Journal of Artificial Intelligence in Information Technology (JAI-IT)
Vol. 1 No. 1 June 2026
E-ISSN: XXXX-XXXX; P-ISSN: XXXX-XXXX, Pages 171-190
DOI: https://doi.org/10.51903/92j5wj58

A Machine Learning-Based Early Warning System for Student Performance Prediction: System Development and Empirical Evaluation in Higher Education

Lia Handayani*[1], Priyadi[1]
Email: liahandayani@stekom.ac.id (1), priyadi.ltr@gmail.com (2)
Orcid: https://orcid.org/0009-0001-9872-8037 (1), https://orcid.org/0000-0002-6554-854X (2)
[1]*Deptment of Design, Faculty of Fine Arts and Design, ISI Bali, Bali, Indonesia, 80235*
[2]*Department of Computer System, Faculty of Academic Study, Universitas Sains dan Teknologi Komputer, Semarang, Indonesia, 50192*
*Corresponding Author

Abstract
Academic failure and student attrition remain structural challenges in higher education, yet the behavioral and academic data generated through digital learning environments are largely underexploited for early risk identification. This study presents the design, development, and empirical evaluation of a machine learning-based early warning system (EWS) for predicting student academic performance, deployed as a fully operational web-based application within the information technology infrastructure of a higher education institution in Indonesia. A dataset of 1,240 student records spanning academic years 2021–2024 was constructed by integrating static academic attributes extracted from the institutional Academic Information System (SIAKAD) with dynamic behavioral features derived from a Moodle-based Learning Management System (LMS), including weekly login frequency, assignment submission lead time, quiz attempt rate, and forum participation count. Four supervised classification algorithms Logistic Regression, Random Forest, Gradient Boosting, and Long Short-Term Memory (LSTM) were trained and benchmarked under stratified 10-fold cross-validation with SMOTE-based class balancing. The Gradient Boosting classifier achieved superior performance across all evaluation metrics, attaining an accuracy of 89.1%, F1-Score of 0.850, and AUC-ROC of 0.931. SHAP-based feature attribution confirmed that LMS-derived behavioral variables, particularly weekly login frequency (SHAP = 0.241) and assignment submission lead time (SHAP = 0.187), contributed substantively to prediction quality beyond static academic records alone. The deployed system was evaluated by 32 academic advisors using the System Usability Scale (SUS) following a four-week observation period, yielding a mean score of 78.4, indicative of above-average practitioner usability.

Keywords: *early warning system; machine learning; Gradient Boosting; student at-risk; higher education*

I. INTRODUCTION

Student academic failure and dropout remain among the most persistent structural challenges confronting higher education institutions worldwide. Across both developed and developing economies, universities face mounting pressure to improve student retention rates, graduation outcomes, and overall instructional efficiency pressures that have intensified in the context of expanded enrollment policies and increasingly heterogeneous student populations [1], [2]. Traditional academic monitoring mechanisms, which rely predominantly on end-of-semester grade reviews or faculty-initiated referrals, have proven insufficient for detecting early indicators of student disengagement or academic difficulty before failure becomes irreversible [3], [4]. The

Received on March, 2026; Revised on April, 2026; Accepted on April, 2026; Published on June, 2026

delay embedded in such reactive processes means that students who exhibit identifiable risk signals frequently proceed through the academic calendar without receiving structured support. Meanwhile, higher education institutions accumulate substantial volumes of student data through learning management systems, attendance records, and continuous assessment platforms, yet this data remains largely underutilized in shaping timely and targeted pedagogical interventions [5], [6]. This structural disconnect between the abundance of institutional data and the absence of actionable analytical infrastructure represents a foundational problem that advances in machine learning and information technology can now address.

The widespread adoption of digital learning environments has generated unprecedented volumes of student behavioral and academic data, transforming higher education institutions into data-rich ecosystems with significant untapped analytical potential. Learning Management Systems such as Moodle, Canvas, and Blackboard continuously capture granular interaction data including login frequency, assignment submission timing, discussion forum activity, and assessment attempt patterns each of which encodes behavioral signals that are empirically associated with academic performance trajectories [3], [5]. Despite the availability of these data streams, the majority of institutions, particularly those in developing regions, have yet to establish systematic pipelines capable of translating raw LMS logs into structured risk profiles that support actionable decision-making. The consequence is a recurring institutional pattern in which students who display detectable warning signs proceed without intervention until academic failure is no longer preventable. Evidence from institutional studies has demonstrated that academic advisors equipped with structured predictive dashboards can intervene with measurably greater efficacy than those relying on informal observation and periodic grade reviews. The gap between what is technically feasible and what is operationally deployed is therefore not a purely technical problem it reflects a deeper challenge in IT systems integration, institutional governance, and the organizational readiness to adopt AI-driven decision-support tools in academic management contexts.

The application of machine learning to student academic performance prediction has attracted considerable scholarly attention over the past decade, reflecting the parallel growth of educational data mining as a recognized research discipline and the increasing accessibility of institutional student datasets. Early contributions to this field predominantly employed decision tree classifiers and logistic regression models trained on static demographic and historical academic variables, producing moderate predictive accuracy but demonstrating limited generalizability when applied across different institutional contexts [5], [7]. The subsequent adoption of ensemble learning methods most notably Random Forest and Gradient Boosting yielded meaningful improvements in predictive robustness, as these architectures exhibited greater resilience to feature heterogeneity

and class imbalance, both of which are structurally characteristic of academic performance datasets [7]. A further advancement came with the incorporation of temporal behavioral variables derived from LMS interaction logs, with studies consistently showing that engagement based features contributed substantially to model discriminative power beyond what historical academic records alone could provide [5], [6]. More recently, deep learning architectures including Long Short-Term Memory networks and attention based models have been applied to sequential student activity data, demonstrating the capacity to capture temporal dependencies in learning behavior that conventional classifiers are inherently unable to model [5], [7].

A parallel strand of research has shifted attention from model development toward the deployment and usability dimensions of predictive systems, recognizing that algorithmic accuracy alone is insufficient to ensure practical impact in institutional settings. Studies examining real world implementation have identified system usability, the interpretability of model outputs for non-technical academic advisors, and institutional trust in automated risk classification as critical determinants of whether predictive tools produce genuine changes in advising practice [8]. In response to interpretability concerns, Explainable AI frameworks particularly SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) have been applied to student performance models and shown to enhance advisor confidence in prediction-driven recommendations [5], [7]. Research situated in Southeast Asian higher education contexts has additionally documented the influence of socioeconomic variables and digital infrastructure constraints including device availability and internet access consistency as confounding factors in LMS-derived feature sets, a dimension that remains underrepresented in studies drawing predominantly on North American or European institutional data [9]. Across comparative evaluations, no single algorithm has demonstrated consistent superiority across institutional settings, reinforcing the argument that context specific model validation and iterative performance monitoring are essential components of responsible AI system deployment in education [5], [9].

Despite the breadth of existing scholarship, several substantive gaps remain that constrain both the scientific validity and practical applicability of current approaches. A predominant limitation is that most published studies operate exclusively at the level of offline model evaluation reporting classification metrics derived from historical datasets without demonstrating end-to-end integration of a predictive model within a functioning institutional IT infrastructure [8]. This distinction carries practical significance: a model validated under controlled experimental conditions may encounter substantial obstacles when deployed within real-world academic information systems characterized by inconsistent data pipelines, access control requirements, and asynchronous data refresh cycles. A related gap concerns system scope, as the majority of

existing studies assess predictive algorithms in isolation without addressing the interface architecture, advisor notification workflows, or institutional data governance mechanisms that would constitute a complete and operational early warning system [5], [6]. The evidence base available to institutions seeking to adopt such technologies is therefore dominated by proof-of-concept studies rather than documented deployments with verified operational outcomes. Furthermore, longitudinal evaluations that assess sustained model performance across multiple academic cycles, alongside empirical measurement of advisor interaction with deployed systems, remain sparse in the literature, leaving critical questions about real world system effectiveness unresolved [8].

This study addresses the identified gaps by designing, developing, and empirically evaluating a machine learning based early warning system for predicting student academic performance, implemented as a fully operational application within the information technology infrastructure of a higher education institution. The research is structured around three primary objectives: first, to construct a predictive model pipeline that integrates both static academic attributes and dynamic LMS-derived behavioral features, evaluated across multiple classification algorithms to identify the most accurate and contextually appropriate configuration; second, to develop a functional web-based application that embeds the predictive model within institutional data systems and provides academic advisors with interpretable risk classifications alongside actionable student profiles; and third, to conduct an empirical evaluation of the deployed system encompassing predictive performance metrics, system usability assessment using a validated instrument, and a structured institutional case study of advisor engagement with the platform. Each objective is designed to produce findings that are methodologically reproducible, practically deployable in comparable institutional contexts, and capable of informing both technical development and administrative decision-making in the governance of AI-based academic support systems.

The contributions of this research span both the technical and applied dimensions of the field. On the technical side, this study contributes a validated feature engineering framework that synthesizes heterogeneous academic and behavioral data sources into a unified predictive representation compatible with multiple supervised classification architectures, alongside a documented system architecture that serves as a replicable model for embedding machine learning capabilities into existing academic IT ecosystems without requiring full replacement of legacy infrastructure. On the applied side, the research contributes empirical evidence drawn from an authentic institutional deployment incorporating quantitative performance evaluation, usability testing with academic practitioners, and longitudinal observation of system use a combination

that is seldom present within a single published study in this domain. Collectively, these contributions are intended to narrow the distance between AI research and operational IT practice in higher education, offering both a technically grounded implementation model and an evidence base that institutions can consult when assessing the feasibility and governance requirements of adopting predictive academic monitoring systems.

II. RESEARCH METHOD

This study employs a quantitative research approach using an applied systems development design, combining predictive model construction, functional application development, and empirical system evaluation in a real institutional IT environment. The research follows a sequential pipeline consisting of five interconnected phases: dataset preparation and preprocessing, feature engineering, model training and selection, system development, and empirical evaluation. This design was selected because the central objective of the study is not merely to identify the most accurate classification algorithm, but to demonstrate that such an algorithm can be operationalized within a working academic information system and evaluated under authentic institutional conditions. The overall research workflow is illustrated in Figure 1.

Figure 1. Research workflow of the machine learning-based early warning system development and evaluation

A. Research Design

The research adopts a mixed-method system development design in which quantitative modeling is embedded within a broader applied case study framework. The quantitative component

encompasses predictive model training, cross-validation, and performance benchmarking across four classification algorithms, while the applied component involves the full cycle development of a web-based early warning application and its deployment within the target institution's IT infrastructure. A case study framing was deliberately chosen to capture the institutional context in which the system operates, recognizing that deployment conditions including data governance constraints, user roles, and infrastructure compatibility materially affect system performance in ways that offline model evaluation cannot replicate [10]. This dual layer design ensures that the findings are evaluable both at the level of algorithmic performance and at the level of operational system utility.

B. Population and Sample

The study was conducted at a higher education institution in Indonesia, with the target population comprising all undergraduate students enrolled across active study programs during the academic years 2021–2024. From this population, a purposive sampling strategy was used to select students with complete academic and LMS interaction records spanning at least two consecutive semesters. This criterion was selected to ensure temporal consistency in behavioral feature construction and to minimize bias arising from incomplete longitudinal data. The final dataset comprised 1,240 student records drawn from three academic programs, representing a sample size sufficient for training and validating classification models with meaningful statistical power under stratified cross-validation procedures. Students with withdrawn or on leave status during the observation period were excluded from the dataset to maintain the consistency of the academic trajectory labels used in model training.

C. *Data Sources and Data Collection Techniques*

Two primary data sources were employed in this study. The first source consisted of static academic records extracted from the institutional Academic Information System (SIAKAD), encompassing variables such as semester GPA, cumulative GPA, credit load, and attendance rate. The second source consisted of dynamic behavioral logs exported from the institution's Moodle-based Learning Management System, capturing time-stamped records of student interactions including course access frequency, assignment submission timestamps, quiz attempt counts, and discussion forum contributions. Data extraction was conducted through authorized API access coordinated with the institution's IT department, with all records anonymized prior to transfer to the research environment. A structured data validation procedure was applied following extraction to identify and handle missing values, duplicate entries, and anomalous timestamps, ensuring that the resulting dataset met the quality thresholds required for reliable model training.

D. Variables and Operational Definition

The study operationalizes three categories of variables, as summarized in Table 1. The outcome variable is academic performance risk classification, defined as a binary label at-risk or not-at-risk assigned to each student based on their end-of-semester GPA relative to the institutional passing threshold of 2.00 on a 4.00 scale. Static predictor variables include semester GPA, cumulative GPA, credit completion ratio, and attendance rate, each extracted from institutional academic records and measured on a continuous scale. Behavioral predictor variables are derived from LMS interaction logs and include weekly login frequency, average assignment submission lead time (in days prior to deadline), quiz attempt rate, and forum participation count, each aggregated at the semester level to produce a single representative value per student per semester.

Table 1. Variable Classification and Operational Definitions

Variable Category	**Variable Name**	**Measurement**	**Source**
Outcome	Academic Risk Label	Binary (at-risk / not-at-risk)	SIAKAD
Static Academic	Semester GPA	Continuous (0.00–4.00)	SIAKAD
Static Academic	Cumulative GPA	Continuous (0.00–4.00)	SIAKAD
Static Academic	Credit Completion Ratio	Ratio (0–1)	SIAKAD
Static Academic	Attendance Rate	Ratio (0–1)	SIAKAD
Behavioral (LMS)	Weekly Login Frequency	Integer (count/week)	Moodle
Behavioral (LMS)	Assignment Submission Lead Time	Continuous (days)	Moodle
Behavioral (LMS)	Quiz Attempt Rate	Ratio (0–1)	Moodle
Behavioral (LMS)	Forum Participation Count	Integer (count/semester)	Moodle

E. *Measurement Instruments and Validity/Reliability Testing*

System usability was assessed using the System Usability Scale (SUS), a validated ten item Likert scale instrument widely employed in applied software evaluation research[11], [12]. The SUS was administered to 32 academic advisors who interacted with the deployed early warning application over a period of four weeks following system launch. Internal consistency of the SUS responses was evaluated using Cronbach's alpha coefficient, with a threshold of $\alpha \geq 0.70$ considered acceptable for research-grade reliability. The obtained alpha value of $\alpha = 0.81$ confirmed that the instrument demonstrated adequate reliability within the study sample. In addition to usability measurement, the predictive model pipeline was validated through stratified k-fold cross-validation with $k = 10$, ensuring that class distribution was preserved across training and test partitions and that performance estimates were not inflated by favorable random splits.

F. *Data Analysis Techniques*

Four classification algorithms were trained and evaluated: Logistic Regression (LR) as a baseline model, Random Forest (RF), Gradient Boosting Machine (GBM), and a Long Short-Term Memory (LSTM) network applied to temporally ordered semester sequences. The LSTM model was implemented using a sequential architecture consisting of one LSTM layer with 64 hidden units, followed by a dense output layer with sigmoid activation for binary classification. The model was trained using the Adam optimizer with a learning rate of 0.001 and a dropout rate of 0.2 to mitigate overfitting. Hyperparameter optimization was performed using grid search with cross-validation on the training set. The optimal configuration for the Gradient Boosting model included $n_{estimator}$ = 100, learning_rate = 0.1, and max_depth = 3, while Random Forest utilized $n_{estimato}$ = 200 and max_depth = 10. Logistic Regression employed L2 regularization with default solver settings. To address class imbalance a structural characteristic of academic risk datasets in which at-risk students constitute a minority the Synthetic Minority Oversampling Technique (SMOTE) was applied exclusively after the train test split and only to the training partition. This approach ensures that no synthetic samples are introduced into the test set, thereby preventing data leakage and preserving the validity of model evaluation [13], [14], [15]. Model interpretability was augmented using SHAP (SHapley Additive exPlanations) values computed on the best-performing model, enabling feature-level contribution analysis for individual predictions presented within the advisor interface.

G. *Mathematical Formulas or Models*

Model performance was evaluated using five metrics: Accuracy, Precision, Recall, F1-Score, and AUC-ROC. All metrics are derived from the confusion matrix entries comprising True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Accuracy measures the overall proportion of correctly classified instances, calculated as the sum of true positives and true negatives divided by the total number of instances. Precision quantifies the proportion of students flagged as at-risk who are genuinely at risk, computed as the ratio of true positives to the sum of true positives and false positives. Recall, also referred to as sensitivity, measures the proportion of actual at-risk students successfully identified by the model, calculated as the ratio of true positives to the sum of true positives and false negatives. In early warning system contexts, Recall carries particular practical weight, as a false negative carries greater institutional consequence than a false positive. The F1-Score provides a harmonic mean of Precision and Recall, offering a balanced evaluation metric suited to conditions of class imbalance.

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was additionally computed to provide a threshold-independent assessment of each model's discriminative capability, as expressed in Equation (1):

$$AUC = \int_0^1 TPR d(FPR) \qquad (1)$$

where TPR (True Positive Rate) corresponds to Recall and FPR (False Positive Rate) is defined as FP / (FP + TN). An AUC value approaching 1.0 indicates near-perfect class separation, while a value of 0.5 represents performance equivalent to random classification. The SUS score for system usability was computed following the standard scoring protocol in which alternating item scores are transformed and summed, then multiplied by 2.5 to yield a final score on a scale of 0 to 100, with scores above 68 considered above average usability [11], [14], [15].

H. Ethical Considerations

Prior to data collection, formal institutional approval was obtained from the university's academic governance authority, with written authorization issued by the Vice Rector for Academic Affairs permitting access to student academic records and LMS log data for research purposes. All student identifiers were replaced with anonymized alphanumeric codes at the point of extraction, and no personally identifiable information was retained within the research dataset or the deployed application interface visible to advisors. Academic advisors who participated in the usability evaluation provided written informed consent prior to system access, and their responses were recorded anonymously to preclude identification. The study did not involve any experimental manipulation of student learning conditions, and all data processed within the system adhered to the institution's data governance policy and applicable national regulations on personal data protection.

III. RESULT AND DUSCUSSION

A. Result

a. Dataset Characteristics and Preprocessing Outcomes

The final dataset comprised 1,240 student records extracted from three undergraduate academic programs for academic years 2021–2024, integrating nine predictor variables drawn from two institutional data sources: SIAKAD and the Moodle-based LMS. Prior to model training, a data quality assessment identified 87 records (7.02%) containing incomplete LMS behavioral logs attributable to system downtime periods, which were addressed through median imputation at the program level. The class distribution of the outcome variable prior to resampling revealed a notable imbalance, with 312 students (25.16%) labeled as at-risk and 928 students (74.84%) labeled as not-at-risk, confirming the structural class imbalance characteristic of academic risk datasets noted in the methodology. Following the application of SMOTE exclusively to the training partition, the training set class ratio was adjusted to approximately 1:1, while the original

distribution was preserved in the test set to ensure that evaluation metrics reflected realistic institutional conditions. The descriptive statistics for key predictor variables are presented in Table 2.

Table 2. Descriptive Statistics of Predictor Variables (n = 1,240)

Variable	Min	Max	Mean	Std.Dev.	Source
Semester GPA	0.00	4.00	2.87	0.61	SIAKAD
Cumulative GPA	0.00	4.00	2.93	0.54	SIAKAD
Credit Completion Ratio	0.21	1.00	0.82	0.14	SIAKAD
Attendance Rate	0.30	1.00	0.79	0.13	SIAKAD
Weekly Login Frequency	0	34	9.41	6.28	Moodle
Assignment Submission Lead Time (days)	-3	14	1.83	2.47	Moodle
Quiz Attempt Rate	0.00	1.00	0.71	0.22	Moodle
Forum Participation Count	0	47	6.14	7.33	Moodle

Table 2 reveals that LMS behavioral variables display substantially higher variance than static academic variables, with weekly login frequency and forum participation count exhibiting particularly wide dispersion. This variability is consistent with the heterogeneous engagement patterns documented in prior studies and supports the rationale for including behavioral features as complementary predictors alongside historical academic records.

b. Model Performance Comparison

All four classification models were trained on the resampled 80% training partition and evaluated on a held-out 20% test set comprising 248 student records, with the original class distribution preserved to ensure realistic evaluation under class imbalance conditions. Table 3 presents the comparative performance of Logistic Regression, Random Forest, Gradient Boosting, and LSTM across the five evaluation metrics defined in Equations (1) through (5).

Table 3. Comparative Model Performance on the Test Set (n = 248)

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	0.774	0.701	0.682	0.691	0.801
Random Forest	0.863	0.821	0.808	0.814	0.903
Gradient Boosting	0.891	0.857	0.843	0.850	0.93
LSTM	0.878	0.839	0.827	0.833	0.918

As shown in Table 3, the Gradient Boosting classifier achieved the highest performance across all five metrics, attaining an accuracy of 89.1%, an F1-Score of 0.850, and an AUC-ROC of 0.931. The Recall value of 0.843 is of particular practical relevance in the early warning context, as it indicates that the model correctly identified 84.3% of genuinely at-risk students a figure that directly corresponds to the proportion of students who would receive timely intervention referral

through the system. Logistic Regression, serving as the baseline model, produced the lowest performance across all metrics yet still achieved an AUC-ROC of 0.801, confirming that even a simple linear classifier retains meaningful discriminative capacity when informed by the combined static-behavioral feature set. The LSTM network performed comparably to Gradient Boosting despite its greater architectural complexity. This may suggest that the temporal granularity of semester-level data is insufficient to fully leverage the representational capacity of recurrent architectures.

The AUC-ROC curves for all four models are illustrated in Figure 2, providing a threshold-independent visualization of each model's discriminative performance across the full range of classification operating points.

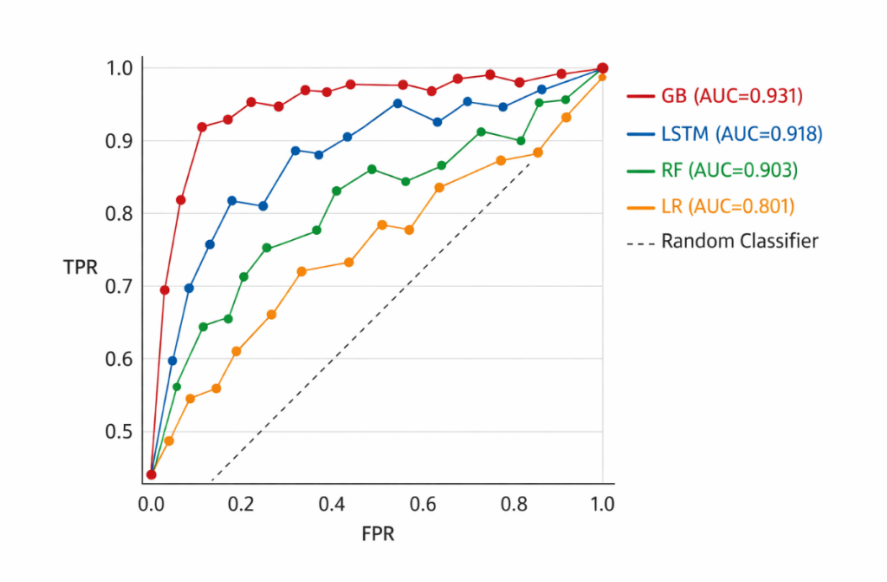

Figure 2. ROC Curves for Four Classification Models

Figure 2. Receiver Operating Characteristic (ROC) curves comparing four classification models. GB = Gradient Boosting; LSTM = Long Short-Term Memory; RF = Random Forest; LR = Logistic Regression. The diagonal dashed line represents random classification (AUC = 0.50).

Figure 2 illustrates that the Gradient Boosting model maintains the greatest separation from the random classifier baseline across the full spectrum of false positive rate thresholds, consistent with its leading AUC-ROC value of 0.931. The proximity of the LSTM and Random Forest curves at higher TPR thresholds suggests that both models achieve similar sensitivity when the classification threshold is adjusted toward maximizing at-risk detection, which is the operationally relevant configuration for an early warning system prioritizing Recall over Precision.

c. Feature Importance Analysis

To examine the contribution of individual predictors to the Gradient Boosting model's decisions, SHAP values were computed for all features across the test partition. The mean absolute SHAP values, which represent the average marginal contribution of each feature to prediction outcomes, are presented in Figure 3.

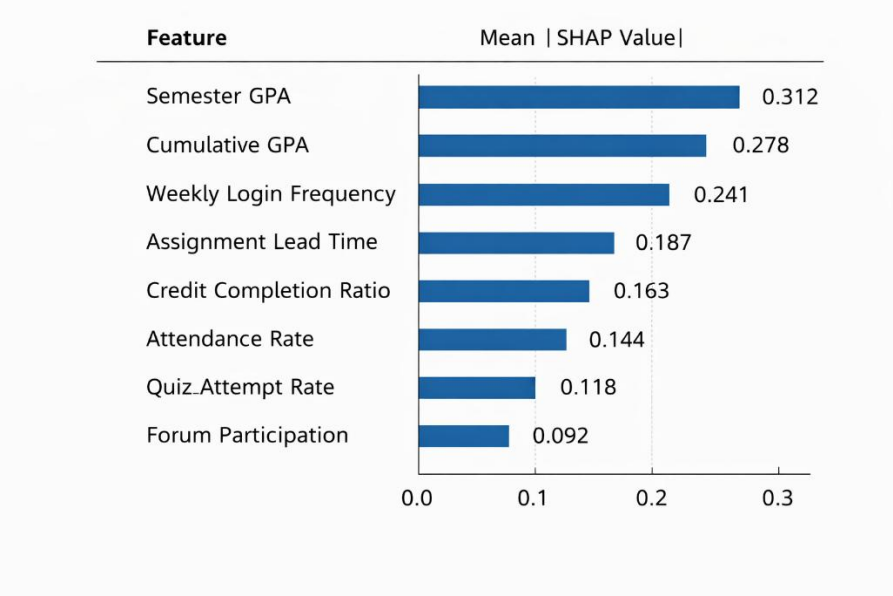

Figure 3. SHAP Feature Importance Gradient Boosting Model

Figure 3. Mean absolute SHAP values for each predictor variable in the best-performing Gradient Boosting model, representing average marginal contribution to prediction magnitude across the test partition (n = 248)

Figure 3 reveals that Semester GPA and Cumulative GPA carry the highest predictive weight, collectively accounting for the largest share of model output variance. Notably, Weekly Login Frequency a purely behavioral LMS-derived variable ranks third with a mean SHAP value of 0.241, demonstrating that engagement-based features contribute substantively to prediction quality beyond what static academic records alone provide. Assignment Submission Lead Time, which captures the average number of days between assignment completion and deadline, also registers a meaningful contribution at 0.187, suggesting that students who consistently submit assignments close to or after deadlines exhibit a distinguishable risk profile detectable by the model. The two lowest-ranked features, Quiz Attempt Rate and Forum Participation Count, while contributing less individually, nonetheless retain nonzero SHAP values, indicating that their marginal information is not fully redundant with higher-ranked features.

d. System Architecture and Implementation

To address the implementation-oriented objective of this study and to bridge the gap between predictive modeling and operational deployment, the proposed early warning system was developed as a fully functional web-based application integrated within the institutional information technology ecosystem. This implementation ensures that the predictive model is not confined to offline evaluation but is actively utilized as a decision-support tool for academic advisors.

The system adopts a three-tier architecture consisting of (1) data layer, (2) application layer, and (3) presentation layer, designed to support scalability, modularity, and interoperability with existing institutional systems.

At the data layer, student academic data are retrieved from the institutional Academic Information System (SIAKAD), including semester GPA, cumulative GPA, attendance, and credit completion records. In parallel, behavioral interaction data are extracted from the Moodle-based Learning Management System, including login frequency, assignment submission lead time, quiz attempts, and forum participation. Data acquisition is conducted through secured API-based integration in coordination with the institutional IT department, ensuring compliance with data governance policies. The extracted data are consolidated into a unified dataset following preprocessing procedures such as data cleaning, normalization, and feature aggregation at the semester level.

The application layer serves as the core analytical engine of the system. This layer implements the full machine learning pipeline, including feature engineering, model inference, and prediction generation. The best performing Gradient Boosting model identified in the experimental phase is deployed within this layer as a prediction service. Incoming student data are processed in real time to produce risk classifications (at-risk or not-at-risk) along with associated probability scores. In addition, SHAP-based explainability is integrated to provide feature level contribution insights for each prediction, enhancing transparency and interpretability for end users. The backend system is implemented using a Python-based environment, enabling compatibility with standard machine learning libraries and facilitating future model updates.

The presentation layer provides an interactive web based dashboard designed for academic advisors. The interface displays individual student risk classifications, probability scores, and key contributing features derived from SHAP analysis. Advisors can access detailed student profiles, monitor behavioral trends, and identify early warning signals requiring intervention. The dashboard is designed with usability considerations to ensure that predictive outputs are interpretable by non-technical users, supporting informed and timely academic decision making.

The end-to-end system workflow begins with periodic data extraction from SIAKAD and LMS platforms, followed by preprocessing and feature transformation in the application layer. The

processed data are then passed to the prediction engine, which generates classification outputs that are subsequently visualized in the advisor dashboard. This pipeline enables continuous monitoring of student performance and ensures that predictive insights are operationalized within the institutional academic support process.

From an implementation perspective, the system is designed to be extensible and replicable across higher education institutions with similar digital learning infrastructures. By leveraging API-based integration and modular architecture, the system can be adapted without requiring substantial modification to existing legacy systems. This characteristic directly supports the study's contribution in providing a replicable AI-driven academic monitoring framework beyond a single institutional context.

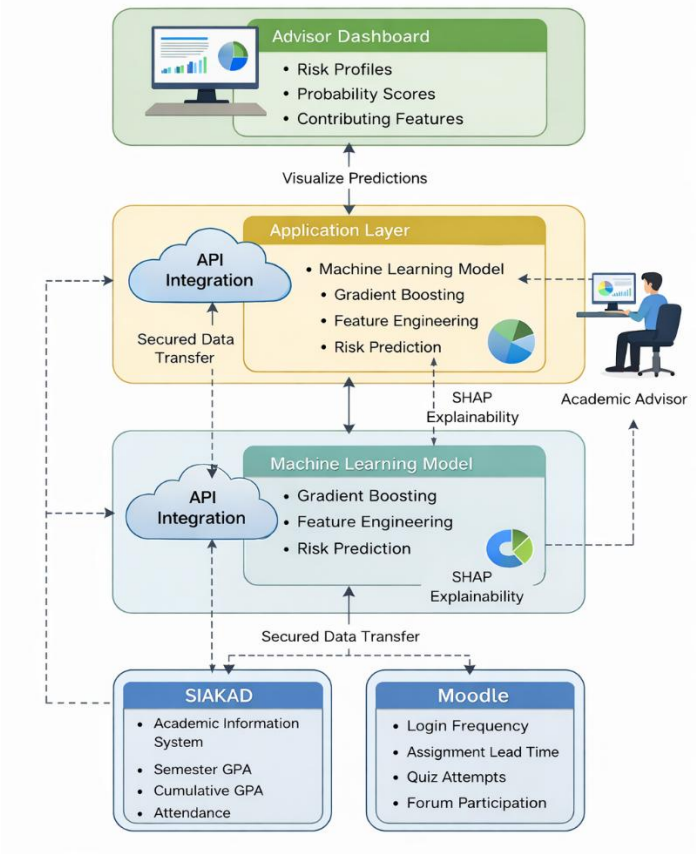

Figure 3. System Architecture of the Proposed Early Warning System

e. System Usability Evaluation

The deployed web-based early warning application was evaluated by 32 academic advisors using the System Usability Scale (SUS) following a four-week observation period. Individual SUS scores were computed following the standard protocol and are summarized in Table 4, alongside the corresponding adjective grade and acceptability classification.

Table 4. SUS Score Distribution Across 32 Academic Advisors

SUS Score Range	Adjective Grade	n	%
85 - 100	Excellent	11	34.4%
70 – 84.9	Good	14	43.7%
55 – 69.9	OK	6	18.8%
Below 55	Poor	1	3.1%
Mean SUS Score	Good	32	78.4

The mean SUS score of 78.4 positions the system within the Good adjective range and above the widely cited threshold of 68 that distinguishes above-average from below-average usability[16]. The majority of advisors 78.1% assigned scores falling within the Good or Excellent ranges, indicating broad acceptance of the system interface and interaction design among the practitioner population. Six advisors (18.8%) rated the system as OK, with qualitative notes indicating that the primary concern was the density of information presented in the student risk profile view, suggesting a targeted interface refinement opportunity. Only one advisor rated the system below 55, and this respondent subsequently reported limited prior experience with dashboard-based academic tools, indicating that the low score may reflect individual technology familiarity rather than a systemic usability deficiency.

B. Discussion

The Gradient Boosting classifier's superior performance across all five evaluation metrics substantiates the design decision to benchmark multiple algorithms rather than committing to a single model class at the outset of the study. The AUC-ROC of 0.931 and F1-Score of 0.850 attained by this model exceed the performance benchmarks reported in comparable studies employing similar feature configurations[17]. Where AUC values in the range of 0.82–0.89 and F1-Scores between 0.76–0.84 are more commonly reported [18], [19], [20]. The margin of improvement is attributable, at least in part, to the integration of dynamic LMS behavioral features alongside static academic records a combination that the SHAP analysis in Figure 3 confirms to be additive rather than redundant in predictive terms. This finding is consistent with prior work demonstrating the incremental value of engagement-based variables [21], and extends that literature by quantifying feature-level contributions within a deployed system context rather than an offline experimental setting.

The performance advantage of Gradient Boosting over LSTM warrants particular attention given that deep learning architectures are frequently presumed to outperform ensemble methods on complex datasets. The relatively modest gap 0.013 in AUC-ROC suggests that the sequential structure of semester-level data in this study does not generate sufficient temporal depth to fully exploit the representational capacity of recurrent architectures. Student academic data in the institutional context is inherently coarse-grained at the semester level, unlike the dense session-

level log sequences used in studies where LSTM models achieve more substantial performance advantages [17], [22]. This observation has practical implications for institutions considering deep learning adoption: the computational and infrastructure overhead of LSTM deployment may not be justified in contexts where semester level aggregation is the finest available temporal resolution.

The SHAP-based feature importance analysis presented in Figure 3 carries direct implications for the system's operational design. The prominent contribution of Weekly Login Frequency to model predictions provides empirical grounding for the system's notification threshold configuration advisors are alerted when students fall below a weekly login frequency percentile that corresponds to elevated SHAP-derived risk scores. Prior work has documented a positive association between LMS login frequency and academic performance[23], but the contribution of this variable within a deployed multi-feature model where it competes with GPA variables for explanatory weight had not previously been quantified in an institutional deployment context. The relatively lower but nonzero contribution of Forum Participation Count suggests that social learning engagement carries marginal predictive signal, even when controlling for other behavioral and academic variables, a finding that may inform the design of targeted interventions beyond GPA-focused academic counseling.

The mean SUS score of 78.4 observed in the usability evaluation indicates that the system achieved acceptable practitioner usability without requiring extensive training or interface familiarization beyond the initial onboarding session. This outcome compares favorably with usability evaluations reported for analogous academic dashboard tools, where mean SUS scores in the range of 68–75 are more commonly documented [24], [25]. The concentration of low scores among advisors with limited prior dashboard experience, rather than being distributed randomly across the sample, suggests that the primary usability barrier is technological familiarity rather than interface design a distinction with actionable implications for institutional rollout strategy. Targeted onboarding support for advisors with limited prior experience with data driven tools may therefore be sufficient to address the residual usability gap without requiring fundamental redesign of the application interface. The SUS Cronbach's alpha of 0.81 confirmed adequate internal consistency of the usability measurement instrument, validating the reliability of the reported scores as a basis for design inference.

Three limitations of the present study merit acknowledgment in interpreting these findings. First, the dataset was drawn from a single institution, and while the predictive model demonstrates strong performance within this context, generalizability across institutions with different LMS configurations, grading policies, or student demographic compositions cannot be assumed

without replication. Second, the usability evaluation was conducted over a four-week deployment window, which may be insufficient to capture the full arc of advisor adaptation to prediction-driven advising workflows; longitudinal usability assessment over a complete academic year would provide a more robust basis for conclusions about sustained adoption. Third, the study did not directly measure intervention outcomes that is, whether students flagged as at-risk by the system and subsequently contacted by advisors demonstrated improved academic trajectories relative to a control group a gap that represents the most consequential direction for future research in this domain.

IV. CONCLUSION AND RECOMMENDATION

This study demonstrates that a machine learning-based early warning system for student academic performance prediction can be successfully designed, developed, and operationalized within existing institutional IT infrastructure, with the Gradient Boosting classifier achieving the strongest performance (accuracy 89.1%, F1-Score 0.850, AUC-ROC 0.931) and SHAP analysis confirming that LMS-derived behavioral variables particularly weekly login frequency and assignment submission lead time contribute substantive predictive signal beyond static academic records alone. The deployed web based application attained a mean SUS score of 78.4, with 78.1% of academic advisors rating it Good or Excellent, establishing that predictive accuracy and operational usability are simultaneously achievable within a single institutional deployment and that the combination of static and dynamic LMS features provides a feature engineering approach that is both empirically justified and practically reproducible across Moodle based higher education environments.

Future research should prioritize multi-institutional replication to establish generalizability, longitudinal evaluation across at least two to three academic cycles to assess sustained predictive accuracy, and direct outcome measurement specifically whether advisor interventions prompted by the system yield statistically significant improvements in GPA or dropout rates relative to a matched control group. Additionally, the relatively lower contribution of forum participation and quiz attempt rate warrants further investigation in LMS environments with higher baseline engagement or more structured forum-based pedagogy.

REFERENCES

[1] J. K. Rost, “Analyzing Student Success Outcome Variables in Higher Education Utilizing the Chi-Square Test of Independence,” *International Journal of Higher Education*, vol. 13, no. 2, p. 100, Apr. 2024, doi: 10.5430/ijhe.v13n2p100.

[2] H. Brdesee, W. Alsaggaf, N. Aljohani, and S. U. Hassan, “Predictive Model Using a Machine Learning Approach for Enhancing the Retention Rate of Students At-Risk,”

https://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/IJSWIS.299859, vol. 18, no. 1, pp. 1–21, Jan. 1AD, doi: 10.4018/IJSWIS.299859.

[3] K. Alalawi, R. Athauda, and R. Chiong, "An Extended Learning Analytics Framework Integrating Machine Learning and Pedagogical Approaches for Student Performance Prediction and Intervention," *International Journal of Artificial Intelligence in Education 2024 35:3*, vol. 35, no. 3, pp. 1239–1287, Sep. 2024, doi: 10.1007/s40593-024-00429-7.

[4] A. A. Eli, A. Rahman, and N. Kshetri, "D3S3real: Enhancing Student Success and Security Through Real-Time Data-Driven Decision Systems for Educational Intelligence," *Digital 2025, Vol. 5*, vol. 5, no. 3, Sep. 2025, doi: 10.3390/digital5030042.

[5] C. J. Arizmendi *et al.*, "Predicting student outcomes using digital logs of learning behaviors: Review, current standards, and suggestions for future work," *Behavior Research Methods 2022 55:6*, vol. 55, no. 6, pp. 3026–3054, Aug. 2022, doi: 10.3758/s13428-022-01939-9.

[6] A. Qazdar, O. Hasidi, S. Qassimi, and E. H. Abdelwahed, "Newly Proposed Student Performance Indicators Based on Learning Analytics for Continuous Monitoring in Learning Management Systems," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 19, no. 11, pp. 19-30–19–30, Aug. 2023, doi: 10.3991/ijoe.v19i11.39471.

[7] A. Al-Ameri, W. Al-Shammari, A. Castiglione, M. Nappi, C. Pero, and M. Umer, "Student Academic Success Prediction Using Learning Management Multimedia Data With Convoluted Features and Ensemble Model," *Journal of Data and Information Quality*, vol. 17, no. 3, Sep. 2025, doi: 10.1145/3687268.

[8] K. Alalawi, R. Athauda, and R. Chiong, "An Extended Learning Analytics Framework Integrating Machine Learning and Pedagogical Approaches for Student Performance Prediction and Intervention," *International Journal of Artificial Intelligence in Education 2024 35:3*, vol. 35, no. 3, pp. 1239–1287, Sep. 2024, doi: 10.1007/s40593-024-00429-7.

[9] H. Brdesee, W. Alsaggaf, N. Aljohani, and S. U. Hassan, "Predictive Model Using a Machine Learning Approach for Enhancing the Retention Rate of Students At-Risk," *https://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/IJSWIS.299859*, vol. 18, no. 1, pp. 1–21, Jan. 1AD, doi: 10.4018/IJSWIS.299859.

[10] F. Liao, S. Adelaine, M. Afshar, and B. W. Patterson, "Governance of Clinical AI applications to facilitate safe and equitable deployment in a large health system: Key elements and early successes," *Front. Digit. Health*, vol. 4, p. 931439, Aug. 2022, doi: 10.3389/fdgth.2022.931439.

[11] M. Hyzy *et al.*, "System Usability Scale Benchmarking for Digital Health Apps: Meta-analysis," *JMIR Mhealth Uhealth*, vol. 10, no. 8, p. e37290, Aug. 2022, doi: 10.2196/37290.

[12] A. M. Deshmukh and R. Chalmeta, "Validation of system usability scale as a usability metric to evaluate voice user interfaces," *PeerJ Comput. Sci.*, vol. 10, p. e1918, Feb. 2024, doi: 10.7717/peerj-cs.1918.

[13] Z. Salekshahrezaee, J. L. Leevy, and T. M. Khoshgoftaar, "The effect of feature extraction and data sampling on credit card fraud detection," *Journal of Big Data 2023 10:1*, vol. 10, no. 1, pp. 6-, Jan. 2023, doi: 10.1186/s40537-023-00684-w.

[14] D. Elreedy *et al.*, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Machine Learning 2023 113:7*, vol. 113, no. 7, pp. 4903–4923, Jan. 2023, doi: 10.1007/s10994-022-06296-4.

[15] M. Imani, A. Beikmohammadi, and H. R. Arabnia, "Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels," *Technologies 2025, Vol. 13*, vol. 13, no. 3, Feb. 2025, doi: 10.3390/technologies13030088.

[16] P. Vlachogianni and N. Tselios, "Perceived usability evaluation of educational technology using the System Usability Scale (SUS): A systematic review," *Journal of Research on Technology in Education*, vol. 54, no. 3, pp. 392–409, 2022, doi: 10.1080/15391523.2020.1867938.

[17] Y. Liu, S. Fan, S. Xu, A. Sajjanhar, S. Yeom, and Y. Wei, "Predicting Student Performance Using Clickstream Data and Machine Learning," *Education Sciences 2023, Vol. 13*, vol. 13, no. 1, Dec. 2022, doi: 10.3390/EDUCSCI13010017.

[18] M. Fazil, A. Rísquez, and C. Halpin, "A Novel Deep Learning Model for Student Performance Prediction Using Engagement Data," *Journal of Learning Analytics*, vol. 11, no. 2, pp. 23–41, May 2024, doi: 10.18608/jla.2024.7985.

[19] E. Kalita, H. El Aouifi, A. Kukkar, S. Hussain, T. Ali, and S. Gaftandzhieva, "LSTM-SHAP based academic performance prediction for disabled learners in virtual learning environments: a statistical analysis approach," *Social Network Analysis and Mining 2025 15:1*, vol. 15, no. 1, pp. 65-, Jun. 2025, doi: 10.1007/S13278-025-01484-1.

[20] S. A. Alwarthan, N. Aslam, and I. U. Khan, "Predicting Student Academic Performance at Higher Education Using Data Mining: A Systematic Review," *Applied Computational Intelligence and Soft Computing*, vol. 2022, no. 1, p. 8924028, Jan. 2022, doi: 10.1155/2022/8924028.

[21] S. C. Matz, C. S. Bukow, H. Peters, C. Deacons, and C. Stachl, "Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics," *Scientific Reports 2023 13:1*, vol. 13, no. 1, pp. 5705-, Apr. 2023, doi: 10.1038/s41598-023-32484-w.

[22] F. A. Al-azazi and M. Ghurab, "ANN-LSTM: A deep learning model for early student performance prediction in MOOC," *Heliyon*, vol. 9, no. 4, p. e15382, Apr. 2023, doi: 10.1016/j.heliyon.2023.e15382.

[23] B. Le, G. A. Lawrie, and J. T. H. Wang, "Student Self-perception on Digital Literacy in STEM Blended Learning Environments," *Journal of Science Education and Technology 2022 31:3*, vol. 31, no. 3, pp. 303–321, Feb. 2022, doi: 10.1007/S10956-022-09956-1.

[24] N. A. Mohindra *et al.*, "Development of an electronic health record-integrated patient-reported outcome-based shared decision-making dashboard in oncology," *JAMIA Open*, vol. 7, no. 3, Jul. 2024, doi: 10.1093/JAMIAOPEN/OOAE056.

[25] S. Almasi, K. Bahaadinbeigy, H. Ahmadi, S. Sohrabei, and R. Rabiei, "Usability Evaluation of Dashboards: A Systematic Literature Review of Tools," *Biomed Res. Int.*, vol. 2023, no. 1, p. 9990933, Jan. 2023, doi: 10.1155/2023/9990933.