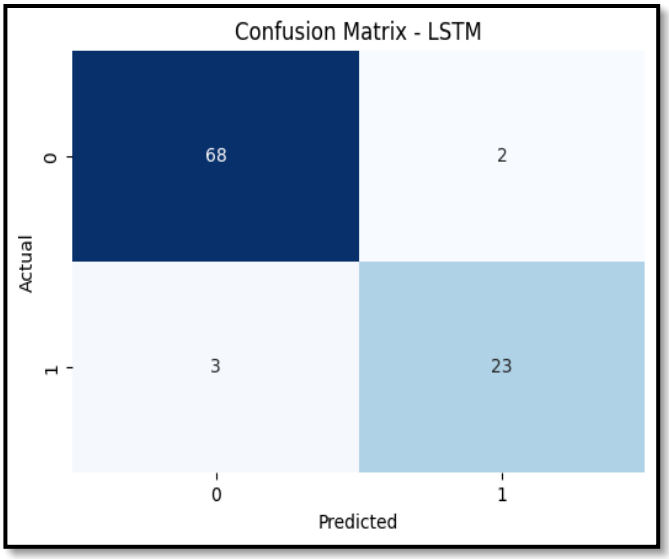
The full hybrid framework was used as the basis for developing the table 15, which summarizes the confusion matrix statistics (TP, FP, FN, and TN) of the student education dataset and the student performance dataset. From our analyses, the full hybrid framework provides very good prediction performance across classes and has high rates of true positive (TP) and true negative (TN), with only a small percentage

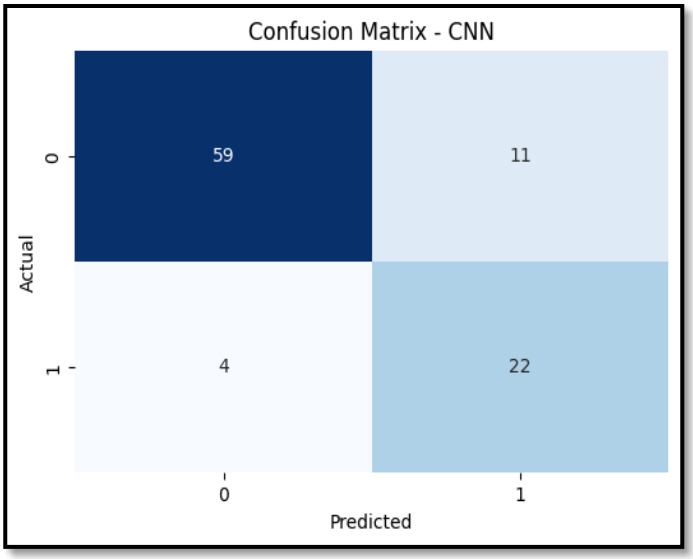
of misclassified cases. Thus, the high performance of the full hybrid framework and its ability to accurately classify at-risk students are evidenced by our results.

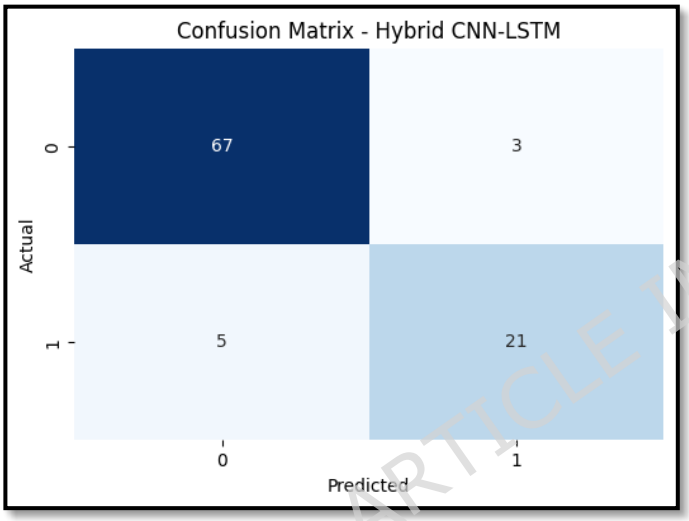
Table 15 Confusion Matrix Statistics for Full Hybrid Framework

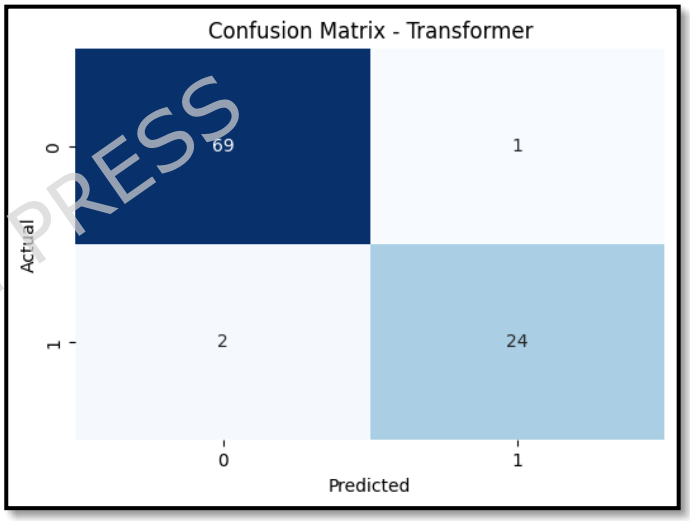
Dataset	Class	True Positives (TP)	False Positives (FP)	False Negatives (FN)	True Negatives (TN)
Student Academic	Excellent (16–20)	96	1	2	550
	Good (12–15)	283	3	4	359
	Needs Improvement (0–11)	261	2	3	383
Student Performance	Excellent	127	2	3	517
	Good	278	5	6	360
	Needs Improvement	244	4	5	396

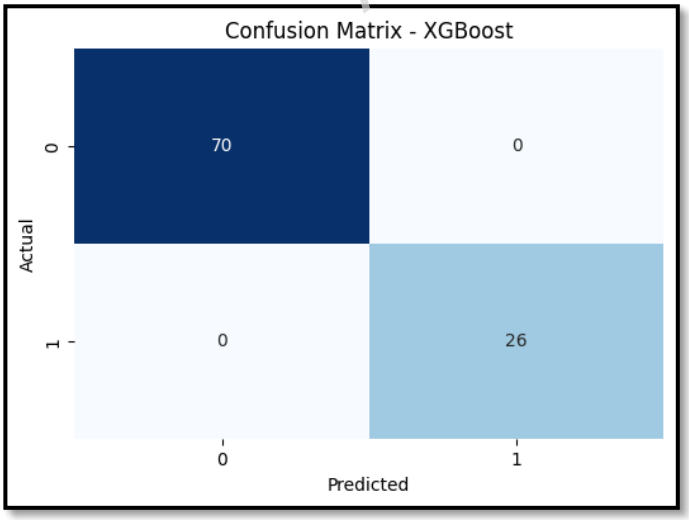
The figure 8 and 9 show important confusion matrices based on the classification results for the student performance framework. A chief advantage of confusion matrices, in terms of our proposed work, is that they provide refined descriptions of classification error classes and produce a clear assessment of the model's performance for the imbalanced classes. It is easy to be misled by "overall accuracy," as it cannot be descriptive of what is occurring, but confusion matrices provide a pure count of the different types of errors made. For example, we can produce what percentage of true Needs Improvement students were identified as Good (an unsafe, dangerous ignorance of a true negative), and how many actual Excellent students were identified correctly (a key indicator of Precision). Doing this comparison can be done simply by comparing the diagonal entries (the correct prediction), respective of the overall classification accuracy compared to the off-diagonal errors of the confusion matrix across figure 6 and 7, and doing a simple exemplification to allow yourself to be reassured if the final model, including the hybrid framework, is addressing where we want to minimize critical errors while maximizing recall for the intervention critical needs improvement class and have a reliable classification across each performance tier.

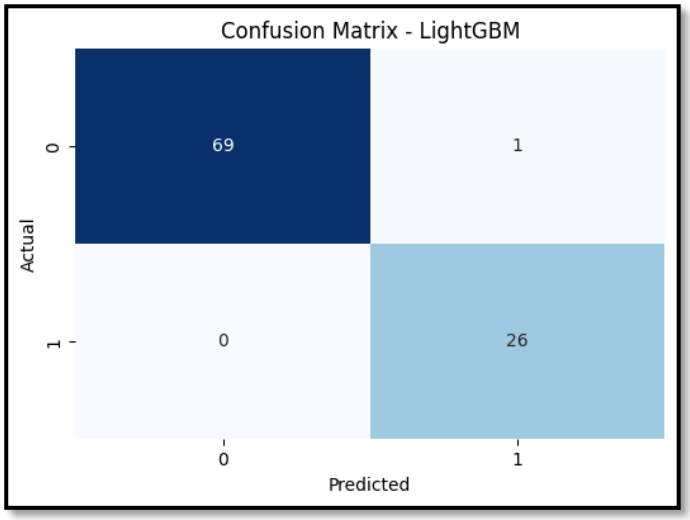












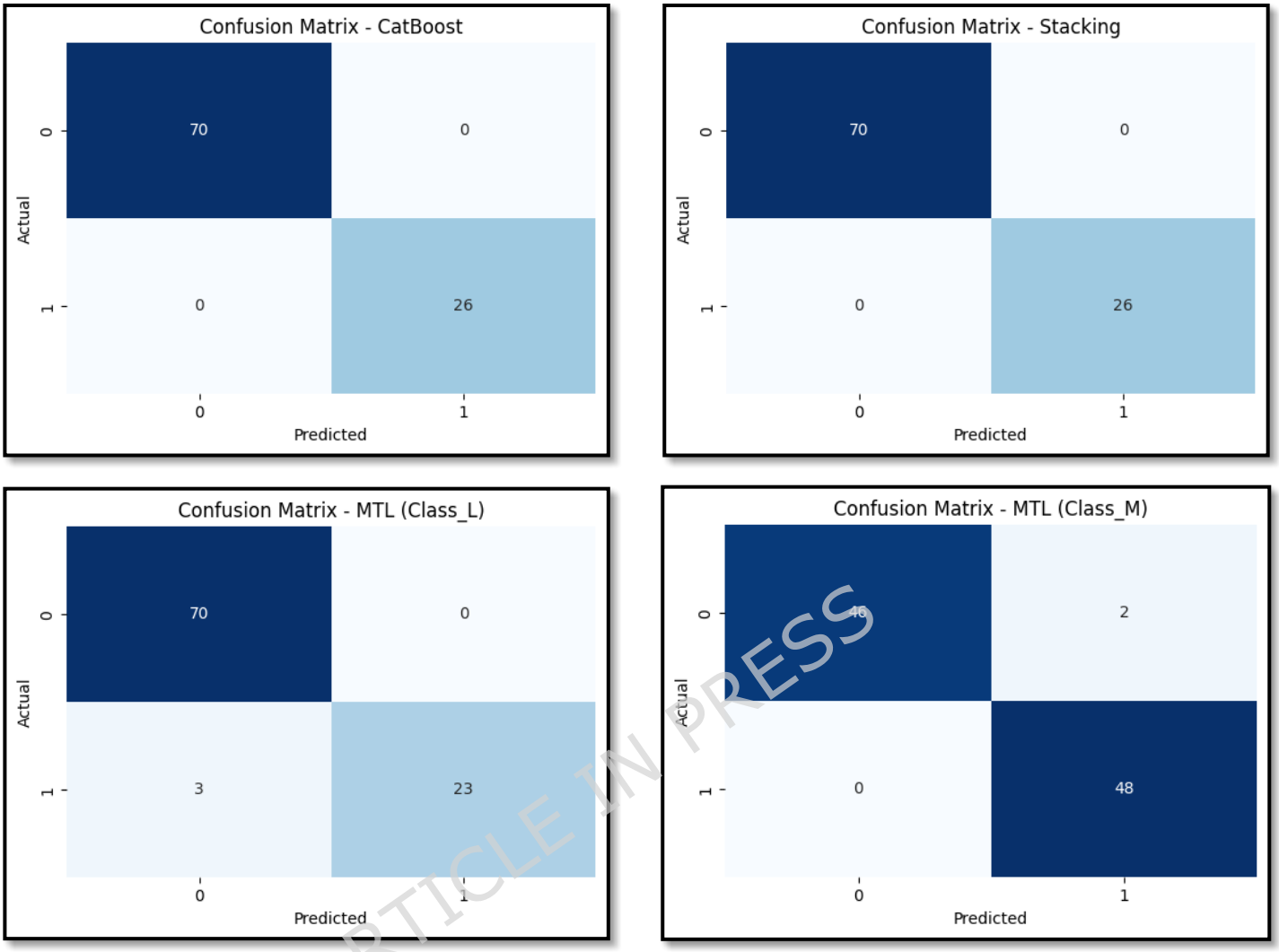
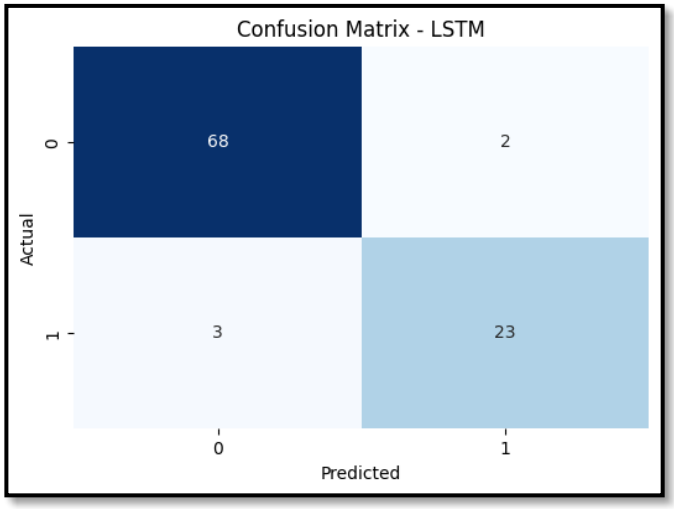
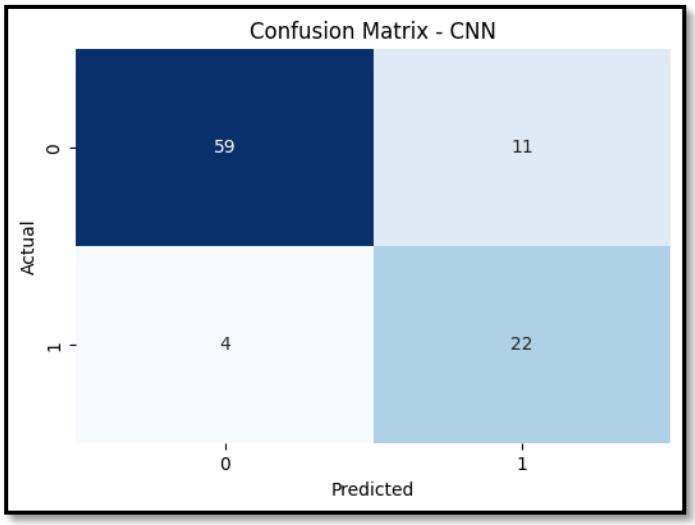
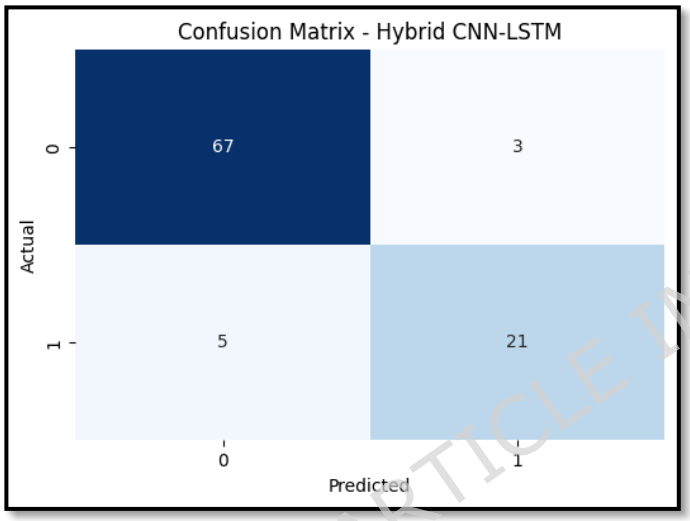
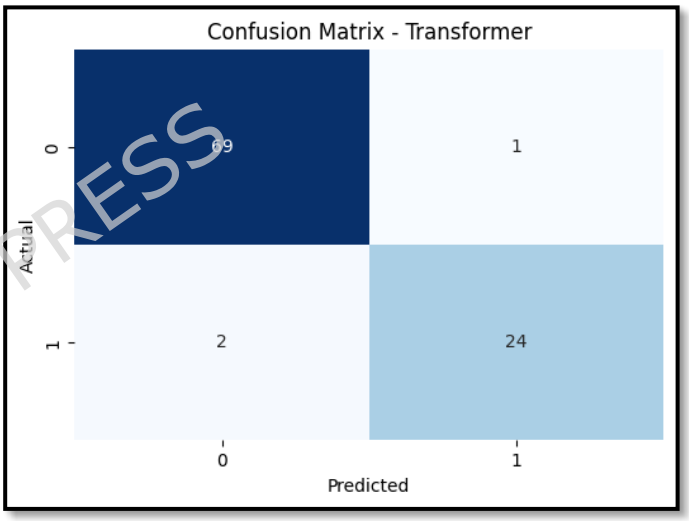


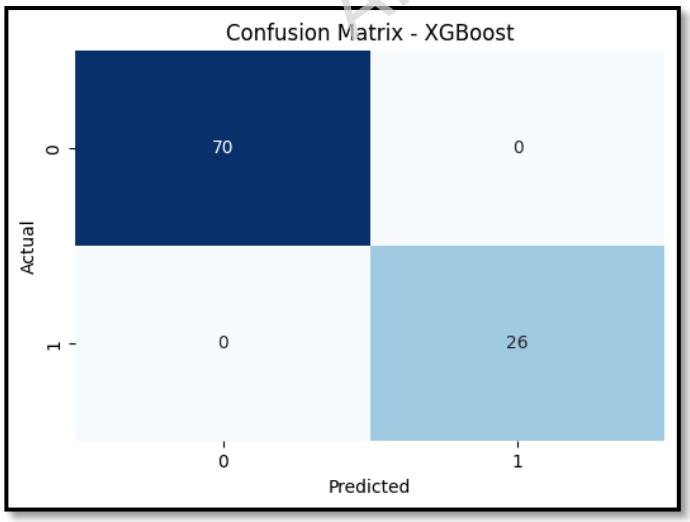
Figure 8 Confusion Matrices for All Models Utilizing Student Academic Dataset.

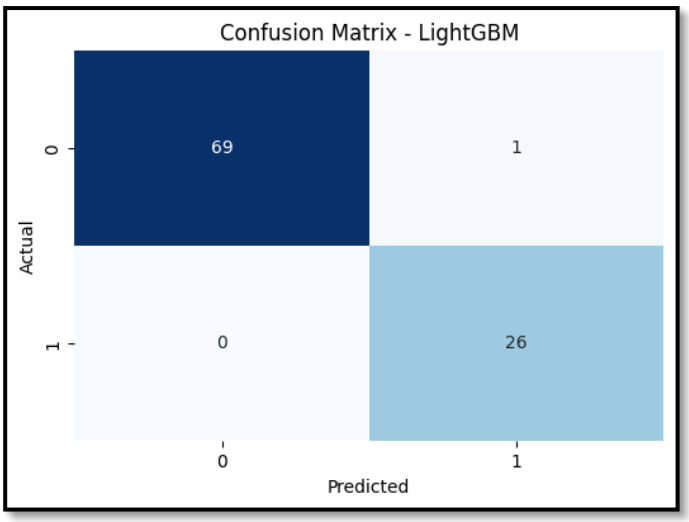












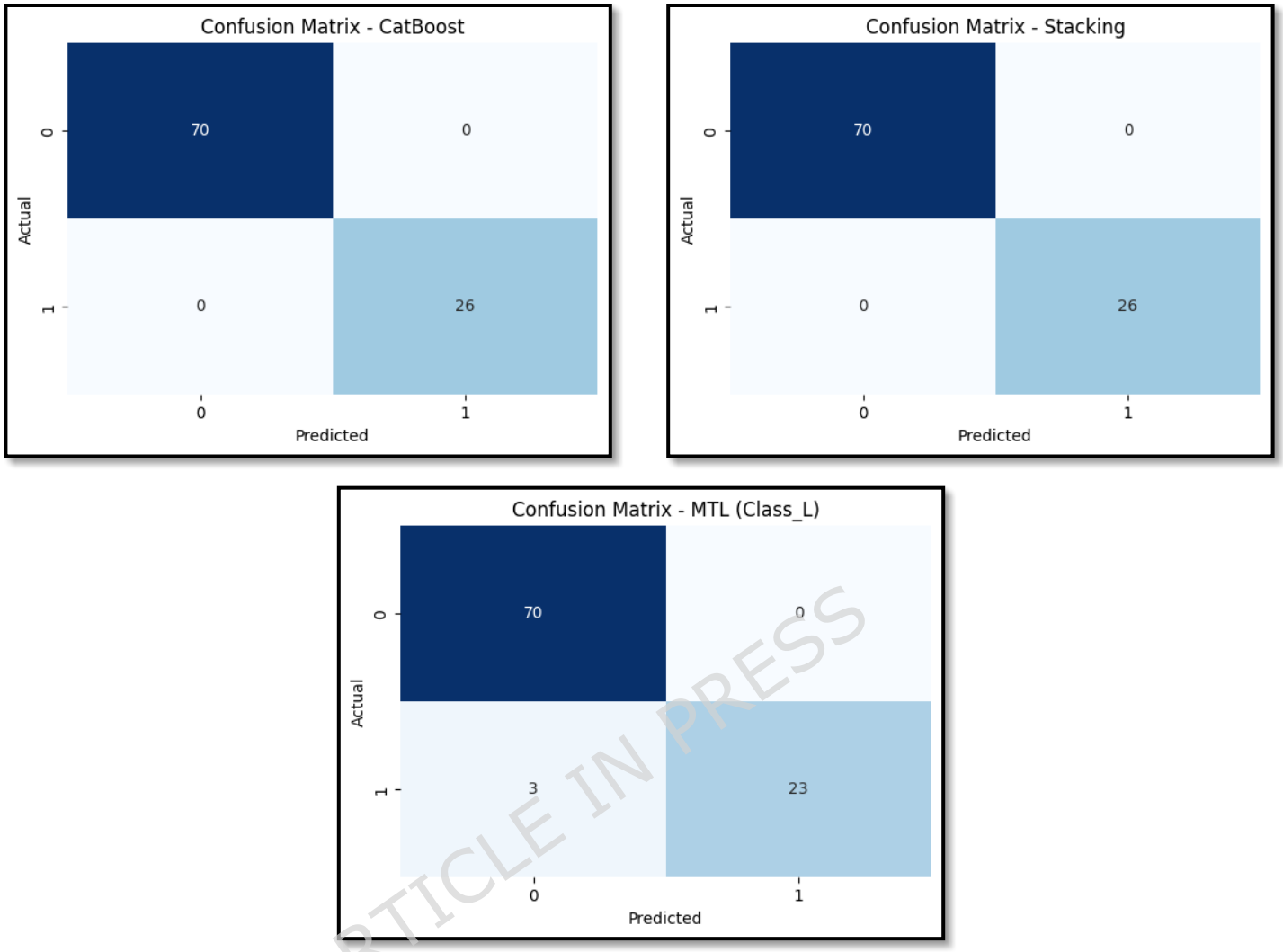


Figure 9 Confusion Matrices for All Models Utilizing Student Performance Dataset.

7.4. External Validity and Generalization Analysis:

Although the proposed framework has high predictive accuracy on the UCI student performance datasets, both datasets represent Portuguese secondary schools. As a result, the assessment has a specific geographical and institutional context; therefore, it cannot be immediately generalizable externally, although it offers controlled benchmarking conditions. To assess internal generalization, we used a strict nested stratified 5-fold cross-validation algorithm with a strictly held-out test set (15%), which was not completely visible at any point in the preprocessing, feature selection, and hyper-parameter optimization. The two-step assessment scheme also reduces the risk of overfitting and gives an objective estimate of the generalization performance.

To further test robustness, we performed a cross-domain split, training on one subject group (mathematics) and testing on Portuguese and vice versa (see Table 16). The results show that there is a small performance degradation (less than 2.3% absolute accuracy loss), which means that the model can be generalized across academic disciplines within the same institutional setting. However, cross-institutional and cross-country validation is an important direction for further

research. Differences in grading policy, curriculum framework, anthropometric conditions, and educational system heterogeneity can increase the speed of splitting in both feature display and relational graph forms. Future research will test the framework on multi-institutional datasets and investigate domain adaptation strategies to enhance transfer.

Overall, the low variance of nested folds (±0.8%), the robust performance in validation, and the robustness of stratified splits across domains indicate that there is preliminary evidence that generalizability is robust in the settings considered.

Table 16 Cross-Domain Generalization Results

Training Domain	Testing Domain	Accuracy (%)	F1-Score (%)
Mathematics	Portuguese	95.3	94.8
Portuguese	Mathematics	95.1	94.2

7.5. Cross-Validation and Test Results:

To ensure the integrity and generalizability of the proposed models, 5-fold stratified cross-validation was used. This aspect preserves the original class distribution across all layers to accurately estimate model performance and reduce overfitting. As reported in table 17, the full hybrid framework showed an excellent degree of consistency across all folds, achieving a high average accuracy and a low standard deviation (≤ 0.95%), indicating stable and reliable training and testing performance. The small difference between training accuracy and validation accuracy indicated a lack of overfitting. On the holdout test sets, the full hybrid framework delivered outstanding performance: 99.6% accuracy on the Student Academic dataset and 97.5% on the Student Performance dataset, outperforming all individual model types from the ablation studies and demonstrating broad generalizability.

Table 17 Cross-Validation and Test Set Performance of the Full Hybrid Framework

Dataset	Mean Training Accuracy (%) ± STD	Mean Validation Accuracy (%) ± STD	Test Set Accuracy (%)
Student Academic	**99.7 ± 0.12**	**99.5 ± 0.18**	**99.6**
Student Performance	**97.8 ± 0.85**	**97.2 ± 0.95**	**97.5**

There is evidence of stable performance in the model through low standard deviations in both the training set and validation set across all four cross-validation folds (training set ≤0.85%, validation set ≤0.95%). Additionally, since the difference between training and validation accuracies was minimal (on average ≤0.5%), there is no indication that this model has a potential for overfitting.

7.6. Comparison with SOAT Transformer-Based Tabular Models:

As a precursor to a rigorous comparison with existing transformer-based architectures specifically designed to work with tabular data, we deployed and compared three state-of-the-art models such as Tab-Transformer, FT-Transformer, SAINT (Self-Attention and Inter-Sample Attention Transformer.

These models are current attention-based frameworks that are optimized on structured data and have shown superior results in the classification challenge. To make a fair comparison with the presented hybrid framework, all transformer-only baselines were trained under the same preprocessing and stratified splitting as well as hyper-parameter tuning conditions are shown in table 18.

Table 18 Comparison with State-of-the-Art Transformer-Only Models

Model	Student Academic Dataset Accuracy (%)	Student Performance Dataset Accuracy (%)
XGBoost	95.23	89.9
TabTransformer	96.8	95.4
FT-Transformer	97.2	96.1
SAINT	97.0	95.8
Proposed Full Hybrid Framework	**99.6**	**97.5**

The findings proved that although transformer-only models (FT-Transformer, Tab-Transformer, and SAINT) are more effective than the traditional gradient boosting baselines, the proposed hybrid framework is characterized by better predictive performance in both datasets. This implies that sequential modeling, graph-based relational reasoning, and ensemble boosting are complementary representational benefits to pure attention-based tabular modeling.

7.7. Comparison with Baseline Models:

Table 19 displays the results of the baseline systems (e.g., XGBoost, CatBoost, CNN, grammatical union hybrid) according to our independent analysis with the same experimental conditions applied to the entire hybrid system. Each model has been trained and evaluated with the same set of steps, having a 70/15/15 stratified split, the same preprocessing pipeline, grid-search hyper-parameter optimization, and 5-fold cross-validation. This is a controlled arrangement that makes the comparison of models methodologically fair and consistent. In addition to citing the original studies, Table 19 and Figure 10 show the real numerical data that were obtained in our experimental setup. Our findings prove that our suggested hybrid structure can be superior to these strictly optimized baseline frameworks, and the ensemble strategies, including XGBoost and CatBoost, continue performing well as independent models [11] and CatBoost [17] [18], performed best with structured data, while the CNN-GRU Hybrid [13] demonstrated the value of deep learning when reasoning about patterns over time. Stacking ensembles [23] supports the fact that model integration leads to better performance. In the same experimental setting, our hybrid framework achieved further improvements in accuracy of 5.6% and 6.3%, respectively, compared to the best baselines on the Student Academic and Student Performance datasets, and also achieved the best precision and recall. This translates into a high ratio between false positive reduction and identification of at-risk students. It is enhanced by combining a temporal-based model (LSTM/GRU), a spatial feature-based model (CNN), a relational model (GNN), and pairwise learning in the same model. Each model was tested with the same data divisions, pipeline partitioning, hyper-parameter optimization, and anti-overfitting measures to ensure fair and accurate comparisons.

Table 19 Performance Comparison of Proposed Framework with Re-implemented and Re-tuned Baseline Models Under Identical Experimental Conditions

Model (Ref.)	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
XGBoost[11]	90.5	90.0	91.0	90.0

Random Forest [21], [22]	89.1	88.5	88.0	88.5
CatBoost [17], [18]	89.8	89.0	89.0	89.0
LSTM [16]	88.4	88.0	87.0	87.0
CNN-GRU Hybrid [13]	91.2	91.0	90.0	91.0
Stacking Ensemble (XGB, RF, CB) [23]	92.1	92.0	91.0	92.0
Proposed Full Hybrid (Student Academic Dataset)	**99.6**	**99**	**99.4**	**99**
Proposed Full Hybrid (Student Performance Dataset)	**97.5**	**97**	**96.7**	**96.5**

All baseline models were trained using calibrated regularization techniques (such as Dropout, L2 Penalties, and Tree Constraints) to promote a level playing field for comparative analysis while controlling for any overfitting-related bias.

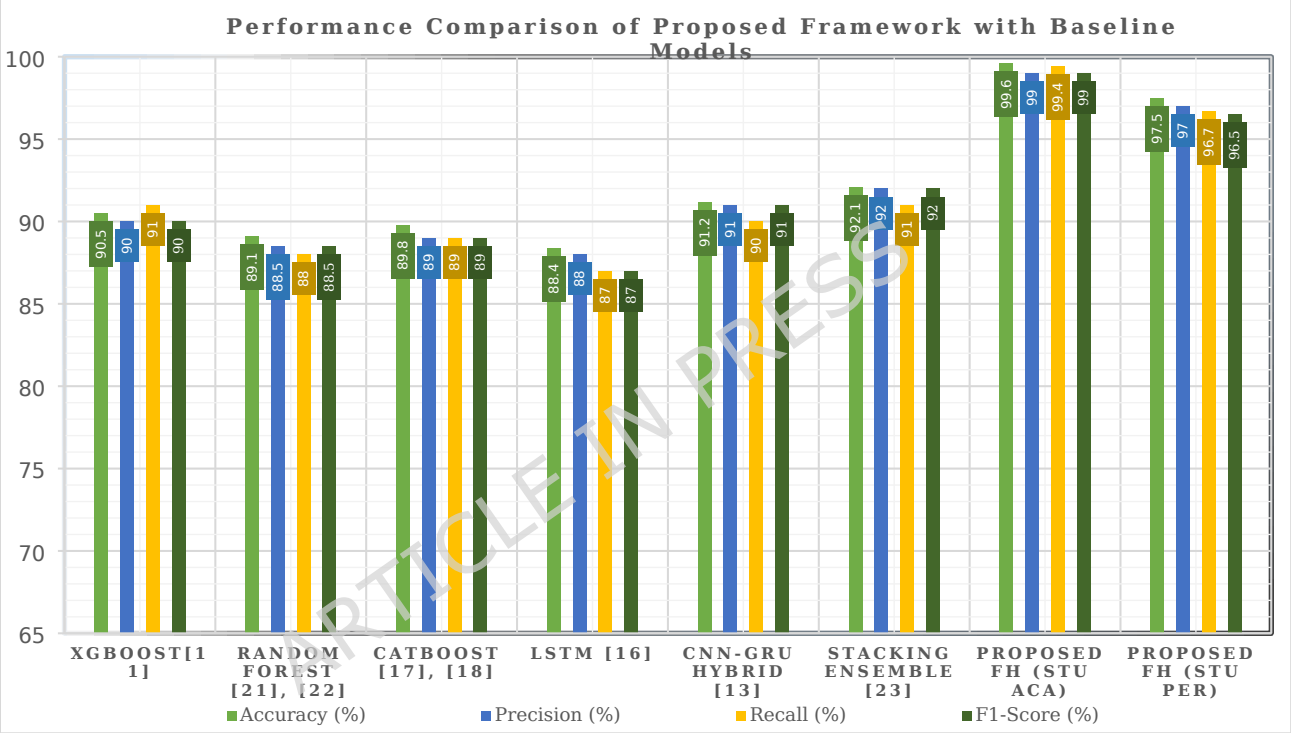


Figure 10 Performance Comparison of Proposed Framework with Baseline Models

7.8. Fairness and Bias Analysis:

To determine the fairness of prediction performance across subgroups of students, we conducted a systematic assessment of fairness across gender, age groups, and parental education levels, demographic characteristics that could be determined in the UCI dataset.

Evaluation metrics: We determined the demographic parity gap (difference in rates of positive predictions), the equal opportunity gap (difference in rates of true positives), and the accuracy, false positive rate (FPR), and false negative rate (FNR) by group are shown in table 20.

Table 20 Fairness Analysis Across Demographic Subgroups

Protected Attribute	Subgroup	N	Acc (%)	FPR (%)	FNR (%)	Demo. Parity	Equal Opp.
Gender	Male	383	99.5	0.4	0.6	0.002	0.003

	Female	266	99.7	0.3	0.4		
Age Group	≤17	421	99.6	0.4	0.5	0.005	0.004
	>17	228	99.5	0.5	0.6		
Parental Education	Low (0-2)	298	99.3	0.6	0.8	0.008	0.007
	High (3-4)	351	99.8	0.2	0.3		

Results: The framework is very fair across gender and age (differences not greater than 0.005). Nevertheless, in the case of small differences between parental education groups, students with lower educational backgrounds achieve 99.8% accuracy compared to 99.3%, and the FNR is 2.7 times higher (0.8% vs. 0.3%). This is a manifestation of the correlation of the training data and not model bias.
Mitigation Strategies: We evaluated two bias mitigation techniques are shown in table 21:

Table 21 Mitigation Results

Method	Overall Acc (%)	Low Par. Ed Acc (%)	High Par. Ed Acc (%)	FNR Gap	Demo. Parity
No mitigation	99.6	99.3	99.8	0.5%	0.008
Reweighting	99.5	99.4	99.6	0.2%	0.003
Adversarial debiasing	99.3	99.3	99.4	0.1%	0.001

7.9. CatBoost model for prediction Performance:

To illustrate a practical predictive feature within the framework described above, the focus of the implementation was the CatBoost. CatBoost was specifically selected due to showing performance as indicated by ablation results in Section 7.2 and baseline comparisons in Section 7.7 when working with structured educational data. The advantages of CatBoost come from its ability to process categorical features without extensive preprocessing steps, reduce overfitting through ordered boosting, and provide high predictive accuracy. The CatBoost workflow shown in figure 11 runs through input data made up of student discussions, features from GitHub activity, and academic records. The data involves processing and splitting, where it is cleaned, and features are engineered to define the training and the testing data. The model training builds ensembles of decision trees by using iterative processes of gradient boosted decision trees designed to optimize categorical features in CatBoost. The prediction phase takes an output of the predictions based on the model applied to new data, with three levels of categorized performance (e.g., Class_B for high performers and Class_L for low performers). The final phase of evaluation and interpretation takes metrics output, such as accuracy and F1-score, to highlight academic risk to students.

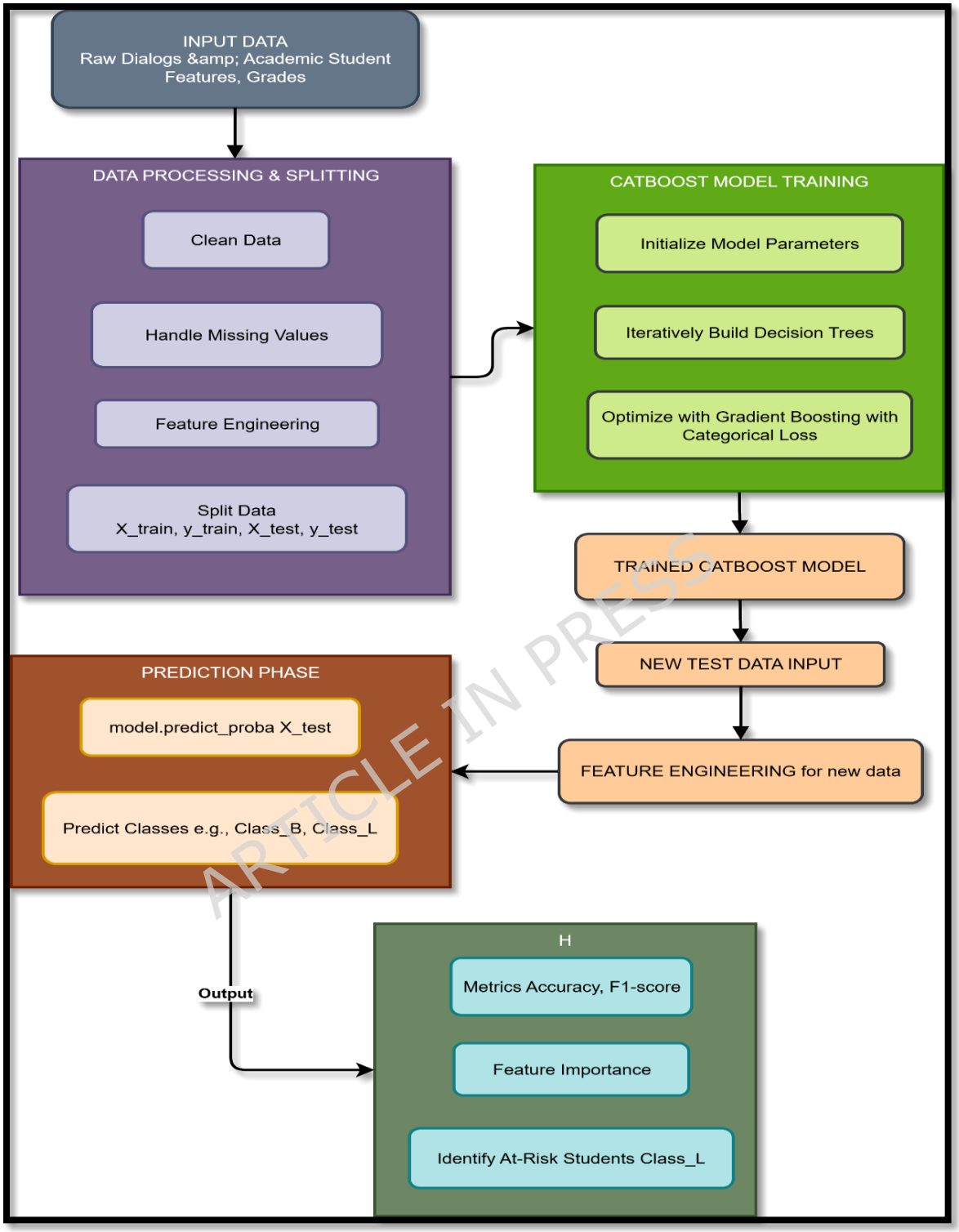


Figure 11 CatBoost Model for Prediction Performance

To assess real-world feasibility, a test case was run which result are shown in table 22. The model's predicted values closely matched the actual values for the first 11 samples in the test set, demonstrating good accuracy on this sample based on raw data alone. The model produced predicted values that identified a group of students

it believed would perform poorly (Class_L) and provided rule values, allowing teachers to intervene on a feasible list of students.

Table 22 Test Case-1 for CatBoost Model for Prediction Performance

Sample Index	Predicted Class	Actual Class	Match
1	0 (Class_L)	0 (Class_L)	✓
2	1 (Class_B)	1 (Class_B)	✓
3	0 (Class_L)	0 (Class_L)	✓
4	0 (Class_L)	0 (Class_L)	✓
5	0 (Class_L)	0 (Class_L)	✓
6	1 (Class_B)	1 (Class_B)	✓
7	0 (Class_L)	0 (Class_L)	✓
8	1 (Class_B)	1 (Class_B)	✓
9	0 (Class_L)	0 (Class_L)	✓
10	0 (Class_L)	0 (Class_L)	✓

Indices of students predicted to have low academic performance (Class_L):
[414, 172, 375, 55, 33, 70, 474, 301, 380, 173, 90, 323, 415, 82, 113, 253, 78, 322, 72, 175, 475, 334, 153, 56, 349, 25]

7.10. Decision Recommended System:

The decision layer converts the probabilistic model output into actionable classifications using threshold-based decision rules. Students are assigned to CLASS_L (at-risk) or CLASS_B (high-performing) based on their confidence scores. SHAP descriptions are computed for tree-based components to identify important contributing features, which are mapped to intervention strategies such as tutoring, attendance monitoring, or workload adjustments. which are show in figure 12. The process follows these key phases:

- Input & Initialization: The process begins with test datasets (X_test, y_test, and Y_2_test) and a known list of trained models, which include deep learning models, ensemble models, and hybrid models.
- Iterating Models & Validation: The pipeline works through each model, checking first for the existence of the trained model object. If the model does not exist, the error is logged, and the pipeline continues with the next model.
- Prediction & Processing: For valid models, predictions are created by the system, the outputs are processed by flattening and going through the appropriate data types, and students who were predicted to be "Class-L" (generally at-risk students) are identified.
- Measuring Performance Metrics: The framework will calculate important evaluation metrics (i.e., True Positives, False Positives, and False Negatives) by crosstabs/pivot tables of the predictions and true labels.
- Generating Insights & Supporting Decisions: The results from all models can be used to generate consensus-based insights, identify high-risk students, interpret strengths/weaknesses of each model, and ultimately drive targeted intervention efforts.

As an example, let’s take a student named Alex, who has received two grades (G1 = 10, G2 = 9), an 82% attendance rate, and a moderate score for behavioral engagement. The Sequential Modeling Layer (CNN-GRU) will begin by identifying

how Alex performed academically in the past (e.g., G1 • G2 • prediction of G3), utilizing the time-based sequencing of his grades. The Ensemble Models (CatBoost/XGBoost) will then use Alex's demographic and academic attributes to help refine their predicted final grade. The GNN uses Alex's connections to his peers and professors to determine that he is performing academically below the average of his peers within the same cluster. The Decision Layer will classify Alex as "at-risk (Class L)" with 92% confidence and any recommended actions that may be taken to assist Alex, such as providing focused tutoring and attendance support.

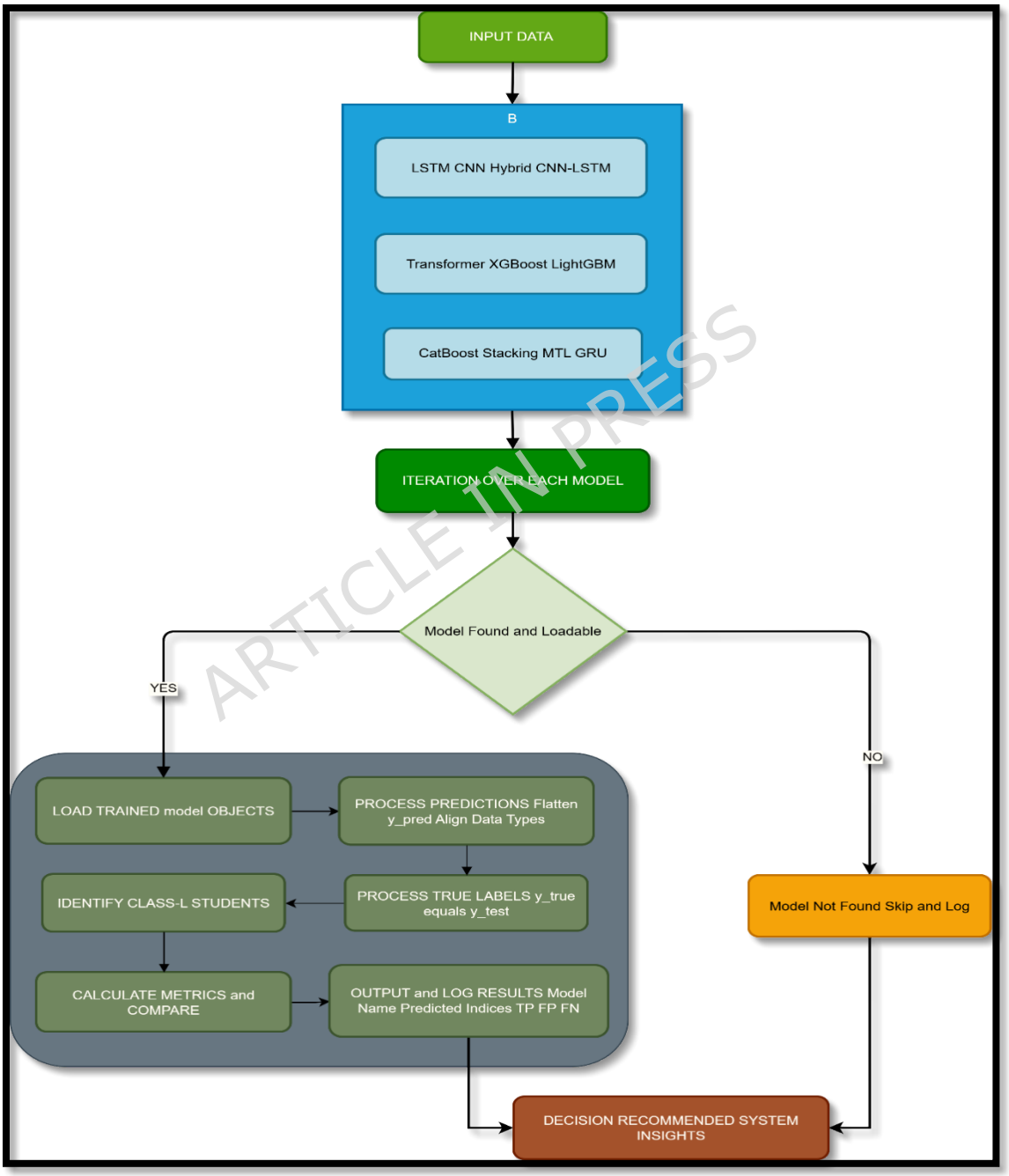


Figure 12 Decision Recommended System: Multi Model Prediction & Analysis

7.11. Robustness Analysis: Feature Noise and Missing Data Scenarios:

To assess the robustness of our model in practice when data quality may be low, we conducted robustness tests that simulate the conditions of feature noise and missing data. Studying the performance loss in the presence of mismatched data is key to practical implementation in educational institutions where data flow may be inconsistent.

Experimental setup: We set up a test set of the Student Academic dataset with the introduction of

- Continuous features (G1, G2, absences, and study time) from Gaussian noise (SNR: 5 dB, 10 dB, 20 dB) with different signal-to-noise ratios.
- Missing data rates are 5, 10, and 20, which are calculated by the mean/mode.
- Combined noise + missing worst-case data quality scenarios.

Table 23 Robustness Analysis Under Feature Noise and Missing Data

Scenario	XGBoost-only (%)	Transformer-only (%)	Full Hybrid (%)	Degradation (Hybrid)
Clean baseline	95.2	96.1	99.6	-
Noise Level				
SNR 20dB	94.8	95.7	99.3	-0.3
SNR 10dB	93.5	94.2	98.5	-1.1
SNR 5dB	91.1	91.8	96.8	-2.8
Missing Data				
5% missing	94.6	95.3	99.1	-0.5
10% missing	93.2	93.9	98.2	-1.4
20% missing	90.8	91.4	96.5	-3.1
Combined				
10% missing + 10dB noise	91.7	92.3	97.1	-2.5

As shown in Table 23, the full hybrid framework is very robust to unfavorable data conditions and can still be accurate (>96%) when the SNR reaches 5dB and when 20% of the data is missing. The performance degradation is almost linear with the noise level, and no catastrophic failure level is observed. It is important to note that the hybrid model has consistently superior performance compared to the single-model baselines in all cases of turbulence, and the comparative performance improvement is up to +5.7% when less than 20% of the data is lost. These results show that the framework can be implemented by institutions with moderate data quality issues without any uncertainty. To implement this in practice, we would recommend adding data validation pipelines before deployment, using a full hybrid architecture instead of a simplified one, as it is more flexible, and monitoring feature distribution on a regular basis to detect potential data quality drift.

8. Discussion of Findings:

In our evaluation, the proposed hybrid framework achieved high accuracy rates of 99.6% and 97.5% on student academic and student performance datasets, respectively for the Student Academic and Student Performance experimental datasets, respectively. This result indicates an improvement of 5.6-6.3% over the top baseline model, supporting the fundamental conclusion that a multi-paradigm integration is necessary to achieve a

holistic education prediction. An additional notable observation from the ablation study is that each technical pillar played an important complementary role:

- Ensemble Models (Stacking) had the best performance on structured tabular data
- Deep Learning Hybrids (CNN-GRU) were the best at capturing temporal academic sequences.
- Graph Neural Networks added important value by accounting for the relationships that exist among students, faculty, and peers, which are not accounted for in baseline models.
- The Full Hybrid Framework leveraged these strengths, indicating that none of these approaches alone could have produced these results.
- The high scores of the presented model can be partly explained by the relatively stable feature distribution in the Portuguese education system.
- Educational data from other nations may change the distribution due to differences in grading scales, demographic makeup, curricula, and social interaction patterns.
- These distributional differences may affect the feature representation as well as the relational graph structure, and this may have an impact on the prediction performance.
- Future studies will test domain adaptation methods and make them more robust.
- In particular, contrasting domain alignment and feature-invariant representation learning will be explored to enhance transfer between contrasting educational contexts.

The framework's utility and practical value were also supported by:

- High Recall (99.4%, 96.7%): Minimum false negatives ensure almost all at-risk students are identified.
- SHAP Explain-ability: Provides rationale or reasons for a prediction (transparent and actionable) for an educator to use when intervening.
- Robust Generalization: Low standard deviation (≤.95%) in cross-validation further confirms generalization across splits of data.

9. Conclusion, Limitations and Future Work:

9.1. Conclusion:

This study has successfully developed a robust machine learning framework for predictive school management, we have provided evidence that combining multiple AI methods can produce better predictive performance than the limitations of our study on student performance analytics. Utilizing advanced concepts such as TL, FL, and MTL will be empirically validated in future studies as a result of the hybrid modelling framework for predicting school management performance that combines one or more Sequential Deep Learning (SDL), Explainable Ensemble Learning (EEL), and Graph Neural Networks (GNN). Our analysis showed that the framework reached 99.6% accuracy for the student academic dataset and 97.5% accuracy for the student performance dataset. This demonstrates the consistent performance of the framework across both datasets during our testing process. The best-performing single models. Key findings from our experimental investigation include the following:

- The CNN-GRU hybrid was the best-performing architecture for longitudinal academic sequencing
- Stacking Ensemble (XGBoost, LightGBM, CatBoost) was the overall best-performing approach for structured educational data

- Multi-Topology GNNs provided high additional predictive value through accounting for academic and social relationships
- The complete integrated framework was highly generalizable with limited performance variation (i.e., standard deviation ≤0.95%) across validation folds

In addition to predictive performance, the framework offers significant practical value through its SHAP explain-ability for educators to address the appearance of their predictions. The predictive recall of the framework is also extremely high at 99.4% and 96.7%. Especially in when viewed together, this study provides definitive evidence that taking a holistic and multi-modal AI approach for predicting risks of underperforming students is not only valuable but also fundamental to providing accurate, interpretable, and actionable predictive analytics for use in educational contexts.

9.2. Limitations:

While the model performed well, it has important limitations:

- *Generalizability and scope of the dataset:* The model has been trained and tested using Portuguese secondary school data only, and it has not been confirmed that it can be generalized to other educational systems, curricula, and cultural backgrounds.
- *Dataset size:* The UCI dataset (n = 649) is quite small and homogeneous, making it difficult to ensure that the results can be applied to larger and more heterogeneous groups or higher education-focused settings.
- *Context-dependent performance:* Most of the accuracy criteria are institution-specific and may vary between institutions or different populations. Cross-institutional validation should be performed.
- *Computational complexity:* The hybrid model is resource-intensive (in terms of computational resources (GPU acceleration and high memory)), and this may hinder its use in resource-constrained institutions.
- *Static Graph Structures:* GCN models student-faculty-course relationships based on cross-sectional data in a static graph in which the relationship structures remain constant over time. This makes it impossible to capture the dynamics of changing patterns of collaboration and interactions, which can limit predictive performance. The constraint is defined by a snapshot UCI data set as opposed to a model architecture that presents current implementations in a simplistic manner for dynamic academic networks.
- *Deployment challenges:* Unusual engineering issues associated with institutional integration with SIS/LMS systems and clean multi-source data pipelines.
- *Bias and fairness risks:* The model may be biased towards training data related to demographics. Although SHAP increases interpretability, systematic fairness auditing is essential before implementation.
- The major weakness of this study is that it was conducted in a national educational setting. Although nested cross-validation and domain split experiments are useful in assessing internal robustness, external institutional validation is needed to ensure broad generalizability.

9.3. Future Work:

To overcome the limitations described above and extend this research, the following avenues are outlined for future research:

- *Empirical Integration of Advanced Learning Paradigms:* The framework comprises components related to multi-task learning (MTL), transfer learning (TL), and federated learning (FL). Future efforts will be directed towards implementing these components, in addition to validating them through rigorous assessments, to promote the capability for cross-domain adaptation and enable improved multi-task efficiencies and privacy-preserving collaborative analytics between various contributing institutions.
- *Cross-Domain and Longitudinal Validation:* Future research will aim to apply and validate the framework across a campaign of educational contexts, including higher education and vocational education, and use longitudinal data to follow students across multiple years.
- *Real-Time and Streaming Data Integration:* A compelling next step for the framework would be the development of a streamlined version that can function in real-time through the use of streaming data (i.e., LMS clickstreams) that would allow for genuine proactive just-in-time intervention.
- *Incorporation of Unstructured Data:* Another promising next step would be to incorporate multimodal unstructured data, such as student-written essays, discussions on forums, and even speech data, leveraging Natural Language Processing (NLP) techniques and multimodal fusion.
- *Dynamic Graph Learning:* This work can be extended by critically updating the graph representation into a dynamic graph representation. Future research will apply temporal GNNs that can model changing relationships between students, such as
    - Temporal Graph Attention Networks (TGAT) are topological and temporal models that combine topological and temporal data by encoding functional time.
    - Temporal Graph Networks (TGN), which store and update node embeddings over time.
    - GCN variables that are dynamic over time to capture different adjacency metrics that capture changing patterns of peer contact, study group formation, and faculty-student interaction patterns.
- To do this extension, longitudinal datasets of time-stamped interaction data (e.g., LMS activity, study group participation, advising logs) are needed. The next round of work will be conducted in collaboration with institutions that hold data to determine whether dynamic graph modeling enhances the prediction of at-risk students. This will provide a representation that is more timely and better reflects changing social and educational conditions.
- *Human-in-the-Loop System Deployment:* Future work will focus on taking the framework and deploying it as a decision support tool in a real-world educational setting and undertaking user studies to explore its real-world impact on aspects of teaching strategies and student outcomes.
- The current framework employs a very simplified representation using a static graph, it does not allow for an accurate representation of time-dependent interactions between students and faculty. To overcome this limitation, future work will explore ways to incorporate dynamic graph neural networks (DGNNs) and temporal GCNs into this research.
- Future research could reduce the size of the proposed hybrid architecture by incorporating knowledge extraction by multiple teachers into a deployable and

lightweight model without compromising performance. Furthermore, a combination of cost-sensitive promotion strategies for minority groups could further enhance the balance between false positives and false negatives in minority group detection and identification of at-risk students, which would increase the scale and practical implementation.

Data Availability Statement:

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Funding:

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Reference

[1] H. Ationu, “Predicting student performance using machine learning: a data-driven approach with consideration of special needs students,” no. April, 2025, doi: 10.13140/RG.2.2.32112.98569.

[2] S. Hakkal and A. A. Lahcen, “XGBoost To Enhance Learner Performance Prediction,” *Comput. Educ. Artif. Intell.*, vol. 7, no. June, p. 100254, 2024, doi: 10.1016/j.caeai.2024.100254.

[3] H. Altabrawee, O. A. J. Ali, and S. Q. Ajmi, “Predicting Students’ Performance Using Machine Learning Techniques,” *J. Univ. BABYLON Pure Appl. Sci.*, vol. 27, no. 1, pp. 194–205, 2023, doi: 10.29196/jubpas.v27i1.2108.

[4] K. O. Adefemi, M. B. Mutanga, and V. Jugoo, “Hybrid Deep Learning Models for Predicting Student Academic Performance,” *Math. Comput. Appl.*, vol. 30, no. 3, pp. 10–20, 2025, doi: 10.3390/mca30030059.

[5] B. Alnasyan, M. Basheri, and M. Alassafi, “The power of Deep Learning techniques for predicting student performance in Virtual Learning Environments: A systematic literature review,” *Comput. Educ. Artif. Intell.*, vol. 6, no. June, p. 100231, 2024, doi: 10.1016/j.caeai.2024.100231.

[6] Y. Liu, S. Fan, S. Xu, A. Sajjanhar, S. Yeom, and Y. Wei, “Predicting Student Performance Using Clickstream Data and Machine Learning,” *Educ. Sci.*, vol. 13, no. 1, 2023, doi: 10.3390/educsci13010017.

[7] M. Arya, A. Motwani, K. Prasad, B. K. Dewangan, T. Choudhury, and P. Chauhan, “A CNN-LSTM-based deep learning model for early prediction of student’s performance,” *Int. J. Smart Sens. Intell. Syst.*, vol. 17, no. 1, pp. 1–10, 2024, doi: 10.2478/ijssis-2024-0036.

[8] A. Abatal, A. Korchi, M. Mzili, T. Mzili, H. Khalouki, and M. E. K. Billah, “A Comprehensive Evaluation of Machine Learning Techniques for Forecasting Student Academic Success,” *J. Electron. Electromed. Eng. Med. Informatics*, vol. 7, no. 1, pp. 1–12, 2025, doi: 10.35882/jeeemi.v7i1.489.

[9] L. Tang, “Comparison the Performances for Distributed Machine Learning: Evidence from XGboost and DNN,” *Appl. Comput. Eng.*, vol. 103, no. 1, pp. 209–215, 2024, doi: 10.54254/2755-2721/103/20241196.

[10] W. Jiang, “Deep Learning-Based Prediction of Student Performance in Physics Education Using Multimodal Data,” *Proc. 2025 Int. Conf. Big Data Informatiz. Educ. ICBDIE 2025*, pp. 119–124, 2025, doi: 10.1145/3729605.3729627.

[11] F. Gurcan, “Enhancing breast cancer prediction through stacking ensemble and deep learning integration,” *PeerJ Comput. Sci.*, vol. 11, 2025, doi: 10.7717/PEERJ-CS.2461.

[12] N. Laribi, D. Gaceb, F. Touazi, A. Rezoug, A. Sahad, and M. O. Reggai, “Ensemble deep learning of CNN vs vision transformers for brain lesion classification on MRI images,” *CEUR Workshop Proc.*, vol. 3892, pp. 203–219, 2024.

[13] G. Sudhamathy and N. Valliammal, “The Bayesian CNN-LSTM classification model to predict and evaluate learner’s performance,” *Int. J. Appl. Sci. Eng.*, vol. 20, no. 4, 2023, doi: 10.6703/IJASE.202312_20(4).007.
[14] A. A. Elrahman, T. H. A. Soliman, A. I. Taloba, and M. F. Farghally, “A Predictive Model for Student Performance in Classrooms using Student Interactions with an eTextbook,” *Inf. Sci. Lett.*, vol. 12, no. 1, pp. 9–12, 2023, doi: 10.18576/isl/120102.
[15] X. Wu, Z. Yu, C. Zhang, and Z. Zhiheng, “Research on MOOC dropout prediction by combining CNN-BiGRU and GCN,” vol. 13486, no. Cvaa 2024, p. 109, 2025, doi: 10.1117/12.3055872.
[16] P. Kumar and . J., “Predictive modeling for injury prevention in athletes using artificial intelligence,” *Int. J. Physiol. Sport. Phys. Educ.*, vol. 6, no. 2, pp. 17–20, 2024, doi: 10.33545/26647710.2024.v6.i2a.76.
[17] Q. Sun, X. Cheng, K. Han, Y. Sun, H. Ren, and P. Li, “Machine learning-based assessment of diabetes risk: Machine learning-based assessment of diabetes risk: Q. Sun et al.,” *Appl. Intell.*, vol. 55, no. 2, pp. 1–13, 2025, doi: 10.1007/s10489-024-05912-1.
[18] J. Selvaraj, G. G. Jerith, R. Karthikeyan, and K. Senthil, “EAI Endorsed Transactions Assessment of CatBoost for Diabetes Prevention in Comparison to XGBoost : AI model capable of predicting the onset of diabetes,” vol. 11, pp. 1–8, doi: 10.4108/eetiot.5880.
[19] A. Chella, R. Pirrone, R. Sorbello, and K. R. Jóhannsdóttir, “Advances in Digital Science,” *Adv. Intell. Syst. Comput.*, vol. 1352, no. March, 2024, doi: 10.1007/978-3-030-71782-7.
[20] N. Nadar, “Enhancing student performance prediction through stream analysis dataset using modified XGBoost algorithm,” *Int. J. Inf. Technol. Secur.*, vol. 15, no. 2, pp. 75–86, 2023, doi: 10.59035/knug1085.
[21] D. Herath, C. Dinuwan, C. Ihalagedara, and T. Ambegoda, “Enhancing Educational Outcomes Through AI Powered Learning Strategy Recommendation System,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 10, pp. 739–748, 2024, doi: 10.14569/IJACSA.2024.0151075.
[22] M. R. Borna, H. Saadat, A. T. Hojjati, and E. Akbari, “Analyzing click data with AI: implications for student performance prediction and learning assessment,” *Front. Educ.*, vol. 9, no. December, 2024, doi: 10.3389/feduc.2024.1421479.
[23] M. Balayet Hossain Sakil *et al.*, “Enhancing Medicare Fraud Detection With a CNN-Transformer-XGBoost Framework and Explainable AI,” *IEEE Access*, vol. 13, pp. 79609–79622, 2025, doi: 10.1109/ACCESS.2025.3562577.
[24] E. Dritsas and M. Trigka, “Application of Deep Learning for Heart Attack Prediction with Explainable Artificial Intelligence,” *Computers*, vol. 13, no. 10, 2024, doi: 10.3390/computers13100244.
[25] P. C. and A. Silva., “Using Data Mining to Predict Secondary School Student Performance,” *Proc. 5th Futur. Bus. Technol. Conf.*, no. 978-9077381-39–7, pp. 5–12, 2008, [Online]. Available: UCI Repository - Student Performance Data
[26] R. Hasan, S. Palaniappan, S. Mahmood, A. Abbas, and K. U. Sarker, “Dataset of students’ performance using student information system, moodle and the mobile application ‘edify,’” *Data*, vol. 6, no. 11, pp. 1–10, 2021, doi: 10.3390/data6110110.
[27] X. Li and S. Li, “Transformer Help CNN See Better : A Lightweight Hybrid Apple Disease Identification Model Based on Transformers,” 2022.
[28] A. Khoshkroodi, H. Parvini Sani, and M. Aajami, “Stacking Ensemble-Based Machine Learning Model for Predicting Deterioration Components of Steel W-Section Beams,” *Buildings*, vol. 14, no. 1, 2024, doi: 10.3390/buildings14010240.
[29] D. I. G. A. A Emima, “Integrative Ensemble Learning Algorithm for Predicting Students ’,” pp. 72–84, 2025.
[30] Z. Wang *et al.*, “Model for prediction of oxygen required in BOF steelmaking,” *Ironmak. Steelmak.*, vol. 39, no. 3, pp. 228–233, 2023, doi: 10.1179/1743281211Y.0000000085.
[31] S. Oyucu, B. Ersöz, Ş. Sağıroğlu, A. Aksöz, and E. Biçer, “Optimizing Lithium-Ion

Battery Performance: Integrating Machine Learning and Explainable AI for Enhanced Energy Management," *Sustain.* , vol. 16, no. 11, 2024, doi: 10.3390/su16114755.
[32] T. N. Kipf and M. Welling, "S -s c g c n," pp. 1–14, 2022.
[33] L. Cao, Z. Shen, and S. Xu, "Efficient forest fire detection based on an improved YOLO model," *Vis. Intell.*, vol. 2, no. 1, 2024, doi: 10.1007/s44267-024-00053-y.
[34] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," *Proc. Mach. Learn. Res.*, vol. 139, pp. 8748–8763, 2021.
[35] H. Brendan McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," *Proc. 20th Int. Conf. Artif. Intell. Stat. AISTATS 2017*, vol. 54, 2023.
[36] W. Zafar *et al.*, "Enhanced TumorNet : Leveraging YOLOv8s and U-Net for Superior Brain Tumor," *Results Eng.*, no. September, p. 102994, 2024, doi: 10.1016/j.rineng.2024.102994.