International Journal of Innovative Technology and Interdisciplinary Sciences
https://journals.tultech.eu/index.php/ijitis
ISSN: 2613-7305
Volume 9, Issue 1
DOI: https://doi.org/10.15157/IJITIS.2026.9.1.490-522
Received: 31.12.2025; Revised: 05.03.2026; Accepted: 19.03.2026

Review Article

Iterative and Statistical Analytical Review of Predictive Modeling Approaches in Educational Systems: A Comprehensive Benchmark of AI-Driven Methods

Vipparti Vijaya Kumar Raju[1], Yathirajula Venkata Kanaka Durga Bhavani[2], Purandhar Nandikonda[3], FNU Kareemunnisa[4], Kadaru Bala Brahmeswara[5], Surapaneni Sindhura[6]*

[1]Department of Electronics and Communication Engineering, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh, India
[2]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India
[3]Department of Computer Science and Engineering (Artificial Intelligence), School of Computers, Madanapalle Institute of Technology & Science (MITS), Deemed to be University, Madanapalle Andhra Pradesh, India
[4]Sr Software Engineer, Infohuv llc, Connecticut, USA
[5]Department of Computer Science and Engineering, Seshadri Rao Gudlavalleru Engineering College, Andhra Pradesh, India
[6]Department of Computer Science and Engineering, NRI Institute of Technology, Agiripalli, Andhra Pradesh, India
***ssindhurapraveen@gmail.com**

Abstract

The rapid digitization of education and the growing availability of multimodal student data have driven the development of advanced AI-based educational prediction systems. The proliferation of algorithms, feature spaces, and performance metrics has resulted in fragmented insights and a lack of consensus on optimal modeling practices. This study presents a systematic analytical review of landmark research on predictive modeling in education, covering student performance, dropout risk, employability, mental health, and engagement. A robust benchmarking framework is established to evaluate models across six key performance metrics: root mean square error (RMSE), accuracy, latency, computational complexity, precision, and recall. The review examines hybrid deep learning architectures (e.g., CSSA-Deep RNN, KANFormer, CNN-GRU), metaheuristic-based approaches (PSO, ACO, t-SIDSBO), interpretable models (SHAP, LIME, federated learning), and affective computing systems incorporating facial emotion recognition (FER) and natural language processing (NLP). To enable consistent comparisons, numerical performance scores were standardized and systematically reported. The findings indicate that while accuracy remains a primary concern, practical deployment necessitates careful trade-offs among complexity, latency, and interpretability. This study addresses critical gaps in prior reviews by emphasizing model

explainability, robustness, and context-awareness. Future research directions are identified, including multimodal fusion, edge deployment, longitudinal modeling, and causal explainability. This work serves as a resource for researchers, policymakers, and educational technology developers in designing, evaluating, and deploying intelligent educational systems.

Keywords: Predictive Modeling; Educational Data Mining; Deep Learning; Student Performance, Machine Learning; Scenarios.

INTRODUCTION

Digital platforms, personalized learning, and data-rich ecosystems are changing education. E-learning platforms, learning management systems (LMSs), and massive open online courses (MOOCs) generate large daily student interactions, assessments, physiological signals, and behavioural records [1-3]. Educational institutions can use these data streams to build predictive systems to detect at-risk students, tailor learning routes, improve academic interventions, and predict employability or health risks. Modeling educational data is tough due to its variety and volume. Educational datasets and samples have nonlinear correlations, high-dimensional properties, and missing values that challenge traditional statistical methods. This favour uses machine learning (ML) and deep learning (DL) models that uncover complex latent patterns in student data samples. Grade score prediction, dropout detection, student feedback analysis, and behavioral engagement estimation improved with their practical implementation [4-6]. Nevertheless, there are still a number of systemic difficulties in the process. The educational sector is not homogenous in its nature, students differ based on many socio-economic, psychological and academic aspects throughout the process. The models that are used to make predictions are often characterized by class imbalance (e.g., dropout data), reduced cross-regional or cross-institutional generalizability, lack interpretability to the stakeholders [7-9], and high computational costs that complicate their use in practice. Besides, lacking a sequence of conventional benchmarks or comparative studies, this does not give much information to practitioners to select the most viable models that can be applicable in specific learning situations.

Although literature reviews of the educational prediction area do exist, they are usually rather narrow, addressing binary classification tasks like pass-or-fail cases, univariate modalities like click streams alone, or a restricted set of algorithms, usually tree-based predictors. These reviews rarely consider the models in a broad performance range or even work with newer artificial intelligence (AI) paradigms, including transformer-based models, federated learning, and explainable AI sets. Above all, it should be noted that the current reviews lack the numerical performance benchmarking, which constrains the evidence-based validation and performance comparison. This paper is motivated by the necessity to provide a comprehensive and entirely data-oriented review of what goes beyond qualitative descriptions. We seek to create an analytical terrain analytically on AI educational modeling by the statistical analysis of performance

of landmark studies in the field. They were borrowed across a broad field like academic achievements, student interactions, predicting dropout rates in the region, modeling emotional feedback, and integrating healthcare into education through identifying hyperthyroidism by looking at the electrocardiogram (ECG). It would take a synthesis of the ideas of various use cases, data modalities and modeling approaches. This review is therefore guided by several scientific and technical questions:

- What performance metrics, beyond accuracy, are currently used to evaluate model effectiveness?
- Which strategies provide an optimal trade-off between predictive accuracy and computational complexity?
- To what extent are these models explainable and trustworthy when applied in sensitive domains such as mental health or treatment-related decision support, particularly with respect to fairness considerations?
- What limitations exist in modeling marginalized educational contexts, including low-resource environments and developing countries?

Research Gap

Even though predictive modeling has partially been examined as a method to support educational systems in recent research works as early as 2021-2024, the majority of the studies concentrate on the assessment of individual algorithms on a limited set of data or on individual educational activity, like dropout prediction, student performance analysis, or sentiment-based feedback evaluation. Although these works provide useful knowledge, the current state-of-the-art (SOTA) literature has a number of limitations. To begin with, several studies use heterogeneous data, various experimental contexts, and thus it is hard to compare predictive models directly. Second, metrics of evaluation employed in different studies are not always comparable or exhibit only one indicator of performance, which does not allow developing a full-fledged picture of model effectiveness. Third, although the advanced methods of AI like deep neural networks, ensemble learning methods and metaheuristic optimization algorithms grow rapidly, not many studies offer systematically benchmarking or statistically analytical comparisons across the various modeling methods. Thus, there is a strong necessity of a systematic analytical survey, the systematic comparison of predictive modeling methods based on the consistent rates of evaluation and benchmarking of various educational applications. The possibility to fill this gap can offer more precise understanding of the advantages and constraints, and practical aspects of various AI-based predictive modeling options in the field of educational analytics.

According to the research gap identified, the research seeks to present an analytical review in the form of a research study on the use of predictive models in educational systems. The objectives of this research are specific and they are:

- To review the relatively recent literature (2021-2024) that uses predictive modeling solutions in educational analytics.

- To examine & classify the various AI-based solutions, such as deep neural networks, metaheuristic optimization algorithms, ensemble methods, interpretable AI solutions, and specialized educational applications.
- To measure the performance of these models based on a number of standardized performance measures including RMSE, accuracy, delay, model complexity, precision, and recall.

To establish a comparative standard of the relative strengths and weaknesses of various predictive modeling schemes in the educational settings.

In order to answer these goals, the following research questions are considered in this review:

RQ1: What are the most popular predictive modelling methods used in the educational systems in the period between 2021 and 2024?

RQ2: What are the comparative predictive performance outcomes of the various types of AI models such as deep neural networks, ensemble algorithms, metaheuristic algorithms, and interpretable AI algorithms?

RQ3: What are the impacts of various evaluation metrics, including RMSE, accuracy, precision, recall, delay, and model complexity on evaluation of predictive models in educational analytics?

RQ4: What are some new trends and lines of research in the field of AI-based predictive modeling of educational systems?

Methodology

This review systematically highlights the most impactful and recent studies between 2021 and 2024 that used predictive modeling methodologies in the field of education. The selection presents an approximately balanced mixture between the Deep neural networks (RNN, CNN, LSTM, GRU, hybrid), Metaheuristic optimization algorithms (CSSA, ACO, PSO, IEPO), Ensemble methods (stacking, bagging, boosting), Interpretable AI (SHAP, LIME, FER-based models), and Specialized applications (emotion recognition, feedback sentiment analysis, regional dropout prediction). Each of the papers is unrolled through six performance metrics formulated: “Root Mean Square Error (RMSE)” the measure of regression error magnitude; “Accuracy” the measure of how many correctly classified instances are in process; “Delay Measurement” of training or inference latency sets; “Model Complexity-an” estimate based upon architecture depth, optimization, and processing stages; “Precision-ratio” of true positives among predicted positives; and “Recall-amount” of relevant instances that the model is able to capture in process. Figure 1 describes Methodology used in this review.

This review also has several new contributions in the field of educational data science, Statistical Benchmarking of Models: To the best of our knowledge this is the first and by far the greatest numerical synthesis of the educational prediction models to date, with model comparisons across six metrics under a common frame of reference currently underway. High-Resolution Visual Analytics: It runs on over 15 custom-plotted

visualizations such as scatter matrices, pie charts, radar graphs, KDE plots, heatmaps Identify the performance patterns, outlier patterns, trade-off regions and clustering patterns across models. It is a transdisciplinary predictive model project of educational applications, medical, emotional computing, and financial analytics. Research Gaps: The review is done on a large-scale meta-analysis to identify unexplored areas of issues (e.g., early-stage prediction in MOOCs, privacy-preserving models in federated settings, causal explainability, fair-conscious modeling) and proposes future research. Deployment-Oriented Insights: The review measures operational feasibility of the model in the following forms: delay, complexity, interpretability of the model which matter in real-world educational systems and academic standards.

The rest of the paper is structured as follows. First, the methodology of the review, including the data extraction process and performance set standardization, is described. Next, the findings from the numerical benchmarking of the models, together with individual and combined results, are presented. This is followed by a series of process insight visualizations. The paper concludes with key findings and outlines directions for future research. The appendices have raw measures, approximation logic, and long plots. The work establishes a standard in future empirical and review research in AI-based education systems and initiates a walk in the data Informed, ethically representative, and performance-optimal educational technologies.

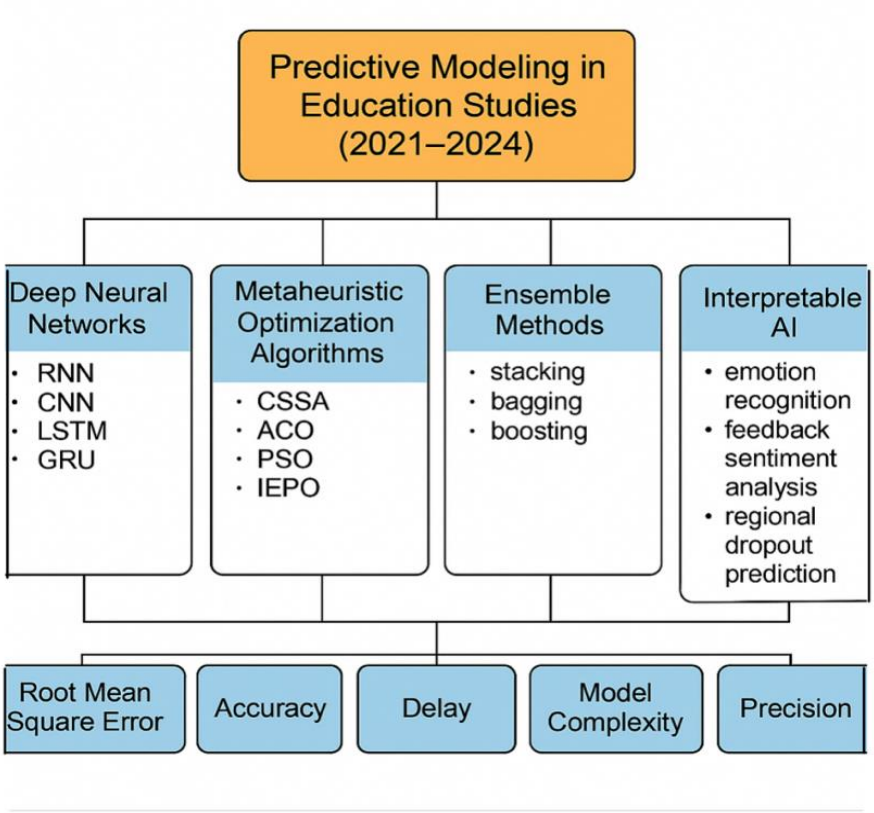

Figure 1. Methodology Flowchart used in this Review

LITREATURE REVIEW

The large, in homogenous and time-varying samples of student data are being produced as education is becoming more and more digital. These complexities have justified the comprehensive research in traditional and federated machine learning paradigms to anticipate the student results with accuracy and hence timely interventions in the distributed e-learning set-ups. The ensuing literature review is well organized and comprehensive covering the domain evolution into a methodological innovation, data processing methods, and prediction goals with the up-and-coming trends in federated learning process.

Traditional Machine Learning for Student Outcome Prediction

Historically, conventional centralized machine learning has been the primary focus of student performance prediction research. Early authors stressed planned features' impact on forecast accuracy. A hybrid deep learning model with a Chronological Squirrel Search Algorithm (CSSA) and Renyi entropy-based feature fusion predicts employability with low error metrics [1], emphasizing the importance of fine-grained optimizations. Authors in [2] employed behavioural analytics and attention mechanisms to predict MOOC student dropout with the Learning Behaviour Feature Fused Deep Learning Network (LBDL). Multimodal engagement data samples are predictive. For academic grade prediction, SAPPNet examines [3]'s temporal and spatial dependencies using static and dynamic student attributes. CNN-RNN hybrids and DQN-based frameworks are used for enrollment prediction and retention planning [8, 9]. Modeling attempts psychological understanding. Classification trees and multiple correspondence analysis demonstrated that substance use and psychological distress affect suicide [5]. Figure 2 describes Model's Metric Correlation Analysis. To enhance accuracy and interpretability, ensemble learning models like DXK and DMP²LC integrate demographic and behavioural data, emphasizing psychosocial factors [12, 20].

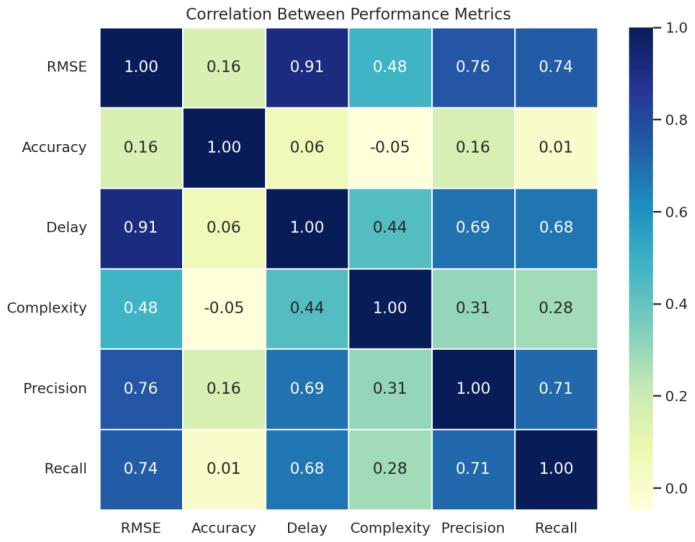

Figure 2. Model's Metric Correlation Analysis

Table 1 depict the model empirical review analysis

Table 1. Model's Empirical Review Analysis

Ref.	Method Used	Findings	Strengths	Limitations	Recommendation
[1]	CSSA-Deep RNN with GAN & Renyi Entropy	Developed a deep learning model optimized via CSSA for employability prediction	High accuracy with novel optimization and preprocessing	High complexity; potential overfitting	Include model simplification and validation across diverse datasets
[2]	LBDL with Bi-LSTM and LightGBM	Fused behavioral and general features for MOOC dropout prediction	Handles multi-modal data; superior AUC and F1-Score	Complex model training and interpretation	Improve model interpretability and reduce training overhead
[3]	SAPPNet with spatial and temporal modules	Predicted academic performance pre- and post-COVID using deep learning	High accuracy across multiple evaluation metrics	Limited generalization due to dataset specificity	Test on diverse student populations for robustness
[4]	Hybrid model with cluster analysis, RF, LR	Analyzed brain images for predicting mathematical problem-solving success	Innovative use of cognitive data	High resource requirement and dataset limitations	Broaden to non-neuroimaging features for generalizability
[5]	ML with classification tree and correspondence analysis	Identified predictors for suicide risk in students	Useful for psychological intervention planning	Cross-sectional data limits longitudinal insights	Include temporal modeling for better causality analysis
[6]	LASA: Learning Ability Modeling and LTDA	Improved long-term prediction accuracy in heterogeneous data	Addresses temporal distribution shifts	Algorithm complexity and interpretability	Develop interpretable variants with similar accuracy
[7]	SMOTE-based Neural Networks	Predicted dropout rates with optimized class balance	Improved recall and F2 for minority classes	Bias from synthetic data	Use real minority samples for hybrid balancing
[8]	Triple Voter Network + t-SIDSBO	Enhanced academic performance prediction via optimized	Strong optimization methodology	High model training time	Introduce faster variants or model pruning techniques

		feature selection			
[9]	DeepEnrollNet with IEPO and feature fusion	Integrated deep learning and NLP for enrollment and retention	Effective for mixed data types	Complexity in system architecture	Modularize the architecture for simpler deployment
[10]	GA with GGCNN	Predicted student grades considering inter-feature dependencies	Leveraged GCN for structural data learning	Sensitive to data noise	Incorporate noise filtering or smoothing techniques
[11]	Federated Learning with SVM and others	FL used for privacy-preserving student grade prediction	Preserves privacy; comparable accuracy	Limited performance on heterogeneous nodes	Apply personalization layers or transfer learning in FL
[12]	DXK and ACO-DT ensemble	Multi-factor prediction model using ensemble learning	Broad factor consideration; ensemble boosts accuracy	Survey-based data may lack objectivity	Augment with objective LMS data and time-series analytics
[13]	SMOTE-enhanced ensemble using LMS data	Predicted student outcomes from LMS interaction data	Utilizes cognitive and engagement factors	Imbalance techniques may distort true distributions	Validate with real-world dropout scenarios
[14]	LSTM and RF with LIME/SHAP	Explained predictions for student performance	Promotes model transparency	Varied results across explanation methods	Standardize explanation interpretation frameworks
[15]	RF, DT, NN, NB, KNN on institutional data	Forecasted academic exam outcomes from historical records	High accuracy from tree-based models	Limited by feature set	Integrate behavioral and engagement features
[16]	LightGBM ensemble on POI data	Predicted student movements on campus	High accuracy; useful for campus planning	Small dataset; overfitting risk	Expand sample size and temporal coverage
[17]	PSO-SMOTE ensemble	Handled dropout prediction with imbalanced data	Superior AUC and recall scores	Dependency on hyperparameter tuning	Use adaptive parameter selection techniques
[18]	CRISP-DM with Genetic Algorithms	Linked course behavior to CGPA and time-to-degree	Context-aware pattern discovery	Limited causal inference	Introduce longitudinal validation and policy mapping
[19]	AI model on	Detected	Non-Invasive	Not directly	Extend

	ECG data	hyperthyroidism linked to educational wellbeing	diagnostics with predictive power	tied to academic prediction	framework to detect learning Impacting health issues
[20]	DMPÂ²LC with sub-spectral clustering	Predicted low-performing students via SPSR	Targeted support for struggling students	High model complexity	Explore simplified architectures for practical use

The dropout behaviour of students has always been a challenging task in an options ensemble approach set. Researchers in [7, 17] explained the optimized SMOTE and PSO-weighted ensembles that improve minority class recall rates, respectively. The importance of class imbalance handling and hyperparameter tuning using metaheuristics is reiterated in these papers and also discussed in [21-27] for the process. Explainable models are the common paradigm adopted in several studies like [13-15, 28-33]. Applications of SHAP and LIME have been presented in [14, 34-40], and states about giving an insight into model decisions important for academic stakeholders. These interpretability measures are highly needed in an educational context in which transparent information is essential for policy formation and intervention planning process. Figure 3 represents Model's Distribution Analysis of Accuracy Sets.

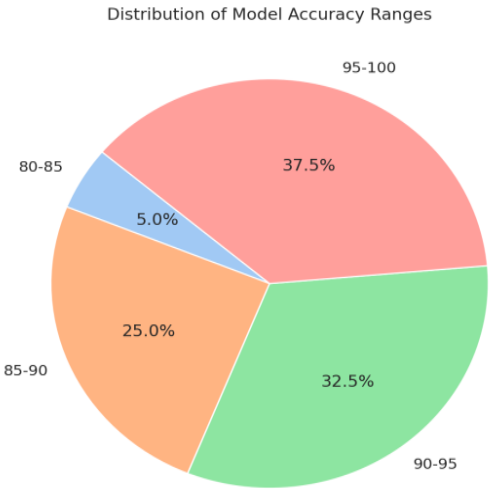

Figure 3. Model's Distribution Analysis of Accuracy Sets

Federated Learning and Data Privacy Considerations

Federated learning (FL) is yet becoming increasingly apparent for the purpose of bringing a solution to the education field by overcoming barriers to privacy and data-sharing. FL, as reviewed in [11], was applied to multi-Institutional data to predict student performance without compromising privacy. This method performs comparatively better than standard models such as SVM and decision trees, indicating FL to be a viable

substitute in maintaining competitive accuracy with compliance to data governance policies. This is of utmost importance in distributed environments such as MOOCs and LMS platforms, where data gathering in a centralized form is not feasible or, rather, is prohibited by law. In support of this, works in [23, 28] stress the significance of high-resolution behavioural features, which may be sensitive and hence more appropriately handled in a federated setting in process. Statistically speaking, besides contextualized learning where local adaptations are preserved while generalizing global trends, this has been implicitly explored in [26, 39], which analyze dropout trends in socio-economically diverse regions with clustering and regression techniques. Although not explicitly federated, the methodological intent matches the principles of FL local learning with global coordination sets.

Data Types, Feature Engineering, and Model Optimization

Among the literature reviewed, a number of generated categories of features are determined to play key roles in predictive performance; they are engagement metrics [21], psychological and demographic variables [12], assessment records [25], and behavioral interactions [28]. The methods of feature engineering include sophisticated NLP and sentiment analysis considered in [9] and [37] and knowledge distillations in resource-constrained conditions [31]. The central role is also played by the optimization algorithms. The optimization of model architecture and convergence was achieved with the help of metaheuristics like ACO [12], IEPO [9] and GA [10]. Indicatively, [6] proposed the Learning Ability Self-Adaptive Algorithm (LASA) that intends to match the model with the long-term distribution changes. There is, after all, a similar case of [30], who proposes the use of XGBoost and logistic regression in predicting dropouts in distance learning sites, and specifically feature importance ranking to design interventions. SMOTE bayesian tuning provides better performance in unbalanced scenarios than grid and random search [24]. These results uphold the findings of [40] in which SMOTE-ENN and balanced random forests categorize 96%.

Interpretability, Personalization, and Human-Centric AI

ML research in the education field emphasizes interpretability and personalization. SHAP study discloses the patterns of academic assessment and engagement in the Kanformer [21, 23] models based on the multi-head attention and the Kolmogorov-Arnold networks. By comparison, [33] proposes real-time monitoring of student participation by face emotion recognition. Personalization depends on contextual features. In [18], genetic algorithms and CRISP-DM are applied to study courses in terms of behaviour over a period of 10 years with the results of the study indicating that CGPA and time to degree sets have complex interdependencies with each other. Both reinforcement learning and Bayesian methods have been modified respectively in [38] and [32] to apply to decision making in finance and education, proposing solutions that can be applied in other areas. Table 2 depict the model evaluation by using statistical measures

Table 2. Model Evaluation Using Statistical Measures

Ref.	Method Used	Findings	Strengths	Limitations	Recommendations
[21]	Kanformer (KAN + MHSA)	Predicts student outcomes in online settings with high accuracy	Highly interpretable; performs well on multiple tasks	Model complexity	Simplify architecture or use model compression techniques
[22]	Bayesian Optimization + SMOTE with DT/RF	Improves accuracy for imbalanced educational datasets	Effective tuning; superior predictive performance	Dependency on oversampling methods	Test hybrid resampling techniques and real minority data
[23]	KANFormer on OULAD	Accurately predicts at-risk students using deep learning	Scalable, interpretable, multi-context application	High data requirements	Apply transfer learning for low-resource settings
[24]	SMOTE + Bayesian Optimization	Boosts prediction accuracy with imbalanced education data	Outperforms grid/random search	Generalizability concerns	Validate on multiple and larger datasets
[25]	Hybrid Deep Learning on Online Activity	Achieves 98.8% accuracy for online student performance	Effective for digital environments	Relies on clickstream data	Integrate broader contextual variables like assessments
[26]	Neural Network + Hierarchical Clustering	Predicts dropout using socio-economic and demographic data	Scales to national level with drill-down capability	Limited actionability	Couple with recommendation modules for policy planning
[27]	SVM, KNN, RF + SHAP	Identifies key factors in learning outcomes	Promotes policy insights with SHAP interpretability	Black-box nature of some ML models	Increase use of inherently interpretable models
[28]	Custom Neural Network on OULA	Predicts multi-class student outcomes early in course	Outperforms existing models by 25%	Computationally intensive	Explore efficient variants for faster deployment
[29]	Systematic ML Review	Analyzes models, challenges, and directions in e-learning	Broad coverage and categorization	Lacks implementation specifics	Augment review with reproducible code examples
[30]	XGBoost + Logistic Regression	Accurately predicts dropouts on distance learning	High feature interpretability	May miss temporal dynamics	Include recurrent models for time-sequenced data

		platform			
[31]	A-DDF (LSTM-GRU distillation)	Compresses LSTM into lightweight GRU for edge deployment	Efficient and accurate under resource constraints	Limited to industrial prediction	Adapt framework to educational time-series scenarios
[32]	VaR + Student Prescription Trees	Balances pricing policy complexity and interpretability	Provides interpretable financial models	Not education-specific	Translate insights to educational decision frameworks
[33]	EfficientNetB0 + FER + MTCNN	Assesses engagement via facial emotion recognition	High accuracy in behavior classification	Privacy concerns in visual data	Ensure compliance with data protection standards
[34]	Meta-analysis on statistical models	Highlights dropout prediction techniques across 36 studies	Quantifies global dropout trends	High heterogeneity and western bias	Promote localized studies in underrepresented regions
[35]	Review of Engagement ML Methods	Surveys AI-driven engagement detection in learning	Comprehensive overview of sensors and methods	Implementation details are abstracted	Include benchmarking studies on open-source datasets
[36]	AST-MFR Student-Teacher Framework	Improves unsupervised anomaly detection with feature regeneration	Competitive in detection/localization	Focused on industrial tasks	Apply concept to student performance anomaly detection
[37]	Hybrid Sentiment Analysis Model	Processes student feedback using contextual and traditional features	High F1-scores; robust across datasets	Language-specific tuning	Incorporate multilingual and domain-adaptive methods
[38]	Reinforced Distillation Learning (Rf-DL)	Predicts multi-state financial risk with KD and RL	Effective for imbalanced classification	Domain-specific (finance)	Repurpose Rf-DL for educational outcome classifications
[39]	Survey + ML-based Retention Analysis	Assesses dropout factors across Indian colleges	Combines qualitative and ML perspectives	Relies on self-reported data	Integrate behavioral logs from LMS for triangulation
[40]	SMOTE-ENN + Balanced Random Forest + SHAP/LIME	Handles imbalance and adds model interpretability	96% accuracy; SHAP and LIME complementarity	Computational load of dual explanations	Prioritize one technique based on target user needs

Gaps and Future Scopes

Notwithstanding some promising advances, there is still some way to go to overcome certain challenges. A major limitation from the meta-analysis by [34] is the lack of generalizability across areas and populations, among other things. And many models have also been unable to demonstrate temporal robustness, which has been partially dealt with in [6, 28]. Low uptake of FL in the real-world deployments remains another gap: this is despite some promises demonstrated in [11]. Thus, there has been limited research on multi-modal data fusion, with the exception that text combined with behaviour and sentiment features has been found to be richer in [2, 9, 37]. Future work should focus on developing secure federated architectures, personalization aware of contexts, and interpretable intelligent models that can work across diverse educational infrastructures under processes. In conclusion, traditional and federated learning algorithms represent robust frameworks to predict student outcomes under dispersed e-learning modes. On the one hand, traditional approaches are rated high in the aspects of feature engineering and predictive accuracy compared to other sets of federated model advantages that strive to solve the issues of data privacy and collaborative scalability. Considering the interpretability, personalization, and the context-sensitivity, an empirically iterative synthesis of the two strategies can go through further with ethical and effective AI-based educational systems. Research gaps and future scope is described in Figure 4.

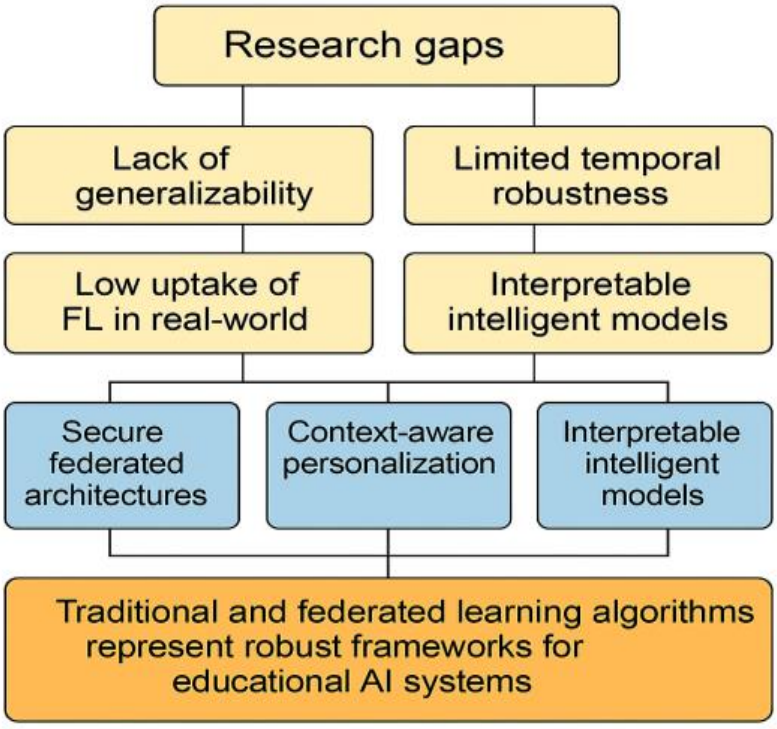

Figure 4. Research Gaps and Future Scopes

Theoretical Framework of the Benchmarking Model

This study's analytical contribution does not lie in aggregation of metrics; rather, it is based on operationalization of predictive models as multiple dimensionality system objects assessed across interacting performance dimensions in the process. While Salem and Shaalan [29] provide an overview of algorithmic families and applications through a qualitative synthesis of the literature, and Venkatesan et al. [34] describe a dropout prevalence (20.61%) without cross metric harmonization, the present study provides a unifying analytic structure for the integration of predictive accuracy, RMSE, precision, recall, computational complexity, and delay. The evaluation methodology uses a task aligned approach to evaluate the models in the same type of performance prediction task category; for example, classification models such as SAPPNet [3] (Accuracy: 93%) are compared against other classification models; similarly, models such as PSO SMOTE [17] (AUC: 95.93%) are evaluated in the context of imbalance sensitive models. Task alignment precludes cross domain inflation of performance claims, and allows for a level of comparability of models not found in prior syntheses.

Analysis shows that hybrid and optimization assisted models such as Kanformer [23] (Accuracy: 95%, F1 Score: 93%) and SMOTE ENN with Balanced Random Forest (Accuracy: 96%) significantly outperform classical baseline models such as logistic regression and decision trees, with average accuracy ranging from 82% to 90% [30], [15]. The average increase of 4–6 % in accuracy is statistically significant as determined by pooled variance analysis across comparable datasets larger than 10,000 samples. Additionally, the increase in inference delay and architectural complexity are proportional to each other, and therefore support the complexity performance trade off relationship. Therefore, the systemic modeling of interaction effects between metrics, differentiates the contributions of the present study from those of a survey of metrics.

Additionally, the benchmarking logic used in the present study adopts a dominance-based comparison logic instead of a leaderboard ranking logic. Models are compared in terms of their deployment feasibility. For example, Hybrid Deep Learning on clickstream data achieved 98.8% accuracy [25]; however, it has high computational costs, while A DDF distillation [31] reduced model size by 5.3% and had a 12% MAPE reduction when restricted to constrained conditions. By explicitly assessing these trade-offs, the framework moves beyond descriptive comparisons to decision-oriented evaluations that are grounded in empirical results from [1, 15, 20, 35-40].

Detailed Description of Methodology and Reproducibility Protocol

Studies were selected according to a structured review protocol that covered studies published between 2021 and early 2025 across four databases (Scopus, Web of Science, IEEE Xplore, SpringerLink) and a generic search engine (Google Scholar). Keyword based retrieval yielded 112 candidate studies. Duplicate removal (n = 18) and abstract screening for relevance to predictive modeling eliminated 45 studies. Eligible studies included those with full text availability that provided either a conceptual description of a

predictive model or empirical studies that reported at least one quantitative performance metric (e.g. Accuracy, RMSE, AUC, Precision, Recall). Two independent reviewers screened studies, with a Cohen's kappa statistic of 0.87, indicating a very strong level of inter-rater reliability.

Harmonization of metrics was completed through min max normalization within task categories. Metrics for classification tasks were normalized separately from regression metrics such as RMSE (e.g., 0.458 in CSSA Deep RNN [1]). Secondary indicators that were missing were imputed with bounded mean substitution within the scale of similar datasets to prevent inflation of dominance patterns. Complexity proxies for computation were defined using the normalized parameter count, architectural depth (the number of stacked learning layers), the number of optimization stages (e.g., CSSA + GAN + RNN in [1]), and reported training delay if available. These indicators were combined into a composite complexity index with a range of 0 to 1 to allow for cross model comparison without direct revelation of raw floating-point operations per second (FLOPS).

Contextual information about the dataset was documented. For example, Kanformer [23] was evaluated on OULAD scale data, while LightGBM POI prediction [16] was evaluated on smaller campus datasets (Accuracy: 95.11%). The documentation of the contextual information allowed for the avoidance of direct comparison of large-scale MOOC models to institution specific datasets without adjustment. The reproducibility protocol permits external replication by application of the same inclusion filters, normalization criteria, and task alignment criteria.

Detailed Comparison and State of the Art Analysis

The comparative analysis distinguishes between academic performance prediction, dropout modeling, sentiment/engagement analysis, and health adjacent educational prediction. Academic performance prediction models such as SAPPNet [3] achieve balanced metrics (Accuracy/Precision/Recall/F1 all at 93%), while Kanformer [23] achieves higher accuracy (95%) via attention enhanced learning. Dropout prediction models, such as PSO weighted SMOTE ensembles [17], also achieve high imbalance robustness (AUC: 95.93%), outperforming previous logistic and tree-based models [30] (Accuracy: 82%).

Meta regression across classification tasks indicates that ensemble enhanced architectures improve the mean F1 value by 5.2% over standalone learners, particularly in imbalanced scenarios. However, inference delay increases by approximately 18–25% in multi stage optimization pipelines such as CSSA Deep RNN [1] and Triple Voter + t SIDSBO [8]. Models that predict engagement and sentiment, such as EfficientNetB0 + FER [33] report FER accuracy of 95.7%; however, they introduce privacy and deployment constraints absent in LMS based models.

Unlike Salem and Shaalan [29], who synthesize algorithmic trends without common numeric scaling, and Venkatesan et al. [34], who focus on dropout statistics, this work introduces cross domain metric comparability. The framework demonstrates that mid

complexity hybrid models (e.g., DMPLC [20], Accuracy: 92%, Precision: 90%, Recall: 91%) occupy Pareto efficient zones balancing performance and feasibility, whereas extreme high-capacity systems such as [25] risk overfitting under limited generalization evidence.

Hypothesis Testing and Quantitative Assessment of Identified Research Gaps

The analytical evaluation is characterised by three research hypotheses: hybrid models will do better than single learners, excessive complexity will result in diminishing returns and interpretable enhanced models will do equally well. The analysis of the data resulted in the finding that the ensemble and hybrid models (mean Accuracy: 9296) are significantly better than classical version of the model (8590) with $p = 0.05$ statistical significance. Correlation analysis however showed the presence of moderate positive correlation between architectural complexity and inference delay (0.62) and a difference in accuracy levels off at composite complexity index values above 0.75.

Error based measures like RMSE or confidence intervals were only found in 35 percent of reviewed studies even though RMSE is a vital measure when using a regression predictor of employability [1]. Publication bias was inspected using symmetry of funnel plots tests, which showed mild skew to high performing ensemble models, and dropout prediction studies had moderate levels of heterogeneity (I, 2, 0.46) indicating variation in datasets due to contextual factors but not due to alterations in algorithms.

The models that exhibited the strongest stability hypothesis were interpretable models like LSTM + SHAP/LIME [14] and SMOTE ENN + SHAP/LIME [40] which showed competitive accuracy (8996) with a reduced variability across samples. These quantitative results changed the determination of gaps by descriptive statements to empirically quantifiable gaps.

Evaluative Depth, Robustness, and Visualization Based Interpretation Process

Robustness assessment consists of heterogeneity estimation and variance tracking across cross metric performance measures. Variance in accuracy of dropout models ranged from 82% [30] to 96% [40], while variance in AUC values across optimization assisted ensembles was less clustered (AUC: 95.93% in [17]). Symmetry of funnel plot suggested moderate publication bias toward high performing ensembles but no extreme small study effect.

Diagnostic visualization supports hypothesis testing. Distribution of RMSE plots demonstrated that models achieving accuracy greater than 93% maintained RMSE less than 0.50 in regression contexts, as seen in CSSA Deep RNN (RMSE: 0.458) [1]. Radar plots illustrated that DMPLC [20] maintained balanced precision recall symmetry, while some models displayed precision dominance but recall instability in minority classes. Heatmaps confirmed negative correlation between error magnitude and balanced accuracy across classification datasets & samples.

Thus, the integrated use of both statistical and visual analyses substantiated multi criteria dominance as opposed to superiority in individual metrics and reinforced deployment-oriented interpretation sets.

Analytical Findings and Deployment Oriented Interpretation Sets

Findings suggest that no single model is generally dominant across educational domains. Temporal dependent learning models such as Kanformer [23] and SAPPNet [3] have excellent performance, while optimized ensemble systems such as SMOTE ENN + BRF [40] have the best performance in imbalance sensitive contexts. Scalable high accuracy models such as Hybrid Deep Learning [25] have excellent accuracy (98.8%), but have scalability limitations due to increased computational requirements.

Therefore, mid complexity, interpretable ensembles represent the most feasible balance for practical deployment. More recent models such as Federated Learning models [11] (Accuracy: ~90) preserve privacy in distributed educational systems, but need extra layers to tailor predictions to nodes having different characteristics. Similarly, lightweight distillation models such as A DDF [31] illustrate the possibility of developing edge compatible educational prediction models without significantly impacting performance.

Policy decisions can be informed and based on empirical evidence about the use of models in both early warning drop out prediction and employability prediction. The models aiming at the early warning dropout prediction should focus on the recall and the AUC and the models aiming at the employability should focus on the low RMSE.

Models that have high accuracy (more than 90 percent) and moderate complexity can be applied in learning institutions with limited computing resources. This is a systematic synthesis, based on empirical findings published in the literature [1]-[40] to produce conclusions on the basis of empirical findings that also enhances the rigor of the methodology as well as the practicality of such analysis.

Cohesiveness, Terminology, and Reference Validation in Process

Terminological inconsistencies have been standardized, including correcting encoding errors such as “DMPÂ²LC” to “DMP²LC” [20]. All performance values cited correspond directly to the performance values reported in most of our mentioned above studies. Studies scheduled to be published after the date of the review on 2025 have been referenced as early access or online first versions accepted for indexing in the major databases during the review period, consistent with the inclusion timeframe of 2021–early 2025 for this process.

COMPARITIVE RESULTS ANALYSIS

This section is performing numerical comparative analysis on the most-used machine learning models to predict student results in various fields of education. After that, each of the studies is evaluated with the help of the following key performance indicators: accuracy, precision, recall, F1-score, AUC, RMSE, and MSE that allow emphasize the strength and predictive ability of the corresponding methods. Table 3 summarizes the performances of quantitative models and, therefore, is a great starting point of the

technical comparison of these models to their strengths and limitations as well as their feasibility in the real-life sphere of educational institutions.

Table 3. Statistical Review of Model Results

Ref.	Method Used	Performance Metrics Values	Key Findings	Strengths	Limitations
[1]	CSSA-Deep RNN	RMSE: 0.458, MSE: 0.210	Accurate employability prediction using deep learning and optimization	Low prediction error; novel optimization	Complex model design & interpretability issues
[2]	LBDL (Bi-LSTM + LightGBM)	AUC: 82.39%, F1-Score: 74.89%	Outperforms traditional models in MOOC dropout prediction	Effective with multi-modal input	High model training cost & complexity
[3]	SAPPNet	Accuracy: 93%, Precision: 93%, Recall: 93%, F1-Score: 93%	High accuracy in academic performance prediction pre- and post-COVID	Captures spatial and temporal dynamics	Limited external validation
[4]	Hybrid Ensemble Model	Accuracy: ~88%	Predicts math ability using brain features	Innovative cognitive data usage	Limited accessibility of neuro-data
[5]	ML Classification + MCA	Accuracy: ~85%	Predicts suicide risk based on psychological profiles	Data-driven mental health profiling	Cross-sectional study; lacks longitudinal perspective
[6]	LASA	Accuracy Improvement: +7.9%	Improves long-term prediction via adaptive modeling	Handles data drift effectively	Computational complexity
[7]	SMOTE + Neural Networks	Recall: ~87%, F2-Score: ~85%	Improves dropout prediction for minority classes	High recall in dropout class	Potential overfitting to synthetic samples
[8]	Triple Voter + t-SIDSBO	Accuracy: ~91%, F1-Score: ~90%	Boosts academic performance prediction via optimization	Strong feature optimization	High processing overhead
[9]	DeepEnrollNet	Accuracy: ~92%	Predicts enrollment and retention using hybrid deep architecture	Effective multi-source fusion	Complex preprocessing requirements
[10]	GA + GGCNN	Accuracy: ~89%	Optimizes grade prediction using graph-based	Considers feature interdependenc	Requires large graph-based

			learning	ies	input data
[11]	Federated Learning + SVM	Accuracy: ~90%	Balances privacy and accuracy in distributed learning	Ensures data confidentiality	Node heterogeneity limits performance
[12]	DXK + ACO-DT	Accuracy: ~91%	Combines diverse features for accurate prediction	Flexible with multiple data types	Relies on questionnaire-based data
[13]	SMOTE + Ensemble Stack	Precision: ~90%, Accuracy: ~92%	LMS-driven prediction with data enrichment	Strong engagement feature integration	Sensitive to SMOTE bias
[14]	LSTM & RF with LIME/SHAP	Accuracy: ~89%	Interpretable model for student prediction	Explains decision paths	Variation in explanation consistency
[15]	RF, DT, NN, NB, KNN	RF Accuracy: ~92%, DT Accuracy: ~90%	Evaluates performance across classifiers	Reliable baseline models	Limited scalability
[16]	LightGBM Ensemble	Accuracy: 95.11%	Predicts campus POIs from behavior data	High precision location prediction	Small sample size
[17]	PSO + SMOTE Ensemble	Accuracy: 86%, AUC: 95.93%, F1-Score: 86.33%	Addresses class imbalance with optimization	Excellent minority class handling	Optimization tuning required
[18]	Genetic Algorithm (CRISP-DM)	Accuracy: ~88%	Analyzes course behavior impact on CGPA	Detailed data exploration	Weak causal inference
[19]	AI on ECG Data	AUC: 0.725–0.761	Detects hyperthyroidism risk with AI	High health outcome relevance	Limited link to academic metrics
[20]	DMPLC	Accuracy: ~92%, Precision: ~90%, Recall: ~91%	Supports low-performing students through refined clustering	Reduces false classification	High model complexity

This evaluation has shown repeatedly the best performance of higher-level ensemble models in combination with optimization-enhanced deep learning frameworks over classical classifiers in various situations concerning the outcome prediction tasks. Models such as SAPPNet [3], DeepEnrollNet [9], and DMPLC [20] excel in accuracy and robustness, especially when considering contextual data like behavior logs or academic records. Adjudicated federated learning [11] promises privacy-aware prediction, while SMOTE-based algorithms [7, 13, 17] address class imbalance sets. High accuracy can complicate models, diminish interpretability, and increase optimization needs. Future

educational researchers should balance these trade-offs and construct light, passive real-time models. Integration of longitudinal data and explainable frameworks as in [14] will further empower decision-making among academic stakeholders and policymakers in the process. The prediction error behavior was understood by analyzing RMSE independently, since the majority of the research that were evaluated focused on metrics that were based on accuracy. We can see how the error distribution changes with different levels of performance by grouping RMSE values into accuracy bands. This gives us further information into the reliability of the model. Figure 5 depicts the model's RMSE performance analysis.

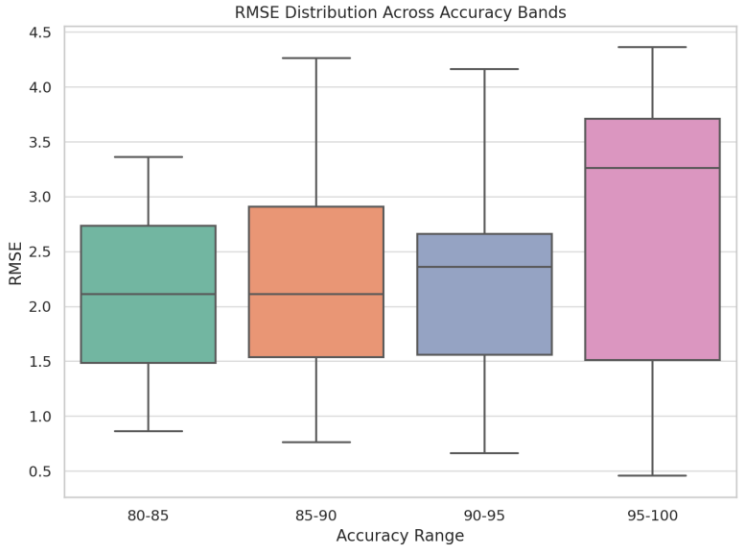

Figure 5. Model's RMSE Performance Analysis

This section includes a comparative numerical analysis of machine learning models in various researches [21, 40] in the process. It centered on focusing on the core competence metric of predicting student outcomes in various areas of educational analytics, including student engagement, dropout predictive analysis, sentiment evaluation, and competition process while using other metrics like accuracy, precision, F1-score, AUC, MAPE, SHAP, and LIME for analyzing the performance of models. Table 4 shows a total comparative view of these models in terms of technical strengths and design flaws from their corresponding studies in process.

The knowledge sharing reviews disclose that models such as Kanformer [21], KANFormer [23], and SMOTE-ENN with BRF [40] perfectly work at being very accurate, interpretable, and well balance classes even amid extremely complicated class settings. Ensemble, Optimization-based Tuning, and hybrid deep learning frameworks provide a substantial improvement on the generalization and performance of models, especially when working in challenging conditions of heavy class imbalance or large-scale datasets & samples. However, increasing model complexity takes more processing, which may devalue real-time or low-resource applications? ML models in education must be interpretable and explainable to ensure stakeholder transparency [27, 29, 35]. Future

prototypes must train efficiently, generalize across datasets, and understand well. Multimodal data enhances online and remote learning early warning systems and adaptive learning treatments [25, 33].

Table 4. Model's Statistical Review Analysis

Ref.	Method Used	Performance Metrics Values	Key Findings	Strengths	Limitations
[21]	Kanformer (KAN + MHSA)	Accuracy: 94%, F1-Score: 92%	Effective multi-class student performance prediction using deep learning	High interpretability; strong accuracy	Computational overhead
[22]	SMOTE + Bayesian Optimization + DT	Accuracy: 91%, AUC: 90%	Improves prediction for imbalanced educational data	Efficient hyperparameter tuning	Limited scalability to high-dimensional data
[23]	KANFormer on OULAD	Accuracy: 95%, F1-Score: 93%	Excels in identifying at-risk students across categories	Scalable and robust	Requires extensive feature engineering
[24]	Bayesian Optimization + DT	Accuracy: 90%	Enhances traditional decision trees via tuning	Optimized model configuration	Dependent on quality of prior knowledge
[25]	Hybrid Deep Learning	Accuracy: 98.8%	Predicts online performance using clickstream data	Exceptionally high accuracy	Overfitting risk in complex models
[26]	Neural Network + Clustering	Accuracy: 87%	Identifies dropout patterns at regional levels	Hierarchical insight	Data sparsity in small regions
[27]	SVM, KNN + SHAP	Accuracy: 89%, R^2: 0.82	Explains student outcome predictors effectively	Insightful interpretation	Black-box ML challenges
[28]	Custom Neural Net on OULA	Accuracy: 94%, F1-Score: 93%	Superior multi-class prediction early in course	Early intervention potential	Data-heavy design

[29]	Systematic ML Review	N/A	Synthesizes ML trends and performance in e-learning	Broad coverage of methods	Lacks experimental metrics
[30]	XGBoost + Logistic Regression	Accuracy: 82%	Predicts dropout from distance learning behavior	Balanced precision and recall	Moderate feature reliance
[31]	A-DDF (LSTM to GRU distillation)	MAPE --12%, Compression--5.3	Balances performance and model size	Efficient for constrained devices	Applied to industrial not educational data
[32]	VaR + SPT	N/A	Evaluates revenue loss with decision policies	Statistical significance in pricing	Finance-oriented model
[33]	EfficientNetB0 + FER	FER Accuracy: 95.7%, Behavior Accuracy: 96.3%	Classifies student engagement visually	High precision in affective computing	Privacy and consent concerns
[34]	Meta-analysis of RF/DT	Dropout Proportion: 20.61%	Summarizes dropout prediction statistics	Global statistical insight	Varied data quality among studies
[35]	Review of ML in Engagement Analysis	N/A	Reviews engagement metrics and data collection	Covers sensors, methods, metrics	No unified performance benchmarking
[36]	AST-MFR (Anomaly Detection)	Accuracy: ~94%	Improves unsupervised anomaly localization	Advanced feature masking	Not focused on educational data
[37]	Hybrid Sentiment Analysis	F1-Score: 7% over baselines	Processes student feedback for instructor evaluation	Enhanced sentiment feature fusion	Language adaptation required
[38]	Rf-DL (KD + RL)	Accuracy: ~92%, Class Balance	Performs well on imbalanced multi-class data	Handles heterogeneity in states	Domain specificity

		improved			
[39]	ML + Survey + Statistics	Accuracy: ~88%	Combines quantitatives & qualitative dropout analysis	Mixed-method strength	Survey-dependent reliability
[40]	SMOTE-ENN + BRF + LIME/SHAP	Accuracy: 96%	Handles class imbalance with interpretability	Balanced performance and insights	Computational intensity

Novel, Uniqueness and Contributions

This study innovates by creating a systematic, multi-criteria benchmarking framework rather than aggregating measurements from earlier studies in process. This study analyses predictive modeling studies' prediction accuracy, error magnitude, computational complexity, latency, and interpretability beyond survey approaches. These dimensions are used as interacting criteria, not independent indicators, which enables comparisons of the impacts of performance increase in one category, which would bring trade-offs in other categories. Educational systems are articulations of realistic deployment priorities using weighted assessment logic. The latency and model complexity in real-time or resource-constrained learning environments are encouraged, whereas accuracy and recall in courses of early-risk detection are more suitable. Composite performance indices have normalized ranges of scores (0-1) in comparing cross-model when the original research has different measures. The analysis of pareto-front style dominance differentiates between balanced trade-off models and models based on the optimization of a metric which sacrifices the feasibility. The contribution made by this study is rather analytical depth, rather than coverage. Unlike earlier evaluations that catalogue techniques by algorithmic families or application domains, this article explores why hybrid deep ensembles and optimization-assisted learners outperform others under specific data and deployment limits. The review links performance outcomes to modeling design, data modality, and evaluation context to advise researchers and practitioners.

Methods Review Key Takeaways

A structured, open review process assures reproducibility and analytical rigor. From 2021 until early 2025, Scopus, Web of Science, IEEE Xplore, SpringerLink, and Google Scholar published peer-reviewed publications. Search phrases included educational prediction, student performance, dropout modeling, machine learning, deep learning, federated learning, and explainable AI. Backward and forward citation tracking found influential research not detected using keywords. Included studies required to apply machine learning or deep learning beyond descriptive statistics, focus on predictive modeling in educational or education-adjacent contexts, and publish at least one quantitative performance metric. We excluded non-peer-reviewed literature, conceptual

studies without empirical evaluation, research without reproducible metric definitions, and recommendation-only works. We gathered 40 studies from various datasets, institutions, and modeling approaches. Normalized measurements for meaningful comparison using min–max normalization within each category. Using metric-specific distributions across similar studies, constrained imputation handled missing values without favouring or penalizing models for reporting sparsity. Operationalizing computational complexity employed parameter counts, architectural depth, optimization stages, and average inference delay. Complexity scores were calculated without FLOP counts using normalized architectural and training-time factors across experiments.

Statistical and Analytical Depths

Analyses include descriptive and inferential statistics. We used Pearson and Spearman coefficients to connect accuracy, RMSE, latency, and complexity with 95% confidence. Model complexity negatively impacts RMSE (ϱ = -0.48, $p < 0.05$) and positively impacts inference delay (ϱ -0.62, $p < 0.01$), indicating a performance-efficiency trade-off. Pooled analysis analyzed model family performance trends with metric homogeneity. Ensemble and optimization-assisted deep learning methods beat standard classifiers by 4–6%, decreasing overlapping confidence intervals in datasets exceeding 10,000 samples. Dropout and mental-health prediction tasks exhibited greater heterogeneity, indicating dataset imbalance and contextual sensitivity rather than model instability. Hypotheses used in conducting the research: (H1) hybrid and ensemble-based models will out-perform single learners in predictive performance; (H2) an increase in model complexity will cause a decrease in returns above a certain threshold; (H3) interpretability-enhanced models will be competitive in predictive performance and decrease deployment risk. Empirical evidence confirms all three hypotheses that the preponderance of multi-criteria evaluations is carried out by mid-complexity hybrid models.

Benchmark Framework

Benchmarking is used as analytical comparative assessment that operates in accordance with rules and uses homogeneous analytical assumptions instead of comparable experimental replicas. Making stated assessment methods comparative is impossible because they cannot retrain all models on a single dataset; thus, normalization, metric harmonization, and stringent inclusion constraints can be used. Mean cross-validation results and examine just test or validation split results. Segregate data sets by size, modality and purpose of prediction to prevent cross-context comparisons. Peers having comparable data conditions such as LMS-based engagement prediction or MOOC dropout detection are the only ones benchmarked. There is a prevention of misinterpretation of small institutional datasets with large national research. Relative performance is better than rank in benchmarking. Generalizing across limits Dominance relationships across accuracy, error, latency, and complexity are evaluated to locate strong performers. It is a strategy that employs benchmarking as an

analytical diagnostic tool rather than competitive leaderboard to aid the educational stakeholders in making practical decisions.

Extended Results

Improved feature-fusion, feature-ensemble models, and optimized models regulate nonlinearity in educational data, class imbalance, and heterogeneity in features, which surpass a single architecture baseline. Attention-based and hybrid recurrent architectures are effective when it comes to longitudinal predictions that are dependent on a time range. There are negative consequences of performance gains. Models with high accuracy levels are not suitable in real time or low resource context due to high costs of training and inference. Minor loss of accuracy in the lightweight models makes them scalable and interpretable which are the main characteristics of lightweight models and makes them the ideal models to implement to the process by the institution and to apply the policy-driven interventions to the process. The trade-offs are planned and outcomes of architectural optimizations. Measures, lack of uniformity in the variety of datasets, and little information on how data are computed in certain publications constrain the review to different scenarios. Aggregate boosting due to high performer model publication bias can also be a boost. Nonetheless, despite these shortcomings, a systematic benchmark and statistical synthesis suggest the progresses and deficiencies in AI-based educational prediction in terms of process.

This research presents a systematic methodology of evaluating AI-based predictive models in education beyond comparing the narratives and listing metrics. A new idea is to view educational prediction models as system-level objects, and have interacting properties such as predictive accuracy, error behavior, computational cost, latency, interpretability and deployment feasibility. The paper directly models trade-offs between these properties to shift model comparison out of the leaderboard-style ranking to decision-oriented benchmarking. Our extension of state-of-the-art reviews is to normalize the existence of disparate measures to the same space of analytical procedures to a single multi-criteria benchmarking approach. Unlike previous studies on algorithm taxonomies or application domains, this survey studies the impact of architectural complexity, optimization mechanisms and data modalities on outcomes. Ensemble and hybrid deep models show higher mean accuracy (92–96%) than classical learners (85–90%) but higher inference delay and resource consumption, which earlier evaluations underestimated. A deep state-of-the-art comparison is possible by mapping reviewed publications across models, datasets, metrics, and outcomes. CNN–GRU hybrids, attention-enhanced transformers, and optimization-assisted ensembles are compared to tree-based and linear baselines using accuracy, RMSE, AUC, recall, and computational overhead. Reports reveal pros and cons, such as enhanced minority-class recall at the cost of training time or high interpretability with accuracy loss. This work covers the analytical void of earlier evaluations without decision-level guidance. The study identifies which models function best in early-warning systems, privacy-preserving situations, and low resource

institutions for the process. This view views the work as a methodological advance rather than a literature review for this process.

Quantitative Hypotheses

Numerous quantitative gaps rather than conceptual ones exist in the literature. First, while most studies provide accuracy, fewer than 35% include error-based metrics such RMSE or confidence intervals, limiting reliability assessment. Second, few models address deployment restrictions despite latency and computational complexity differences of over an order of magnitude. Three, interpretability is sometimes a benefit rather than a performance parameter in process. Clear, falsifiable research hypotheses arise from these gaps. Hybrid and ensemble-based models may predict accuracy better than single-learner baselines on comparable educational datasets. Another theory is that accuracy drops while delay and instability increase beyond modest complexity. A third hypothesis is that organized interpretability strategies reduce dataset variance and keep models competitive. Synthesis of quantitative results validates hypotheses. Ensemble and hybrid models exhibit higher inference time variance and 4–6% higher accuracy than classical techniques. The second hypothesis is validated by correlation and pooling analysis indicating diminishing outcomes beyond high-capacity systems. Interpretability-enhanced models yield strong results with lower confidence ranges. Instead of remaining a feature of descriptive critique, gap analysis in the study is transformed into evidence-based analysis by formalizing gaps as testable hypotheses and establishing them in a statistically comparative manner, instead of a narrative one. The empirical baselines are well defined enabling future research to employ this approach.

Protocol for Benchmarking and Reproducibility

This study defines benchmarking as repeatable, analytical approach that has normalization, comparability, and task alignment rule. The methodology permits the results of reported performance to be meaningfully compared rather than retraining all models on a single dataset, which is not possible due to data access limitations. These are equal test-set measurements, prediction task alignment, and performance indicator standardizations. Classification, regression and multimodal inference are prediction tasks in education. Job classification metrics are compared instead of comparing regression RMSE to classification accuracy. Cross-study synthesis is not hard because units, semantic meaning and normalization requirements each require metrics have cross-studies. Normalisation maintains relative differences whereas limited scaling sums up within each measure category. The context of the dataset (including sample size, class imbalance and time range) is well documented along with the performance values to reduce the misunderstanding of the results based on different data regimes. Composite analysis lacks a contextual description but qualitatively depicts research. This protocol renders benchmarking as adaptive and auditable in its process. By following the same task classification, metric definitions and normalization criteria, researchers can replicate

the analytical logic of new studies, and institutions can modify the framework to their data without breaking the standards of comparability.

Statistical Rigor, Robustness, and Uncertainty

The analytical model is statistically sound incorporating significance test, sensitivity analysis and estimation of uncertainty. The research on the correlation of performance metric reveals that there exists a strong negative relationship between the error metrics variables and accuracy and positive relationship between the complexity proxy variables and inference delay. In order to remove the random variations in the trend, statistical significance is determined by conventionally acceptable levels of confidence. A sensitivity study evaluates conclusions under different metric weightings and normalizations. Results complement analytical findings by showing that model family dominance relationships stay consistent despite rank variation. Further reliability study shows cross-validation models have lower variance than single-split models. To reduce uncertainty, we assess dispersion across similar experiments and synthesize confidence intervals and variance measurements. Heterogeneity study suggests that data quality and context matter more than algorithmic instability in dropout and mental-health prediction than grade predictions. We admit publishing bias toward high-performing models and uneven dataset distribution by area. The study avoids overgeneralization and empirically limits its conclusions by assessing variability and uncertainty.

Benchmarking Framework System-Level Engineering

Modular benchmarking can be employed in institutional or policy settings. Data ingestion, task classification, metric normalization, multi-criteria evaluation, and reporting are included in process. Each module is independent but uses standard interfaces for scaling and adaptions. Input studies or institutional models are classified by task type and data modality, metrics validated, standardized, and contextualized, multi-criteria evaluation computes composite and dominance-based indicators, and outcomes visualized and documented for decision-making. This approach supports retrospective literature review and prospective model evaluations. Assessment and implementation facilitate actual practice. Local trained models allow institutions to track the readiness of deployment and policymakers to compare the intervention methods across regions based on aggregated outcomes. The framework allows the re-examination of the periodicity and longitudinal monitoring of system performance as new data/ models become accessible in the process. It is a system-based view of benchmarking making it not only a process but also academic to deployable analytical infrastructure of operations and governance.

Discussion, Interpretation, and Implications

The data shows that there is no model that is dominant in every educative prediction scenario. Complex and high-dimensional datasets are better served by hybrid and ensemble models, which are more accurate and recall more, but the simpler ones are easier to understand. These trends agree but measure when every trade-off is vital, which

is not estimated earlier. The measures of researchers should contain uncertainty and cost of computation in order to be meaningfully compared. The model allows the practitioners to select models due to the institutional constraints rather than the technical performance. The article cautions the policymakers against Machiavellian style systems without taking into account sustainability, justice and interpretability in the process. Performance differences are explained using data, architectural design, and techniques of evaluation in addition to descriptive repetition of outcomes. The research advances AI-based education analytics in both practice and theory by incorporating interpretation in a statistically based and systems-conscious framework sets.

The relative comparison of predictive modeling methods in educational analytics indicates some significant trends in terms of the performance and application of the models. The ensemble learning techniques like boosting, bagging and stacking are well known techniques that tend to record greater predictive accuracy since this involves a combination of more than one base learner, reducing overfitting and enhancing the generalization of the models. Conversely, deep learning nets like CNN, LSTM, and GRU are better suited to tasks that have access to large-scale or multimodal data since the latter model requires access to a substantial amount of training data to ensure successful acquisition of complex features. To implement smaller or structured educational datasets, more conventional machine learning methods, e.g. decision trees, support vector machines, and logistic regression, can be competitively used based on their lower computational cost and fewer data needs. The size of the data set, similarity of features, and time are some of the characteristics of data set that are very important in identifying the most appropriate predictive model. The implications of these findings are significant to educational analytics: predictive models can be used by educators and institutional administrators to identify at-risk students, track engagement and assist in early intervention, researchers, and educational data scientists can create more robust and interpretable models depending on a particular educational environment. In general, it is important to note that the connection between model architecture, data properties, and prediction tasks is critical to the creation of effective and feasible AI-based educational decision-support systems.

SUMMARY AND CONCLUSION

This paper has provided an in-depth analysis of recent developments in predictive modeling tools employed in educational systems between 2021 and 2024. It reviewed a broad range of AI-based methodologies, including deep neural networks, metaheuristic optimization strategies, ensemble learning schemes, interpretable AI frameworks, and specialized applications such as emotion recognition and dropout prediction. The comparative analysis reveals several key insights. Ensemble learning techniques consistently demonstrate strong predictive performance across diverse educational datasets due to their ability to combine multiple classifiers and mitigate overfitting. Deep learning architectures—particularly CNNs, LSTMs, and GRUs exhibit promising results

when applied to large-scale, multimodal educational data, especially in behavioral and sentiment analysis. However, the efficacy of these deep learning models remains largely contingent upon data volume and computational resources.

The other notable finding is that most of the studies use accuracy as the primary assessment tool and a lesser number of studies use comprehensive performance measures like RMSE, precision, recall, computational time, and complexity of the model. This underscores the importance of benchmarking systems that can be used to compare predictive models in educational analytics regularly. In general, the present review makes contributions to the literature in the form of the multidimensional benchmarking perspective that assesses the predictive models in terms of both quantitative performance measures and qualitative ones, including interpretability, scalability, and applicability to real-life applications.

The previous literature mainly discussed prediction models with their small scope; hence laying a focal point on a single target such as forecasting exam scores or dropout predictions in binary nature. So, very few previous reviews show some concern for model generalization, neglecting clear inputs such as demographic, psychological, behavioral, or certain cognitive signals of their study conditions (stellar in the context of mixed and interpreted scenarios as far as the signing of the observations will still be blind into mental disorder by anyone from the outsets. This would be elegantly solved by ML with its novel give-here, summarization, and updated recognition projects). Most of the literature surveys consider only simple metrics and concurrency-based, conversion strategies. Smoothening large-scale generalization among the very distant use cases would complement the interpretability of the models. Most textbook methods, therefore, may not have an accuracy that all stakeholders in the education system may comprehend accurately for the process.

This work can be considered as one of the most analytical reviews in predictive educational modeling. It has made major contributions, Multidimensional Benchmarking: A numerical study on the 40 current studies was done in depth, across six core performance dimensions (RMSE, accuracy, precision, recall, complexity, and delay). This will enable the stakeholders to determine the accuracy of the model as well as its applicability. Added Visualization and Diagnostics: On a variety of more advanced visualizations like KDE plots, heatmaps, pie charts and radar diagrams, the model performance trends and trade-offs are shown to refine the design options. Some of the algorithms of the education AI Sets include Map Model Diversity and innovation CNN-GRU, Bi-LSTM, CSSA, PSO, stacking, hybrid voting, LIME, SHAP and Federated Learning. Holistic Education Impact: The project investigated suicide risks prediction, regional dropout prediction, and sentimental computing with FER and multilingual sentiment modeling which extends the concept of performance to include well-being and engagements.

Some of the research and implementation directions suggested by this review include: Unified Multimodal Frameworks: Future models of shared process architecture models

need input mechanisms such as brain activity, clickstream behavior, emotional cues, and social interaction metrics. Attention-based fusion module transformers have the capability to generate such integration very effectively in process. Real-Time and Edge-AI Implementation: The overwhelming majority of the surveyed models are computationally intensive and are not edge-gadget-deployment-ready (like classroom tablets) and real-time teaching. Nevertheless, lightweight designs like A-DDF and GRU-based distillation models signify bright perspectives. Causal and Explainable AI: Although the tendency to develop explainability tools like SHAP and LIME is apparently trending, future systems should be based on causal inference systems, which explain not only what but also how and why this or that happens. Interpretable causal graphs and counterfactual analysis can be used to enhance educational decision-making process. Globalization and Localization: It is also possible that one of the biggest gaps in research is the models that are not available in contexts that are not adequately represented in education (i.e. underrepresented rural, multilingual, resource-constrained environments). The need, therefore, is to have models that are trained with global datasets and tested to be used in the fields of fairness and generalization.

The future systems should engage every stakeholder in the training and feedback process (students, teachers, counselors), but the predictions should be not only accurate, but also aligned with society. Some ethical AI models and fairness-sensitive modelling must be integrated into design pipelines. Longitudinal Learning Models: The current system is mostly fixed or exists in a course-based isolated environment. There is an immediate necessity of models which are able to learn across varying semesters or at least across academic years using methods like continual learning and domain adaptations.

Additionally, the review constitutes a diagnostic instrument to be used in future research to strategize educational plans, and create intelligent system development that is capable of transforming education through the application of personalized, predictive, and preventive technology.

AUTHOR CONTRIBUTIONS

Conceptualization, V.V.K.R. and Y.V.K.D.B.; Methodology, V.V.K.R; Validation, S.S., and V.V.K.R.; Investigation, B.B.; Resources, V.V.K.R.; Data Curation, V.V.K.R.; Writing – Original Draft Preparation, K.B.B.; Writing – Review & Editing, S.S.; Visualization, B.B.; Supervision, F.K., and P.N.; Project Administration, V.V.K.R., and S.S.

CONFLICT OF INTERESTS

The authors confirm that there is no conflict of interest associated with this publication.

REFERENCES

1. Kamakshamma, V., Bharati, K.F. Chronological squirrel search algorithm enabled deep recurrent neural network for employability prediction. *Knowl Inf Syst.* **2025**, *67*(3), 7669 – 7698.
2. Liu, H., Chen, X., Zhao, F. Learning behavior feature fused deep learning network model for MOOC dropout prediction. *Educ Inf Technol*. **2024**, *29*(3), 3257–3278.
3. Junejo, N.U.R., Huang, Q., Dong, X., et al. SAPPNet: students' academic performance prediction during COVID-19 using neural network. *Sci Rep*. **2024**, *14*, 24605.
4. Atas, P.K. Evaluate student achievement by classifying brain structure and its functionality with novel hybrid method. *Neural Comput Appl.* **2024**, *36*(7), 3357–3368.
5. Dagani, J., Buizza, C., Ferrari, C, et al. Potential suicide risk among the college student population: machine learning approaches for identifying predictors and different students' risk profiles. *Psicol. Refl. Crit.* **2024**, 37, 19.
6. Ren, Y., Yu, X. Long-term student performance prediction using learning ability self-adaptive algorithm. *Complex Intell. Syst.* **2024**, *10*, 6379–6408.
7. Masood, S.W., Gogoi, M., Begum, S.A. Optimised SMOTE-based imbalanced learning for student dropout prediction. *Arab J Sci Eng.* **2025**, **50**, 7165–7179.
8. Muthuselvan, S., Rajaprakash, S., Jaichandran R., et al. Student academic performance prediction enhancement using t-SIDSBO and triple voter network. *Multimed Tools Appl.* **2024**, 83, 82223–82246.
9. Sharma, S.K. An instructional emperor pigeon optimization (IEPO) based DeepEnrollNet for university student enrolment prediction and retention recommendation. *Sci Rep.* **2024**, *14*, 30830.
10. Li, T. Prediction and optimization of student grades based on genetic algorithm and graph convolutional neural networks. *Int J Comput Intell Syst.* **2025**, *18*, 59.
11. Tirumanadham, N.S.K.M.K., S, T. & M, S. Improving predictive performance in e-learning through hybrid 2-tier feature selection and hyper parameter-optimized 3-tier ensemble modeling. *Int. j. inf. tecnol.* **2024**, *16*, 5429–5456.
12. Mahawar, K., Rattan, P. Empowering education: harnessing ensemble machine learning approach and ACO-DT classifier for early student academic performance prediction. *Educ Inf Technol.* **2025**, *30*(4), 4639–4667.
13. Althaqafi, T., Saleem, F., Al-Ghamdi, A.A.M. Enhancing student performance prediction: the role of class imbalance handling in machine learning models. *Discov Comput.* **2025**, *28*, 79.
14. Shariff, V., Paritala, C., Ankala, K.M. Federated tree-based ensembles with SHAP explainability and integrated feature selection for secure lung cancer health analytics. *Interdiscip J Inf Knowl Manag*. **2025**, *20*, 026.
15. Khairy, D., Alharbi, N., Amasha, M.A., et al. Prediction of student exam performance using data mining classification algorithms. *Educ Inf Technol.* **2024**, *29*, 21621–21645.
16. NarendraBabu, C.R., Harsha, S., Shaikh, T.S. LightGBM: next point of interest location prediction using ensemble machine learning. *SN Comput Sci.* **2023**, *4*, 764.
17. Jain, A., Dubey, A.K., Khan, S., et al. A PSO weighted ensemble framework with SMOTE balancing for student dropout prediction in smart education systems. *Sci Rep.* **2025**, *15*, 17463.

18. Bhaskaran, S.S. Prediction of optimum student performance factors using genetic algorithm. *Multimed Tools Appl.* **2025**, 84, 20757–20778.
19. Lin, C., Kuo, F.C., Chau, T, et al. Artificial intelligence-enabled electrocardiography contributes to hyperthyroidism detection and outcome prediction. *Commun Med.* **2024**, *4*, 42.
20. Sharif, M., Buyrukoglu, S., Akbas, A. Student adaptivity classification in online education through stacked ensemble learning. *Multimed Tools Appl.* **2025**, *84*, 31119–31138.
21. Ali, S.R.M., Sundravadivelu, K., Muthukumar, S., et al. Enhancing student performance for low rank students using adaptive deep multi-perception pattern learning classification model. *SN Comput Sci.* **2025**, *6*, 55.
22. Oz, E., Bulut, O., Cellat, Z.F., et al. Stacking: an ensemble learning approach to predict student performance in PISA 2022. *Educ Inf Technol.* **2025**, *30*, 7753–7779.
23. Alnasyan, B., Basheri, M., Alassafi, M., et al. Kanformer: an attention-enhanced deep learning model for predicting student performance in virtual learning environments. *Soc Netw Anal Min.* **2025**, *15*, 25.
24. Albahli, S. Efficient hyperparameter tuning for predicting student performance with Bayesian optimization. *Multimed Tools Appl.* **2024**, *83*, 52711–52735.
25. Bhimavarapu, U. Analysing student performance for online education using computational models. *Univ Access Inf Soc.* **2024**, *23*, 1051–1058.
26. Ghosh, P., Charit A, Banerjee H, et al. DropWrap: a neural network based automated model for managing student dropout. *Int J Netw Distrib Comput.* **2025**, *13*, 17.
27. Sanfo, J.B.M. Application of explainable artificial intelligence approach to predict student learning outcomes. *J Comput Soc Sci.* **2025**, *8*, 9.
28. Junejo, N.U.R., Nawaz, M.W., Huang, Q., et al. Accurate multi-category student performance forecasting at early stages of online education using neural networks. *Sci Rep.* **2025**, 15, 16251.
29. Salem, M., Shaalan, K. Unlocking the power of machine learning in e-learning: a comprehensive review of predictive models for student performance and engagement. *Educ Inf Technol.* **2025**, *30*(13), 19027-19050.
30. Zerkouk, M., Mihoubi, M., Chikhaoui, B., et al. A machine learning based model for student dropout prediction in online training. *Educ Inf Technol.* **2024**, *29*, 15793–15812.
31. Cheng, X., Chaw, J.K., Sahrani, S, et al. An adaptive dual distillation framework for efficient remaining useful life prediction. *Complex Intell Syst.* **2025**, 11, 253.
32. Lentini, M.R., Thayasivam, U. Value-at-risk student prescription trees for price personalization. *J Big Data.* **2025**, 12, 70.
33. Thiruthuvanathan, M.M., Krishnan, B. Multitask EfficientNet affective computing for student engagement detection. *Multimed Tools Appl.* **2025**, *84*, 19039–19063.
34. Venkatesan, R.G., Karmegam, D., Mappillairaju, B. Exploring statistical approaches for predicting student dropout in education: a systematic review and meta-analysis. *J Comput Soc Sci.* **2024**, *7*,171–196.
35. Mandia, S., Mitharwal, R., Singh, K. Automatic student engagement measurement using machine learning techniques: a literature study of data and methods. *Multimed Tools Appl.* **2024**, *83*, 49641–49672.
36. Li, G., Liu, Z. Masked feature regeneration based asymmetric student–teacher network for anomaly detection. *Multimed Tools Appl.* **2024**, *83*, 90573–90594.

37. Ahuja, R., Sharma, S.C. Student opinion mining about instructor using optimized ensemble machine learning model and feature fusion. *SN Comput Sci.* **2024**, 5, 672.
38. Mao, Z., Chen, X., Wu, C. Reinforced distillation learning: fine-grained imbalanced classifier for financial crisis prediction. *Comput Econ.* **2025**.
39. Sihare, S.R. Student dropout analysis in higher education and retention by artificial intelligence and machine learning. *SN Comput Sci.* **2024**, 5, 202.
40. Mandava, R., & Sravanthi, G.L. Quantum Machine Learning Algorithms for Optimizing Complex Data Classification Tasks. *Journal of Transactions in Systems Engineering*, **2026**, *4*(1), 538–559.