

Machine Learning Framework for Student Academic Failure Prediction: A Systematic Literature Review and Research Framework

Babandi Usman¹, Salim Ahmad², Zahraddeen Sufyan³

¹Department of Computer Science, Jigawa State Polytechnic Dutse

²Department of Information Technology, Federal University Dutse

^{1,3}Department of Computer Science, Federal University Dutse

Abstract- The early identification of students who are likely to experience academic failure has become an important application of machine learning (ML) and learning analytics in higher education. Although numerous studies have developed predictive models for student performance, the literature remains fragmented across performance prediction, at-risk identification, dropout prediction, and student success modelling, with substantial variation in outcome definitions, predictor availability, validation strategies, and evaluation metrics. This systematic literature review synthesizes research on ML-based prediction of student academic outcomes, with particular emphasis on failure-oriented and early-warning applications. A PRISMA-aligned evidence-synthesis strategy was used to examine systematic reviews, empirical studies, established educational datasets, and recent developments in explainable and responsible artificial intelligence. The reviewed evidence indicates that classification-based approaches, particularly decision trees, random forests, support vector machines, logistic regression, artificial neural networks, and ensemble methods, dominate the field. Academic history, assessment performance, learning-management-system activity, attendance, engagement, and demographic or contextual variables are frequently used predictors. However, high predictive accuracy does not necessarily indicate practical usefulness because information leakage, class imbalance, temporal instability, inadequate calibration, limited external validation, and weak interpretability remain important methodological concerns. Recent studies further demonstrate the importance of explainable AI, fairness assessment, privacy-preserving data generation, and temporal modelling. Based on these findings, this study proposes a Failure-Oriented Explainable Temporal Machine Learning Framework (FET-MLF), integrating leakage-aware temporal feature engineering, predictive modelling, probability calibration, explainability, fairness auditing, and intervention feedback. The framework shifts evaluation from accuracy-centric modelling toward trustworthy, actionable, and institutionally deployable early-warning systems for preventing academic failure.

Keywords— machine learning; student academic failure; student performance prediction; educational data mining; learning analytics; early warning system; explainable artificial intelligence; fairness; predictive modelling; higher education

I. INTRODUCTION

The increasing digitization of higher education has generated large volumes of student-related data through student information systems, learning management systems (LMS), electronic assessment platforms, attendance systems, online discussion environments, and other educational technologies (Teeroovengadum et al., 2017). These data provide opportunities for institutions to identify patterns associated with academic achievement and to

develop predictive systems capable of identifying students who may require additional academic support. Machine learning has consequently become an important methodological component of educational data mining and learning analytics. The central objective of student prediction research is not merely to forecast examination results. In practical educational environments, predictive analytics can potentially provide instructors and academic administrators with early indications that a student is experiencing academic difficulty, allowing

support to be provided before failure becomes irreversible. Systematic reviews have demonstrated substantial growth in this research area. For example, Alalawi et al. reviewed 162 studies published between January 2010 and October 2022 and found that 129 studies addressed assessment, course, or programme performance and the identification of students at risk of failure, while 33 focused on dropout or retention (Alalawi et al., 2025). Their review also showed that classification was substantially more common than regression and clustering approaches.

Despite this growth, the literature contains an important conceptual problem. Terms such as *student success*, *academic performance*, *at-risk student*, *retention*, *dropout*, and *failure* are frequently treated as related or interchangeable outcomes even though they represent different educational phenomena. A student who receives a low course grade is not necessarily a dropout risk, while a student who eventually withdraws from a programme may not previously have been classified as academically failing (Jamil et al., 2023). Consequently, a framework specifically designed for academic failure prediction requires a clearly defined failure outcome and a prediction point that occurs sufficiently early to permit meaningful intervention. Earlier systematic reviews have identified academic history, e-learning activity, demographic characteristics, and social information as major predictor categories. Abu Saa et al., for example, reviewed 36 studies from a pool of 420 publications covering 2009–2018 and identified previous grades and class performance, e-learning activity, demographic characteristics, and social information as major factor categories. Decision trees, Naïve Bayes, and artificial neural networks were among the most frequently used techniques.

More recent evidence suggests that the field is moving beyond conventional performance prediction toward increasingly sophisticated learning analytics architectures. Chen et al. synthesized empirical research on ML in higher education, covering student performance, engagement, and self-efficacy, while Yuan et al.'s 2026 state-of-the-art review focuses specifically on

models, predictive factors, and evaluation metrics for university student success (Apolinar, 2024). Recent work has also incorporated explainable artificial intelligence (XAI), fairness assessment, privacy-preserving synthetic educational data, and temporal deep learning. The methodological challenge is therefore no longer simply whether machine learning can predict student outcomes. The more consequential question is whether the resulting prediction is early enough, valid enough, explainable enough, fair enough, and sufficiently reliable to support an educational intervention.

This study addresses that problem through a systematic literature review focused on machine-learning-based student academic failure and at-risk prediction. It synthesizes the principal data sources, predictor categories, machine-learning algorithms, evaluation practices, methodological weaknesses, and emerging responsible-AI requirements. Based on the synthesis, a Failure-Oriented Explainable Temporal Machine Learning Framework (FET-MLF) is proposed.

II. LITERATURE REVIEWED

Student Academic Failure and Performance Prediction

Student academic performance prediction is an established research area within educational data mining. Existing studies have approached prediction as regression, classification, clustering, ranking, and sequential modelling. In classification settings, students may be divided into categories such as pass/fail, at-risk/not-at-risk, or multiple achievement levels (Xu, 2024). Regression approaches instead predict continuous outcomes such as marks, GPA, or final scores. The distinction between prediction of performance and prediction of failure is particularly important. Performance prediction can seek to estimate a student's final grade, whereas failure prediction normally requires a predefined threshold representing an academically undesirable outcome. The latter creates a more direct connection between prediction and early-warning intervention (Khan & Alvi, 2024). This concentration on classification is understandable because institutional decision-making often requires actionable categories.

Nevertheless, binary classification can simplify a continuum of academic risk.

Predictor Variables and Data Sources

The predictive variables used in the literature can broadly be classified into academic, behavioural, demographic, socioeconomic, engagement, and institutional categories.

Table 1: Predictor Variables

Predictor category	Examples	Potential predictive value
Academic history	Previous GPA, previous grades, admission scores	Indicates prior academic trajectory
Assessment behavior	Quiz, assignment, test and midterm results	Captures emerging course performance
LMS activity	Logins, clicks, resource access, time-on-task	Represents learning engagement
Attendance	Lecture attendance, laboratory attendance	Captures participation
Interaction	Discussion posts, forum activity, peer interaction	Represents online engagement
Demographic/contextual	Age, gender, socioeconomic indicators	May capture contextual differences
Programme information	Department, course load, year level	Captures programme-specific structure
Administrative	Registration status, academic probation, course history	Represents institutional trajectory

The systematic review by Abu Saa et al. grouped predictors into previous grades/class performance, e-learning activity, demographic characteristics, and social information (Putri & Rudyanto, 2026). Alwarthan similarly reported that university-level academic characteristics, including assessment results, attendance and CGPA, were frequently used, with academic variables often providing strong predictive information. The availability of LMS data has particularly expanded the temporal dimension of prediction. The Open University Learning Analytics

Dataset (OULAD), for example, combines demographic information with aggregated clickstream activity from students interacting with a virtual learning environment. Such datasets allow researchers to study how student behavior changes throughout a course rather than relying exclusively on static demographic information.

Riestra-González et al. demonstrated this potential through large-scale LMS-log analysis for early student-performance prediction. Their study investigated student interactions at different stages of course progression and demonstrated that LMS behavioral patterns could be used to distinguish students at risk from higher-performing students (Chen, 2026).

Machine-Learning Algorithms

The literature contains a broad range of algorithms. Traditional methods remain highly relevant because educational datasets often have limited sample sizes, heterogeneous variables, and structured tabular characteristics. Logistic regression provides a transparent baseline for binary failure prediction. Decision trees offer interpretable rules, while random forests improve robustness by aggregating multiple trees. Support vector machines are effective in high-dimensional classification problems, and k-nearest neighbours provides a relatively simple instance-based approach. Ensemble learning has received considerable attention because combining complementary learners can improve predictive robustness. Injadat et al. investigated ensemble model selection using educational datasets at different stages of course delivery and developed a systematic procedure for selecting ensemble learners based on statistical and feature-selection criteria.

More complex models have also been introduced. Neural networks and deep-learning architectures can represent nonlinear relationships and sequential patterns, while recurrent neural networks, LSTMs, transformers, and graph neural networks can potentially model temporal or relational educational data. However, greater model complexity does not automatically imply greater practical value. If a complex model produces marginally higher

discrimination but substantially reduces transparency or portability, its institutional utility may be limited.

Temporal Prediction and Early Warning

One of the most important dimensions of student-failure prediction is when the prediction is generated.

A model predicting failure after the final examination has limited intervention value because the outcome is already known. Conversely, a prediction generated very early in the semester may provide substantial intervention time but have access to relatively little behavioural evidence (Atuahene, 2024). Research has demonstrated the feasibility of progressive prediction. Xu et al. developed a machine-learning approach for tracking and predicting performance in degree programmes while explicitly addressing the evolving nature of students' academic records. Similarly, Hung et al. proposed a multistage approach for at-risk identification and evaluated prediction at different stages of learning, demonstrating the value of considering the timing and relative nature of engagement indicators. Consequently, a failure-prediction framework should treat prediction time as a core experimental variable rather than an incidental methodological choice.

Explainable Artificial Intelligence

Explainability has become increasingly relevant as student prediction systems move toward practical deployment. A model that identifies a student as high-risk without explaining the factors contributing to that prediction may be difficult for instructors and academic advisers to use responsibly (Mathew et al., 2025). SHAP provides a game-theoretic framework for estimating feature contributions to individual predictions and aggregate model behaviour. Lundberg and Lee introduced SHAP as a unified framework for interpreting model predictions. In educational applications, SHAP can identify whether variables such as recent assessment performance, attendance, or LMS engagement contribute positively or negatively to predicted risk (Hassija et al., 2024). Recent learning-analytics research demonstrates an additional role for explainability: model stability assessment. Tiukhova et al. found

that SHAP can be used not only to explain predictions but also to examine whether the importance of learning indicators remains stable as educational contexts change. Their findings indicate that learning-context changes can alter the importance of predictors, reinforcing the need for monitoring and model updating.

Thus, XAI should not be regarded merely as a visualization layer added after model development. It should form part of the model-audit and decision-support process.

Fairness, Privacy and Responsible Prediction

Student prediction systems operate on sensitive educational information and can potentially reproduce historical inequalities. If socioeconomic characteristics, demographic variables, or institutional differences are associated with historical academic outcomes, an ML model may learn these associations and produce systematically different error rates across groups. Recent research explicitly integrates fairness analysis into student academic-performance prediction (Akrami et al., 2023). For example, a 2025 study evaluated logistic regression, random forest, and XGBoost models while examining demographic and socioeconomic attributes using demographic-parity and equalized-odds measures. Privacy is equally important. Student records may contain personally identifiable information, academic histories, behavioural traces, and sensitive contextual variables. Recent research has consequently explored privacy-preserving synthetic educational data and federated learning (Hetigan et al., 2025). Agal et al. proposed a privacy-preserving synthetic-data framework and reported that synthetic data could be particularly useful as a pre-training resource before fine-tuning with a smaller quantity of real educational data. Other recent research has investigated privacy-preserving federated learning for interpretable student at-risk prediction across institutions. These developments demonstrate that future student-failure prediction systems must address not only predictive accuracy but also data governance, privacy, fairness, transparency, and institutional accountability.

Synthesis of the Literature

The literature demonstrates significant progress but also reveals a persistent methodological fragmentation.

Dimension	Dominant practice	Identified limitation
Outcome	Pass/fail, grades, performance levels	Failure definitions vary substantially
Data	Academic records and LMS activity	Limited cross-institutional datasets
Models	DT, RF, SVM, LR, ANN, ensembles	Algorithm comparison often lacks temporal control
Validation	Random train/test or cross-validation	Temporal and external validation are less consistently reported
Metrics	Accuracy, F1, AUC	Calibration and intervention utility are often underemphasized
Timing	Course-stage prediction	Prediction horizons are inconsistently defined
Explainability	Increasing use of SHAP/LIME	Often added after modelling rather than integrated
Fairness	Emerging research area	Limited subgroup reporting
Privacy	Increasing concern	Institutional data sharing remains difficult
Deployment	Early-warning prototypes	Intervention effectiveness is rarely evaluated

The most important gap is therefore not the absence of machine-learning algorithms. There is already substantial evidence that numerous algorithms can predict academic outcomes. The gap lies in integrating temporal validity, predictive discrimination, calibration, explainability, fairness, privacy, generalization, and intervention evaluation into a single failure-oriented framework.

III. METHODOLOGY

Review Design

This study adopts a systematic literature-review design informed by the PRISMA 2020 reporting framework. PRISMA 2020 provides updated guidance for transparent reporting of systematic reviews, including specification of eligibility criteria, information sources, selection procedures, synthesis, and limitations (Page et al., 2021). The review was designed to synthesize evidence concerning machine-learning-based prediction of student academic performance, failure, and at-risk status, with particular emphasis on higher education. The synthesis incorporated peer-reviewed empirical studies, systematic reviews, methodological studies and established educational datasets that are directly relevant to predictive modelling. The review also incorporated recent 2025–2026 literature because explainability, fairness, privacy-preserving learning, and temporal modelling have developed rapidly and materially affect the design of contemporary student early-warning systems.

Search Strategy

The evidence searches used combinations of terms relating to four conceptual components: the population, outcome, computational approach, and educational setting. Search concepts included combinations of:

"Student academic performance prediction", "student failure prediction", "at-risk student prediction", "student success prediction", "early warning system", "machine learning", "educational data mining", "learning analytics", "explainable AI", "fairness", and "higher education".

The evidence base was expanded through backward and forward identification from highly relevant systematic reviews and seminal empirical studies. This approach is particularly useful in a field where terminology is inconsistent and studies may describe failure prediction using terms such as *performance prediction*, *at-risk identification*, or *student success prediction*.

Importantly, this manuscript does not fabricate numerical database counts. The evidence search was conducted through scholarly web and publisher/index sources rather than through a reproducible institutional export from Scopus, Web of Science, IEEE Xplore, or ACM Digital Library. Consequently, the review reports the characteristics and findings of the identified evidence rather than presenting invented PRISMA record counts.

Eligibility Criteria

The literature was considered relevant when it addressed machine-learning or data-mining approaches for predicting student academic performance, academic risk, failure, or success, particularly in higher education.

Table 2: Eligibility Criteria

Criterion	Inclusion	Exclusion
Population	University/tertiary students	Studies unrelated to student learning
Outcome	Performance, failure, at-risk status, success, retention where relevant	Purely administrative prediction
Method	ML, data mining, learning analytics, XAI	Non-computational studies without predictive modelling
Data	Academic, LMS, assessment, behavioural or contextual student data	Unrelated datasets
Publication	Peer-reviewed research and authoritative methodological sources	Unverified informal material
Relevance	Direct contribution to prediction methodology	Studies with no meaningful relevance to predictive modelling

Dropout and retention studies were not treated as equivalent to academic failure. They were included only when they provided methodological insights

relevant to early identification of academically at-risk students.

Data Extraction

For each selected study, the following dimensions were considered: publication year, educational setting, dataset characteristics, target variable, prediction stage, predictor categories, algorithm families, validation strategy, evaluation metrics, explainability methods, fairness considerations, privacy mechanisms, and practical intervention implications.

The extraction framework was deliberately broader than conventional algorithm-comparison reviews because a high-performing classifier cannot be considered institutionally useful if its target is poorly defined or its predictors contain future information.

Quality-Appraisal Framework

A structured methodological appraisal was developed around eight dimensions. Each dimension can be scored from 0 to 2, giving a maximum score of 16.

Table 3 The Quality-Appraisal Framework

Criterion	0	1	2
Outcome definition	Unclear	Partially defined	Explicit and reproducible
Prediction timing	Not reported	Approximate	Explicit prediction point
Leakage control	Absent	Partially described	Clearly controlled
Preprocessing	Poorly reported	Partial	Fully described
Class imbalance	Ignored	Mentioned	Properly addressed
Validation	Weak	Standard CV/random split	Temporal/external or rigorous validation
Evaluation	Accuracy-dominated	Multiple metrics	Discrimination + calibration +

			imbalance-aware metrics
Trustworthiness	No XAI/fairness/generalization	Partial	Integrated assessment

This framework does not replace established risk-of-bias tools but provides a domain-specific method for evaluating methodological maturity in student-failure prediction.

IV. RESULTS AND DISCUSSION

Overall Evidence Pattern

The reviewed literature confirms that machine learning has substantial predictive potential in educational environments. The strongest evidence comes from the convergence of several systematic reviews and empirical studies rather than from any single algorithm.

Alalawi et al.'s review of 162 studies provides evidence of the scale of the field and demonstrates that performance and at-risk prediction substantially outnumbered dropout/retention studies. Abu Saa et al. similarly demonstrated the diversity of predictors and algorithms in higher-education performance prediction, while Alwarthan identified classification as the predominant modelling formulation in its review of 45 studies. The evidence therefore supports the feasibility of ML-based academic-risk prediction. However, the literature does not support the proposition that one universal algorithm is optimal across all educational contexts. Differences in institution, discipline, data availability, outcome definition, prediction horizon, and student population strongly influence model performance.

Evidence Concerning Predictive Variables

Academic history consistently emerges as an important predictor category. Previous grades, GPA, assessment marks, and admission-related information can provide strong signals about subsequent performance. Behavioral variables from LMS platforms add a dynamic dimension because they capture changes in student engagement during the course. The OULAD dataset is an important

example of this transition toward behavioural learning analytics. It combines student demographic information with VLE interaction data, enabling analysis of student behaviour over time. Riestra-González et al. demonstrated the value of large-scale LMS logs for early prediction and showed that predictive models could identify different performance patterns at multiple stages of course progression.

However, the use of assessment marks requires particular caution. A model predicting final failure from assessments conducted immediately before the final outcome may achieve excellent predictive performance while offering little actionable lead time. Consequently, predictive feature importance should always be interpreted together with the time at which the feature becomes available.

Evidence Concerning Algorithm Selection

The literature supports the continued relevance of traditional machine-learning algorithms. Decision trees, random forests, support vector machines, logistic regression, neural networks, k-nearest neighbours, and ensemble methods are repeatedly represented across studies. Ensemble methods are particularly attractive because student-performance datasets may contain nonlinear relationships and interactions that are not easily represented by a single classifier. Injadat et al. demonstrated the usefulness of systematic ensemble selection at different stages of course delivery. Nevertheless, algorithmic complexity should be proportional to the research problem. If a dataset contains a relatively small number of structured variables, an interpretable tree-based model or calibrated gradient-boosting model may provide a more useful institutional solution than a deep neural network.

Prediction Timing as a Core Evaluation Dimension

A major result of the synthesis is that prediction timing should be treated as an evaluation axis rather than a secondary methodological detail.

A model with an F1-score of 0.90 at week 12 and a model with an F1-score of 0.82 at week 4 do not necessarily have the same practical value. The latter

may provide substantially more time for intervention.

This motivates an early-warning evaluation matrix:

Table 4: Early-warning evaluation matrix

Prediction stage	Information availability	Intervention window	Expected uncertainty
Very early	Low	Very high	High
Early	Moderate	High	Moderate
Mid-course	High	Moderate	Lower
Late-course	Very high	Low	Low

The ideal model should therefore be evaluated at several prediction checkpoints rather than only once at the end of the academic period.

Research on progressive performance prediction and multistage at-risk identification supports this approach.

Accuracy Is Insufficient

A central finding from the synthesis is that reporting accuracy alone can produce an overly optimistic assessment of student-failure prediction systems.

Suppose a dataset contains 1,000 students and only 100 ultimately fail. A classifier that labels all 1,000 students as non-failing achieves 90% accuracy while identifying none of the student's requiring intervention. Consequently, the recommended minimum evaluation should include accuracy alongside precision, recall, F1-score, specificity, ROC-AUC and PR-AUC. For imbalanced failure outcomes, PR-AUC and class-specific recall can be particularly informative. Calibration should also be considered. A model that assigns a failure probability of 0.80 should ideally correspond to approximately 80% observed failure among comparable cases. Calibration allows institutions to interpret predictions as meaningful risk probabilities rather than arbitrary model scores.

Explainability and Actionability

Prediction without interpretation creates a significant gap between machine learning and educational decision-making. Consider two students

who are both assigned a high-risk prediction. For Student A, the dominant contributors may be declining assessment performance and repeated absence. For Student B, the dominant factors may be declining LMS engagement and missed assignments. The appropriate intervention may differ substantially.

This is where SHAP, LIME, counterfactual explanations, and related XAI techniques become useful. Recent research has shown that explainability can also reveal changes in feature importance over time, providing a mechanism for detecting model instability and contextual drift.

The objective should therefore not be to explain the model merely for technical transparency. Explanation should support an actionable academic response.

Fairness and Bias

Fairness represents a particularly important extension of conventional predictive modelling. If a model produces systematically different false-negative or false-positive rates across demographic or socioeconomic groups, an institution may unintentionally distribute academic interventions inequitably. The 2025 fairness-oriented study using the UCI Student Performance dataset illustrates how demographic-parity and equalized-odds measures can be incorporated into student prediction evaluation.

However, fairness cannot be reduced to removing sensitive attributes. Proxy variables can reproduce similar information, and excluding sensitive variables may prevent researchers from detecting subgroup disparities. Therefore, sensitive attributes may need to be retained under appropriate governance for auditing purposes, even if they are excluded from prediction features.

Privacy-Preserving Learning

Educational prediction systems face a structural data-access problem. Institutions possess rich student information but are often unable to share it because of privacy, governance, and ethical constraints. Recent research suggests two emerging directions. The first is privacy-preserving synthetic

educational data. Agal et al. reported that synthetic educational data can provide useful pre-training information but that direct transfer to real educational data may be weaker than fine-tuning with a relatively small amount of real data.

The second direction is federated learning, where models can be trained across institutional environments without centralizing raw student records. Recent work has specifically investigated privacy-preserving federated learning for interpretable student at-risk prediction. These approaches could become increasingly important for cross-university validation and generalisation.

Proposed Failure-Oriented Explainable Temporal Machine Learning Framework

Based on the literature synthesis, this study proposes the Failure-Oriented Explainable Temporal Machine Learning Framework (FET-MLF).

The framework is designed around seven interconnected principles: clearly defined failure outcomes, temporal feature validity, leakage-aware preprocessing, predictive modelling, calibrated risk estimation, explainable and fair decision support, and intervention feedback.

The framework begins with heterogeneous student data. Rather than immediately feeding all available variables into an ML algorithm, the proposed approach establishes the prediction point first. Every predictor must then pass a temporal-validity test. The second component addresses class imbalance. Any resampling technique such as SMOTE should be applied only within training folds after the data have been partitioned. Applying oversampling before train/test splitting can create information leakage because synthetic observations may contain information derived from observations subsequently appearing in the test set. The third component is model development. Multiple algorithms should be evaluated rather than assuming that one method is universally optimal. At minimum, a transparent baseline such as logistic regression should be compared with tree-based and ensemble models. More complex temporal architectures should be

justified by the structure and scale of the available data. The fifth component combines explainability and fairness. SHAP or related methods can provide global and local explanations, while subgroup analysis can determine whether the model behaves differently for different student populations.

The sixth component is human-led intervention. The model should not autonomously make high-stakes academic decisions. Instead, its role is to provide evidence that can help qualified academic personnel decide whether further assessment, mentoring, tutoring, attendance support, or other assistance may be appropriate.

The final component is outcome monitoring. After intervention, the institution should determine whether identified students actually improved and whether the intervention produced unintended consequences. Model drift should also be monitored because relationships between learning behaviour and academic outcomes can change over time. The importance of this monitoring is supported by research showing that changing learning contexts can alter the stability of predictive indicators.

Proposed Evaluation Protocol

The FET-MLF framework recommends evaluating a student-failure model across multiple dimensions rather than reporting a single performance metric.

Evaluation dimension	Recommended measure	Purpose
Discrimination	ROC-AUC	Overall ranking ability
Imbalanced classification	PR-AUC	Failure-class performance
Positive identification	Recall	Detecting failing students
Precision	Precision	Controlling unnecessary interventions
Balance	F1-score	Joint precision-recall assessment
Reliability	Brier score / calibration curve	Probability reliability

Stability	Temporal validation	Performance over time
Generalisation	External validation	Cross-context performance
Explainability	SHAP/LIME/counterfactual analysis	Interpretability
Fairness	Equal opportunity/equalized odds/subgroup calibration	Equity
Operational utility	Intervention uptake and outcome change	Real-world effectiveness

The framework therefore proposes a distinction between model performance and system effectiveness.

Model performance concerns whether the algorithm predicts accurately. System effectiveness concerns whether those predictions produce useful educational outcomes when incorporated into an intervention process. This distinction is critical. A highly accurate model that produces predictions too late for intervention may be less useful than a slightly less accurate model that identifies risk sufficiently early.

Research Gaps Identified

The synthesis reveals several major gaps.

First, there remains considerable variation in the definition of academic failure. Some studies use final grades, some use performance categories, and others use at-risk classifications. A standardized operational definition is therefore necessary for reproducibility. Second, temporal validation remains insufficiently central to evaluation. Many studies demonstrate strong predictive performance without adequately demonstrating that the model could have generated the prediction early enough to influence an intervention. Third, external validation remains a major challenge. A model trained at Institution A may exploit institution-specific relationships that do not transfer to Institution B. This is particularly important because curriculum structure, assessment practices, LMS configuration, student demographics, and institutional policies differ substantially. Fourth, calibration is less

frequently emphasized than discrimination. A ranking model can distinguish high-risk from low-risk students without producing reliable probabilities. For intervention prioritization, however, calibrated probabilities are valuable.

Fifth, explainability and fairness remain unevenly integrated. Although recent studies increasingly use SHAP, LIME, fairness metrics, and responsible-AI methods, these components are often added after predictive modelling rather than incorporated into the original system design.

Sixth, relatively little research directly evaluates whether predictive alerts actually improve student outcomes. This represents a crucial distinction between prediction research and intervention research.

Implications for Future Research

Future studies should move toward longitudinal, multi-institutional and intervention-aware research designs. Rather than training a model on one semester and reporting a static test score, researchers should evaluate whether the model maintains calibration and discrimination across subsequent cohorts. A particularly promising research direction is the combination of temporal deep learning with explainable ensemble models. Sequential models can capture evolving behaviour, while explainability mechanisms can translate predictions into human-understandable evidence.

Privacy-preserving learning also provides an opportunity for multi-institutional research without requiring institutions to exchange raw student records. Synthetic-data pre-training and federated learning are emerging approaches that could help address the scarcity of shareable educational datasets.

Another important direction is causal evaluation of interventions. If a model identifies 200 high-risk students, the research question should not end with how accurately those 200 students were identified. The subsequent question should be whether targeted interventions changed their academic

outcomes compared with an appropriate comparison group.

V. CONCLUSION

Machine learning has established substantial potential for predicting student academic performance and identifying students who may be at risk of academic failure. The systematic evidence reviewed in this study demonstrates the widespread use of decision trees, random forests, support vector machines, logistic regression, neural networks, and ensemble approaches, supported by academic, behavioural, LMS, assessment, demographic, and contextual data. Large-scale learning analytics datasets have further enabled researchers to investigate student behaviour dynamically rather than relying exclusively on static academic records. However, the synthesis demonstrates that predictive accuracy alone is insufficient for a trustworthy academic-failure prediction system. The practical value of a prediction depends on whether it is generated early enough for intervention, whether the predictors were legitimately available at prediction time, whether the model performs reliably under class imbalance, whether predicted probabilities are calibrated, whether the model generalizes across cohorts and institutions, and whether educators can understand and appropriately act upon its predictions. The literature also indicates a clear transition toward responsible and explainable educational artificial intelligence. Recent studies increasingly address explainability, fairness, privacy-preserving learning, synthetic educational data, and federated learning, suggesting that future systems must be evaluated as socio-technical systems rather than isolated classifiers. To address these requirements, this study proposed the Failure-Oriented Explainable Temporal Machine Learning Framework (FET-MLF). The framework integrates temporal feature engineering, leakage prevention, class-imbalance management, machine-learning modelling, calibrated risk estimation, explainability, fairness auditing, human-led intervention, and continuous outcome monitoring. The principal implication is that the next generation of student-failure prediction research should move beyond the question of "Which algorithm achieves

the highest accuracy?" toward the broader question of "Can the system reliably identify academic risk early, explain why the risk is present, operate fairly and privately, generalize across educational contexts, and support interventions that measurably improve student outcomes?"

A genuinely useful student-failure prediction system is therefore not simply a high-performing classifier. It is a temporally valid, calibrated, explainable, fair, privacy-conscious, generalizable, and intervention-oriented decision-support system. This perspective provides a stronger foundation for future empirical implementation and offers a research direction capable of connecting machine learning methodology with the practical objective of reducing preventable academic failure in higher education.

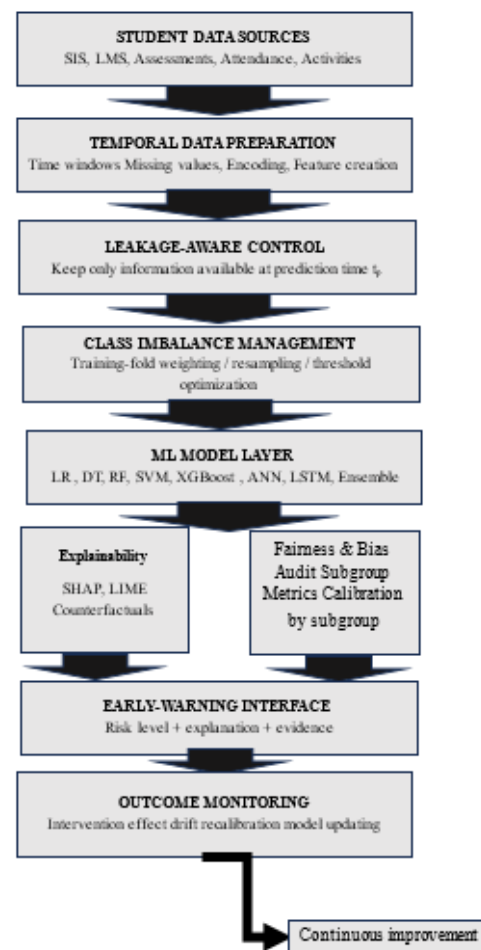


Figure 1. Conceptual architecture of the proposed FET-MLF framework

REFERENCES

1. Akrami, N. El, Hanine, M., Flores, E. S., Aray, D. G., & Ashraf, I. (2023). Unleashing the Potential of Blockchain and Machine Learning: Insights and Emerging Trends From Bibliometric Analysis. *IEEE Access*, 11(July), 78879–78903. <https://doi.org/10.1109/ACCESS.2023.3298371>
2. Alalawi, K., Athauda, R., Chiong, R., & Renner, I. (2025). *Evaluating the student performance prediction and action framework through a learning analytics intervention study*. 2887–2916.
3. Apolarin, M. (2024). *Using Decision Tree Algorithm to Predict Student Performance*. February 2019. <https://doi.org/10.17485/ijst/2019/v12i5/140987>
4. Atuahene, S. (2024). *Comparative Study of Traditional vs . Digital Learning Materials in Ghanaian Schools*. 1–18.
5. Chen, K. (2026). *Auditing GenAI – Student Grade Claims on Public Datasets: Nested Controls , Frozen Thresholds , and Claim Labels*. 306, 1–17.
6. Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., & Huang, K. (2024). Interpreting Black - Box Models: A Review on Explainable Artificial Intelligence. In *Cognitive Computation*. Springer US. <https://doi.org/10.1007/s12559-023-10179-8>
7. Hetigan, J., Abrenica, M. C., & Lolong, L. (2025). *Leveraging Predictive Analytics on LMS Logs to Examine the Impact of Engagement on Academic Performance among College Students Enrolled in Centro Escolar University*. XII(2321), 563–573. <https://doi.org/10.51244/IJRSI>
8. Jamil, L. A., Shaqsi, A., & Tarhini, A. (2023). *Using machine learning to predict factors affecting academic performance: the case of college students on academic probation*. 12407–12432.
9. Khan, S., & Alvi, S. T. (2024). *Impact of drug abuse on student 's academic performance at University Level The Impact of drug abuse on students ' academic performance at university Level*. November. <https://doi.org/10.33282/rr.vx9il.162>
10. Mathew, D. E., Ebem, D. U., Ikegwu, A. C., Ukeoma, P. E., & Dibiaezue, N. F. (2025). Recent Emerging Techniques in Explainable Artificial Intelligence to Enhance the Interpretable and Understanding of AI Models for Human. *Neural Processing Letters*, 57(1), 1–32. <https://doi.org/10.1007/s11063-025-11732-2>
11. Page, M. J., Mckenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-wilson, E., Mcdonald, S., ... Yhec, C. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Journal of Clinical Epidemiology*, xxxx, 1–12. <https://doi.org/10.1016/j.jclinepi.2021.03.001>
12. Putri, N. S., & Rudyanto, R. (2026). *Machine learning approaches for predicting mathematics achievement in k – 12 education*. 11, 13–33. <https://doi.org/10.23916/086824011>
13. Teeroovengadum, V., Heeraman, N., & Jugurnath, B. (2017). Examining the antecedents of ICT adoption in education using an Extended Technology Acceptance Model (TAM) Viraiyan Teeroovengadum , Nabeel Heeraman and Bhavish Jugurnath University of Mauritius. *International Journal of Education and Development Using ICT*, 13(3), 4–23.
14. Xu, Z. (2024). AI in education: Enhancing learning experiences and student outcomes. *Applied and Computational Engineering*, 51(1), 104–111. <https://doi.org/10.54254/2755-2721/51/20241187>