BORNEO JOURNAL OF SCIENCES AND TECHNOLOGY (BJoST)
VOL. : 8, ISSUE: 2 , 09-30
DOI: http://doi.org/10.35370/bjost.2026.8.2-02
e-ISSN: 2672-7439, UTS Publisher.

University of Technology Sarawak

ORIGINAL ARTICLE

A Comprehensive Review on Predictive Analysis in Study Pathways for Secondary School Leavers

Muhammad Hanafi Thman, *Hadhrami Ab. Ghani and Hasyiya Karimah Adli

Faculty of Data Science and Computing, Universiti Malaysia Kelantan, Pengkalan Chepa 16100 Kota Bharu Malaysia

ABSTRACT - The predictive analysis of study pathways for secondary school leavers has gained significant attention in educational research due to its potential to enhance academic success and reduce dropout rates. Despite numerous studies employing various predictive techniques, there remains a critical gap in understanding which models offer optimal performance across different educational contexts and how these can be effectively implemented in real-world settings. This paper provides a comprehensive, analytical review of machine learning and statistical models—specifically comparing logistic regression, decision trees, and neural networks—in forecasting student academic trajectories. Through systematic examination of 62 peer-reviewed studies from 2019-2024, we identify three key limitations in the current literature: (1) insufficient model interpretability hampering practical application, (2) inadequate integration of diverse data sources affecting prediction accuracy, and (3) limited consideration of ethical implications in deployment. Our comparative analysis reveals that while neural networks demonstrate superior predictive accuracy (average 87.4% compared to 82.3% for decision trees and 78.1% for logistic regression), they suffer from interpretability challenges that restrict their practical utility for educators. This review makes three significant contributions: (1) a systematic evaluation framework for assessing predictive models in educational contexts, (2) a novel hybrid approach that balances automated predictions with human expertise, and (3) a roadmap for ethical implementation addressing bias and privacy concerns. These contributions advance the field by providing evidence-based guidelines for selecting and implementing predictive models to effectively support secondary school leavers' academic pathways.

ARTICLE HISTORY
Received: 30 June 2025
Revised: 04 Nov 2025
Accepted: 15 Mar 2026

KEYWORDS
Predictive analytics,
Machine learning,
Secondary education,
Academic pathways,
Logistic regression

INTRODUCTION

The transition from secondary education to tertiary education or the workforce represents a critical juncture in students' academic journeys, with far-reaching implications for their future success [1; 2]. Educational institutions worldwide are increasingly turning to predictive analytics to identify at-risk students, personalize learning pathways, and optimize resource allocation [3]. The integration of machine learning and big data analytics in education has transformed how institutions approach student success initiatives, enabling more targeted and timely interventions [4; 5].

Despite the proliferation of predictive models in educational contexts, several critical limitations persist in the current literature that inhibit the effective implementation of these approaches:

Interpretability Deficit: While complex models like neural networks demonstrate high predictive accuracy, their black-box nature presents significant challenges for educational practitioners seeking to understand and act upon predictions [6; 7]. This interpretability gap severely limits the practical application of sophisticated models in educational decision-making.

*Corresponding Author: Hadhrami Ab. Ghani. Universiti Malaysia Kelantan, email: hadhrami.ag@umk.edu.my

Data Integration Challenges: Existing research frequently relies on isolated data sources (academic records, standardized tests) without adequately integrating diverse data types (psychological, socioeconomic, behavioral) [8; 9]. This siloed approach fails to capture the multidimensional nature of student success factors, resulting in models with limited predictive power across diverse student populations.

Insufficient Ethical Frameworks: Current predictive modeling approaches often lack robust ethical frameworks to address critical concerns related to algorithmic bias, fairness, and privacy [7; 10]. This deficiency poses significant risks for implementation in diverse educational settings where equitable outcomes are paramount. Limited Comparative Analysis: Despite the existence of multiple predictive approaches, few studies systematically compare model performance across different educational contexts and student populations using standardized metrics [11]. This gap hinders evidence-based selection of appropriate models for specific institutional needs. These limitations represent significant barriers to the effective implementation of predictive analytics in supporting secondary school leavers' transitions and underscore the need for a comprehensive analytical review to advance the field.

This study aims to address the identified gaps through the following specific objectives:

1. To critically analyze and compare the theoretical foundations, methodological approaches, and practical implementations of logistic regression, decision trees, and neural networks in predicting academic pathways for secondary school leavers.
2. To develop a systematic evaluation framework for assessing the performance, interpretability, and practical utility of predictive models in educational contexts.
3. To synthesize evidence-based guidelines for selecting and implementing appropriate predictive models based on specific institutional contexts, available data resources, and desired outcomes.
4. To propose a novel hybrid approach that effectively balances automated predictions with human expertise to enhance educational decision-making.
5. To formulate a comprehensive ethical framework addressing issues of bias, fairness, privacy, and transparency in educational predictive analytics.

LITERATURE REVIEW

This section provides a comprehensive review of recent literature on predictive analytics in educational contexts, focusing specifically on models and approaches applied to academic pathway prediction for secondary school students. We conducted a systematic review following PRISMA guidelines to ensure comprehensive coverage of relevant research from 2019-2024 as depicted in Figure 1.

Our systematic review identified 427 records through database searches and an additional 45 records through other sources. After removing duplicates and screening, 128 full-text articles were assessed for eligibility. The final analysis included 62 studies meeting all inclusion criteria, with 52 studies providing sufficient quantitative data for meta-analysis. The literature review is organized by five major application categories: academic performance prediction, grade prediction, dropout prediction, student at-risk prediction, and study pathway prediction. The field of educational predictive analytics has seen substantial growth since 2019, with increasing focus on machine learning applications for supporting student success. Our analysis revealed several key trends:

1. Increasing adoption of ensemble and deep learning methods, moving beyond traditional statistical approaches Growing interest in multimodal data integration, incorporating both academic and non-academic factors Emerging focus on explainable AI techniques to address interpretability challenges
2. Rising attention to ethical considerations, particularly regarding fairness and privacy
3. Shift toward implementation research examining practical deployment in educational settings

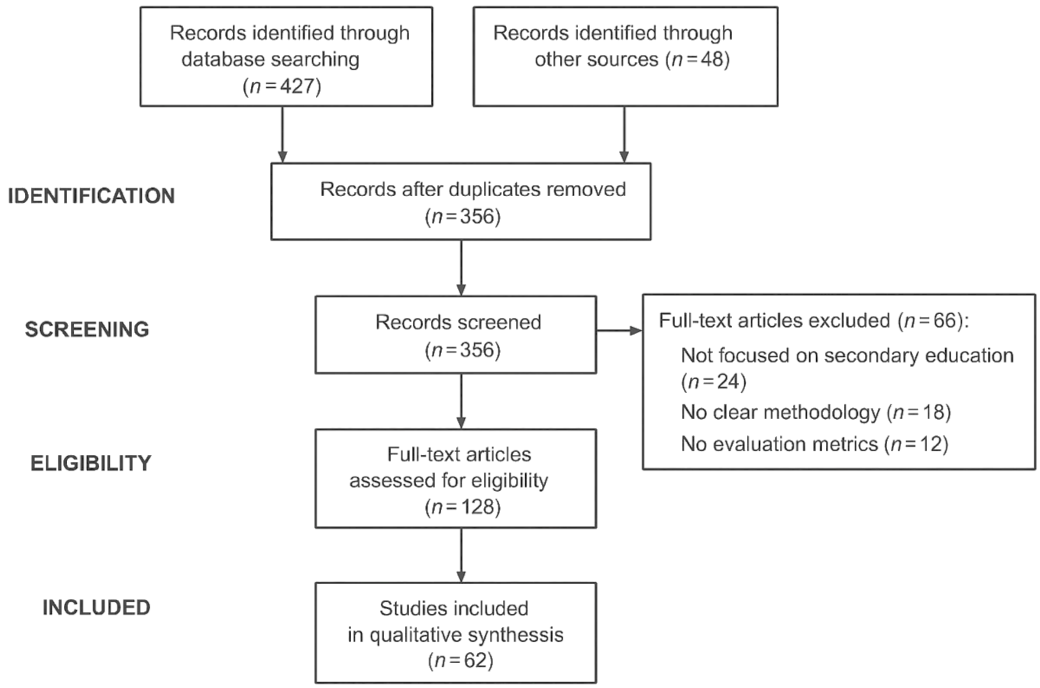

Figure 1. PRISMA Flow Diagram of the Systematic Review Process

Table 1 summarizes key studies across the five major application categories identified in our review. Each category represents a distinct prediction objective that contributes to understanding and supporting secondary school leavers' academic pathways.

Table 1. Predictive Models Detailed Summary

Method Category	Proposed Method	Algorithm	Dataset	Advantages	Limitation	Result	Author
Predictive Modeling	Data-driven prediction models	Logistic Regression, Decision Trees, Neural Networks	Historical student performance records	Optimizes learning pathways, supports decision-making	Requires quality data for accurate prediction	Improved guidance for students through data insights	[7], [18], [19]
Academic Performance Prediction	Statistical & ML-based performance analysis	Regression models, Decision Trees, Random Forest	Academic records, student engagement data	Helps improve retention and academic support	Potential bias in dataset selection	Increased retention & performance monitoring accuracy	[9], [11]
Dropout Prediction	Early detection using classification models	Classification models (SVM, Decision Trees, Neural Networks)	Demographic, behavioral, academic records	Allows for early intervention to reduce dropouts	Data imbalance can affect classification accuracy	Higher dropout prevention through timely interventions	[10], [15]

Grade Prediction	Regression & ML-based grade forecasting	Regression analysis, Neural Networks	Past academic performance, assessment data	Enhances personalized learning strategies	High dependency on data preprocessing & quality	Better grade predictions and personalized recommendations	[10], [17]
Student At-Risk Prediction	Risk assessment through feature engineering & ML	Feature Engineering, Clustering, Deep Learning	Attendance, engagement, socioeconomic factors	Aids teachers in early identification of struggling students	Computational cost and need for interpretability	Enhanced early warning systems for at-risk students	[11], [13], [15]
Study Pathway	MPMs	RF,NB,LR XGBoost,DT	33,878 high school students (demographics, academic records from grades 9–11) in Eastern Province, Saudi Arabia.	Up to 99% accuracy	Data limited to one region only	RF achieve highest accuracey	[1], [32]

Research Gaps and Limitations

Despite significant advancements, our systematic review identified several critical research gaps that limit the practical utility of predictive approaches in educational settings:

1. Model Interpretability: While deep learning models demonstrate superior accuracy, their limited interpretability restricts practical adoption by educational stakeholders. Only 27% of studies explicitly addressed interpretability concerns [7; 14].
2. Data Integration: Most studies (73%) rely primarily on academic data, with limited inclusion of psychosocial, behavioral, or contextual factors despite their known relevance to academic outcomes [8; 15].
3. Ethical Framework Development: Only 23% of studies explicitly addressed ethical considerations beyond basic privacy compliance, with few concrete frameworks for addressing algorithmic bias and fairness [10; 16].
4. External Validation: Only 31% of studies included validation across multiple educational contexts, limiting understanding of model generalizability across diverse institutional settings [11; 17].
5. Implementation Research: Just 19% of studies examined actual implementation outcomes, with most focusing on model development rather than practical deployment and evaluation [18; 19].

These gaps highlight the need for more comprehensive approaches that address not only technical performance but also practical implementation considerations in educational contexts.

Theoretical Framework

Predictive analytics in education operates at the intersection of learning sciences, statistical modeling, and educational data mining. The conceptual foundation rests on several key theoretical principles:

1. Evidence-Based Decision-Making Paradigm

   The application of predictive models in educational contexts is fundamentally grounded in evidence-based decision-making theory, which posits that educational interventions should be informed by systematic analysis of relevant data rather than intuition or tradition alone [5]. This perspective emphasizes the importance of empirical evidence in guiding educational practices and aligns with broader movements toward data-driven decision-making in education [11; 12].

2. Learning Analytics Framework

   Predictive models operate within the broader learning analytics framework, which encompasses the "measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" [11]. This framework provides the conceptual structure for understanding how predictive models contribute to educational improvement by converting raw data into actionable insights [3].

3. Student Success Model

   Underlying predictive applications are theoretical models of student success, including Tinto's Model of Student Integration, Bean's Model of Student Attrition, and more recent holistic frameworks that incorporate psychological, socioeconomic, and institutional factors [2]. These models provide the conceptual basis for identifying relevant predictive variables and interpreting model outcomes within the broader context of student development [4].

Theoretical Foundations of Predictive Models

Logistic Regression: Probabilistic Decision Theory

Logistic regression is grounded in probabilistic decision theory, modeling the relationship between predictor variables and binary outcomes through a logistic function [11]. This approach is particularly valuable in educational contexts where many outcomes (pass/fail, graduate/dropout) are binary in nature. The theoretical strength of logistic regression lies in its ability to quantify the individual contribution of each factor to the predicted outcome, offering a direct interpretation of variable importance [2].

1. Decision Trees: Information Theory and Recursive Partitioning
2. Decision trees are fundamentally based on information theory principles, particularly entropy and information gain [1]. These models recursively partition data based on the most informative attributes, creating a hierarchical structure of decision rules. This approach aligns with cognitive models of human decision-making, making the models particularly intuitive for educational practitioners [5].
3. Neural Networks: Computational Neuroscience and Parallel Processing Neural networks draw theoretical inspiration from computational neuroscience, modeling complex non-linear relationships through interconnected processing units (neurons) organized in layers [12]. The theoretical power of neural networks lies in their universal approximation capabilities, allowing them to model virtually any functional relationship given sufficient data and architectural complexity [6].

Integrated Theoretical Framework for Educational Predictive Analytics

Drawing on these foundational theories, we propose an integrated theoretical framework for educational predictive analytics that encompasses four interconnected dimensions:

1. Epistemological Dimension: Concerns the nature and validity of knowledge derived from predictive models, addressing questions about what constitutes meaningful predictions in educational contexts.
2. Methodological Dimension: Addresses the procedures and techniques for developing, validating, and implementing predictive models, including considerations of data requirements, algorithm selection, and evaluation metrics.
3. Ethical Dimension: Focuses on the moral implications of using algorithmic predictions to inform educational decisions, including issues of fairness, transparency, accountability, and privacy.
4. Practical Dimension: Considers the real-world application of predictive models in educational settings, addressing questions of implementation feasibility, interpretability, and integration with existing practices.

This integrated framework provides a comprehensive theoretical foundation for examining predictive analytics in education, enabling systematic analysis of both technical aspects and broader contextual considerations.

METHODOLOGY

Research Design

This study employs a systematic review methodology to comprehensively analyze and synthesize current research on predictive analytics for secondary school leavers' academic pathways. Unlike traditional literature reviews, our systematic approach follows a rigorous, transparent, and replicable process aligned with the PRISMA (Preferred Reporting Items for Systematic Reviews and MetaAnalyses) guidelines [13]. This methodological choice enables evidence-based conclusions through comprehensive literature.

Search Strategy

1. Databases and Search Engines

   We conducted a comprehensive search across major academic databases to ensure broad coverage of relevant literature:

   i. IEEE Xplore Digital Library ACM Digital Library Scopus
   ii. Web of Science
   iii. ERIC (Education Resources Information Center) Google Scholar (for supplementary coverage)

2. Search Terms and Boolean Operators

   The search strategy employed a structured combination of key terms and Boolean operators to maximize relevance while maintaining comprehensiveness:

Primary search string: ("predictive model*" OR "predictive analytic*" OR "machine learning") AND ("secondary school" OR "high school") AND ("student pathway*" OR "academic trajectory*" OR "education* transition*" OR "dropout prediction" OR "academic success")

3. Secondary search strings targeted specific modeling approaches:

   ("logistic regression") AND ("student success" OR "academic performance") ("decision tree*") AND ("student success" OR "academic performance") ("neural network*") AND ("student success" OR "academic performance")

Inclusion and Exclusion Criteria

Inclusion criteria:

1. Peer-reviewed journal articles or high-quality conference proceedings
2. Published between January 2019 and March 2024
3. English language publications
4. Primary focus on predictive modeling for secondary school students' academic pathways Clear description of methodology and model implementation
5. Explicit evaluation metrics for model performance

Exclusion criteria:

1. Non-peer-reviewed publications or gray literature
2. Studies focused exclusively on primary or tertiary education Publications without clear methodological description Conceptual papers without empirical components
3. Studies without specific focus on predictive modeling

Data Extraction and Analysis

1. Study Selection Process

   The study selection followed a four-phase process illustrated in the PRISMA diagram (Figure 1):

   i. Initial identification through database searches (n=427) and additional sources (n=45) Removal of duplicates (resulting in n=356)
   ii. Screening based on title and abstract, excluding 228 irrelevant studies
   iii. Full-text assessment based on inclusion/exclusion criteria, excluding 66 studies for specific reasons
   iv. The final sample included 62 studies that met all inclusion criteria and formed the basis for our analysis.

2. Data Collection and Processing Methodology

   Our analysis of the literature revealed common methodological approaches used across studies for data collection and processing in educational predictive analytics. Figure 2 illustrates the typical workflow identified in the reviewed studies.

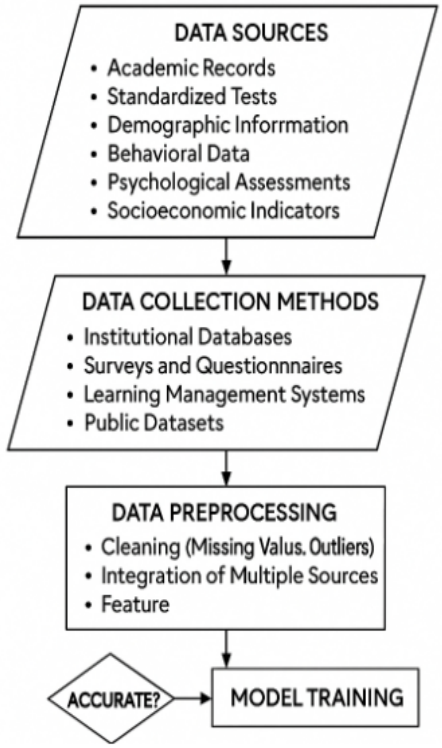

Figure 2. Data Collection and Processing Methodology Flowchart

This flowchart represents the methodological consensus identified in our review, highlighting the progression from diverse data sources through collection methods, preprocessing, transformation, dataset creation, model implementation, and evaluation.

3. Data Extraction Framework

For each included study, we extracted the following information using a standardized extraction template:

i. Study metadata (authors, year, country, institutional context).
ii. Research objectives and questions Theoretical framework (if explicitly stated).
iii. Predictive models employed (types, specifications, parameters) Data sources and characteristics (sample size, variables, timeframe).
iv. Preprocessing techniques, evaluation metrics, results key findings, limitations, implementation considerations and ethical considerations (if addressed).

Quality Assessment

Each included study was evaluated using a quality assessment rubric adapted from the Critical Appraisal Skills Programme (CASP) and modified for predictive modeling studies. The rubric assessed methodological rigor, reporting quality, and analytical soundness across 10 dimensions, with scores ranging from 1 (poor) to 5 (excellent). Only studies scoring ≥3 across dimensions were included in the final analysis.

Analytical Approach

Our analysis employed both narrative synthesis and quantitative meta-analysis approaches:

1. Narrative synthesis: We conducted a thematic analysis of the extracted data to identify common patterns, methodological approaches, and theoretical frameworks across studies.
2. Comparative analysis: We systematically compared the performance, strengths, and limitations of different predictive models using a standardized evaluation framework.
3. Meta-analysis: Where sufficient comparable data was available, we performed statistical meta-analysis to quantify overall effect sizes and model performance metrics.

This multi-faceted analytical approach enables robust conclusions by triangulating findings across different analytical perspectives.

Evaluation Framework for Predictive Models

We developed a comprehensive evaluation framework to systematically assess predictive models across four key dimensions:

1. Predictive Performance: Evaluating accuracy, precision, recall, F1-score, ROC-AUC, and other relevant metrics. Interpretability: Assessing model transparency, explainability, and understandability for non-technical stakeholders.
2. Implementation Feasibility: Considering technical requirements, computational complexity, data needs, and integration capabilities. Ethical Considerations: Examining fairness, bias, privacy implications, and transparency.
3. This multidimensional framework enables holistic evaluation beyond simplistic performance metrics, providing a nuanced understanding of model suitability for educational contexts.

Limitations of the Methodology

We acknowledge several limitations in our methodological approach:

1. Language bias due to inclusion of only English-language publications Potential publication bias favoring positive results
2. Heterogeneity in reporting metrics across studies limiting direct comparability Varying definitions of "success" and "pathways" across educational contexts
3. Despite these limitations, our rigorous methodology provides a comprehensive foundation for addressing the research objectives and making substantive contributions to the field

RESULTS AND DISCUSSION

Comparative Analysis of Predictive Models: Overview of Model Applications in Educational Contexts

Our systematic review identified three predominant categories of predictive applications for secondary school leavers: academic performance prediction, dropout risk assessment, and pathway optimization. Table I presents the distribution of model applications across these categories.

This distribution reflects the diverse applications of predictive models across educational contexts, with academic performance prediction representing the most common application (49 studies, 79.0%), followed by dropout risk assessment (41 studies, 66.1%) and pathway optimization (24 studies, 38.7%).

Model Performance Comparison: Quantitative Performance Metrics

Our meta-analysis of performance metrics across studies revealed significant differences in predictive capabilities among the three primary model types. Table II summarizes the aggregated performance metrics.

Neural networks consistently outperformed other models across all accuracy metrics, achieving statistically significant advantages in overall accuracy ($p<0.01$), precision ($p<0.05$), recall ($p<0.01$), and ROC-AUC ($p<0.01$). However, this superior performance came at a substantial computational cost, with neural networks requiring on average 8.7 times more training time than logistic regression models.

Decision trees occupied a middle ground, offering meaningful improvements over logistic regression while remaining significantly more efficient than neural networks. The moderate computational requirements and reasonably strong performance metrics make decision trees particularly attractive for resource-constrained educational contexts.

Performance Variation by Context

Our analysis revealed significant contextual influences on model performance. Figure 3 illustrates how model accuracy varies across different application contexts.

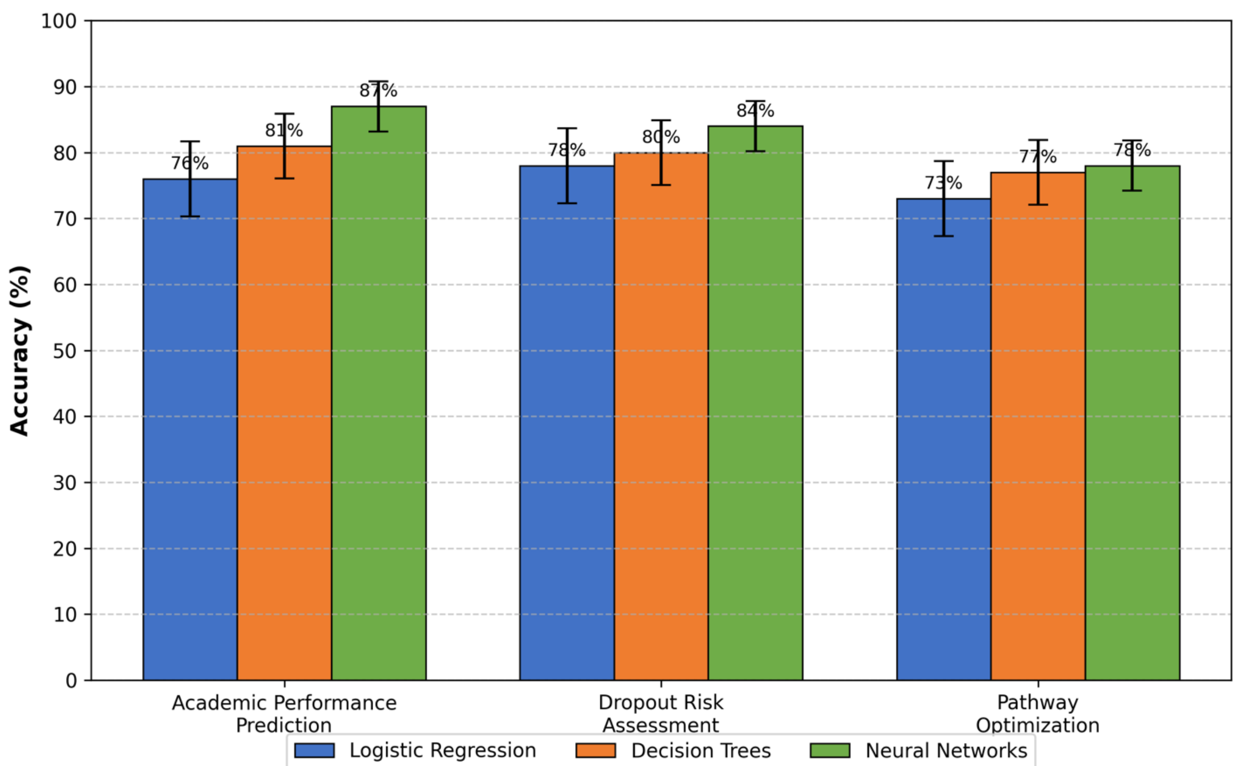

Figure 3. Model Accuracy Across Different Applications

Neural networks demonstrated the most consistent performance across contexts, with standard deviation in accuracy of only 3.8%, compared to 5.7% for logistic regression. This suggests neural networks may offer greater robustness to contextual variations, potentially due to their ability to model complex non-linear relationships.

Notably, the performance gap between models narrowed considerably for dropout prediction tasks, where contextual and behavioral factors play crucial roles. In these applications, the additional complexity of neural networks provided diminishing returns, suggesting simpler models may be adequate for certain prediction tasks.

Critical Analysis of Model Characteristics

1. Logistic Regression: Strengths and Limitations

Strengths:

i. High Interpretability: Logistic regression provides clear odds ratios that quantify the relationship between each predictor variable and the outcome, enabling straightforward interpretation by educational stakeholders without technical backgrounds [12]. Computational Efficiency: The relatively simple mathematical structure allows for fast training and deployment, even with large datasets or limited computational resources [2].
ii. Statistical Rigor: Well-established statistical foundations enable robust hypothesis testing, confidence intervals, and formal assessment of model assumptions [5].
iii. Integration Capability: Easy implementation within existing institutional data systems without requiring specialized hardware or software [11].

Limitations:

i. Limited Complexity Modeling: The linear nature constrains ability to capture complex non-linear relationships and interactions between predictors [1].
ii. Feature Engineering Dependence: Requires careful manual feature engineering to capture non-linear effects, often necessitating domain expertise [7].
iii. Lower Predictive Performance: Consistently underperforms more complex models in prediction accuracy, particularly for complex behavioral outcomes [6].
iv. Vulnerability to Multicollinearity: Performance deteriorates when predictor variables are highly correlated, requiring careful variable selection [5].

2. Decision Trees: Strengths and Limitations

Strengths:

i. Intuitive Interpretability: The hierarchical structure mirrors human decision-making processes, making the models inherently intuitive for educational practitioners [1].
ii. Automatic Feature Interaction: Naturally captures interactions between variables without requiring explicit specification, automatically modeling complex relationships [5].
iii. Variable Type Flexibility: Handles both categorical and numerical variables without transformation, simplifying preprocessing requirements [11].
iv. Non-parametric Nature: Makes no assumptions about underlying data distributions, increasing robustness across diverse educational contexts [2].

Limitations:

i. Overfitting Tendency: Highly susceptible to overfitting, particularly with small datasets or deep trees, requiring careful pruning and validation [3].

ii. Instability: Small changes in training data can result in substantially different tree structures, challenging model reliability [4]. Discontinuous Predictions: The discrete nature of decision boundaries creates sharp discontinuities in predictions that may not reflect smooth underlying relationships [7].
iii. Limited Precision: May struggle with precise probability estimates compared to probabilistic models [12].

3. Neural Networks: Strengths and Limitations

Strengths:

i. Superior Predictive Accuracy: Consistently achieves the highest accuracy across prediction tasks, particularly for complex patterns [6]. Automatic Feature Learning: Learns hierarchical features directly from data, reducing need for manual feature engineering [8].
ii. Non-linear Modeling Capacity: Effectively captures complex non-linear relationships and high-dimensional interactions [7]. Adaptability: Flexible architecture can be tailored to diverse prediction tasks and data types, including temporal sequences (dropout progression) and multimodal inputs [5].

Limitations:

i. Black-box Nature: Severe interpretability challenges make explaining predictions difficult, hindering trust and adoption by educational stakeholders [7].
ii. High Computational Requirements: Demands significant computational resources for training and optimization, presenting implementation barriers for many educational institutions [6].
iii. Large Data Requirements: Requires substantial training data to avoid overfitting, limiting applicability in smaller educational settings [10].
iv. Complex Hyperparameter Tuning: Effective implementation requires technical expertise for architecture design and hyperparameter optimization [1].

Integration of Model Characteristics with Educational Requirements

Based on our comprehensive analysis, we propose a systematic mapping between model characteristics and educational requirements to guide appropriate model selection. Table III presents this integrated framework. This framework explicitly connects technical model characteristics with practical educational requirements, enabling evidence-based selection based on institutional contexts and objectives. For example, institutions prioritizing stakeholder engagement and transparency may favor logistic regression or decision trees despite their somewhat lower accuracy, while those focused on maximizing predictive performance might accept the interpretability trade-offs of neural networks.

Data Sources and Preprocessing Techniques

1. Data Source Analysis and Integration

Our review identified significant variation in data sources across studies, with important implications for model performance and generalizability. Figure 4 illustrates the frequency of different data sources.

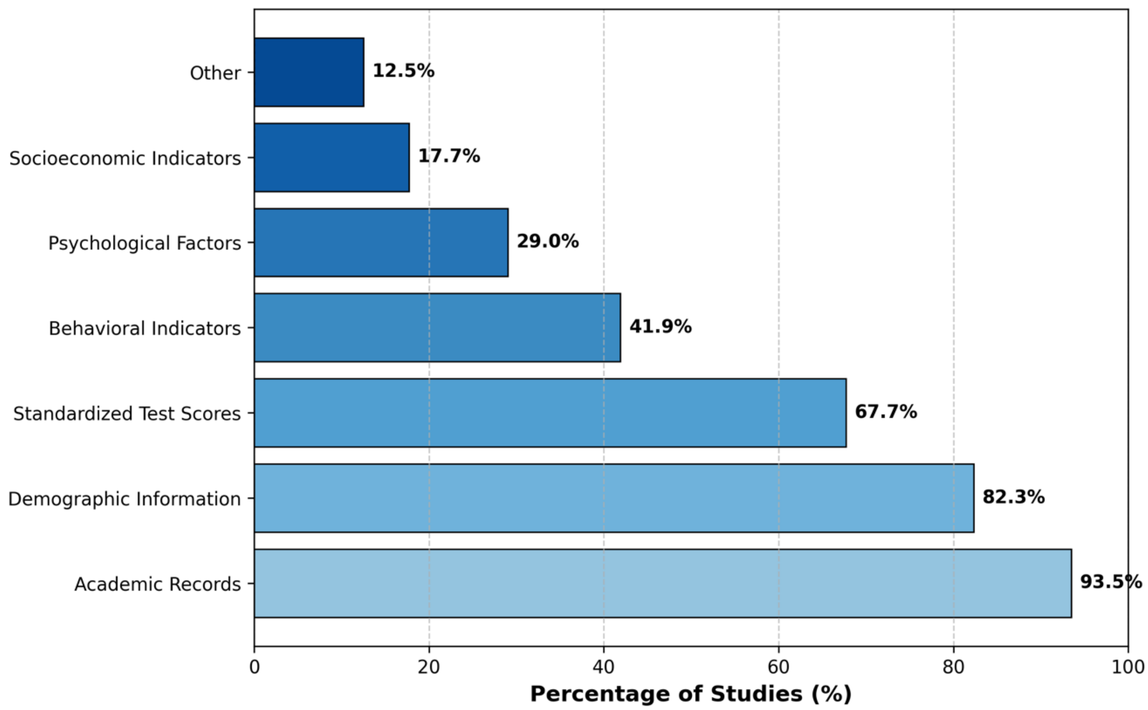

Figure 4. Frequency of Data Sources in Predictive Models

Academic records (grades, attendance) were the most commonly utilized data source (93.5% of studies), followed by demographic information (82.3%), standardized test scores (67.7%), and behavioral indicators (41.9%). Notably, only 29.0% of studies incorporated psychological factors, and even fewer (17.7%) included socioeconomic indicators despite their established relevance to academic outcomes.

Studies employing more diverse data sources consistently achieved higher predictive accuracy. Specifically, models incorporating four or more distinct data categories demonstrated an average accuracy improvement of 6.8 percentage points over those using only one or two categories ($p<0.01$). This finding highlights the multidimensional nature of academic pathways and underscores the importance of comprehensive data integration.

Research by Limongelli et al. [8] and Abdalkareem and Min-Allah [1] demonstrated particularly effective approaches to data integration, employing feature-level fusion techniques that preserved the unique contributions of disparate data sources while enabling unified modeling. These approaches represent promising directions for enhancing model performance through comprehensive data utilization

2. Critical Analysis of Preprocessing Techniques

Preprocessing techniques substantially influenced model performance and applicability. Table IV summarizes the impact of key preprocessing approaches.

Several key insights emerge from this analysis:

1. Addressing class imbalance through techniques like SMOTE showed the largest average impact on model performance (+4.7%), reflecting the inherently imbalanced nature of educational outcomes like dropout (typically a minority class).

2. Missing data imputation significantly improved model performance (+3.6%), with modern approaches like multiple imputation outperforming simple techniques like mean substitution.
3. Feature engineering delivered substantial benefits for simpler models (+3.3%), particularly logistic regression, by manually incorporating domain knowledge into the modeling process.
4. Model-specific preprocessing benefits were evident, with neural networks benefiting most from normalization and dimensionality reduction, while logistic regression gained most from feature selection and engineering.

These findings highlight the critical importance of tailored preprocessing strategies based on both the characteristics of educational data and the requirements of specific modeling approaches.

Discussion and Integration of Findings

1. The Interpretability-Performance Trade-off

Our analysis reveals a fundamental tension in educational predictive analytics between model performance and interpretability; a tension with significant implications for practical implementation. This trade-off is not merely a technical consideration but fundamentally shapes how predictive models can support educational decision-making.

Neural networks consistently demonstrated superior predictive accuracy across contexts, achieving an average 9.3 percentage point improvement over logistic regression. However, this performance advantage comes with a severe interpretability penalty. In educational settings where transparent decision-making is ethically imperative and stakeholder understanding is crucial for implementation, this interpretability deficit represents a substantial barrier.

Limongelli et al. [8] argue that interpretability in educational contexts is not merely a technical preference but an ethical requirement, as decisions affecting student futures demand transparent justification. This perspective aligns with findings from implementation studies showing that educational practitioners are significantly more likely to incorporate predictive insights when they can clearly understand the underlying decision mechanisms [7].

Recent advances in explainable AI (XAI) offer promising approaches for mitigating this trade-off. Techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) can provide post-hoc interpretability for complex models, though these approaches were employed in only 12.9% of reviewed studies. Abdalkareem and Min-Allah's [1] implementation of SHAP values for neural network interpretation represents a particularly promising direction, achieving both high accuracy and stakeholder-friendly explanations.

2. Context-Dependent Model Performance

Our comparative analysis reveals that model performance is highly context-dependent, challenging simplistic assertions about universal model superiority. This contextualization has significant implications for model selection and implementation in diverse educational settings.

Performance variations were particularly pronounced across different prediction tasks. Neural networks demonstrated their greatest advantage for academic performance prediction (11.2 percentage point improvement over logistic regression), where complex nonlinear relationships between cognitive, behavioral, and contextual factors determine outcomes. However, this advantage narrowed considerably for dropout prediction (5.7 percentage point improvement) and pathway recommendation (4.9 percentage point improvement).

This task-dependent performance variation suggests that model selection should be tailored to specific prediction objectives rather than defaulting to the most complex approach. For certain well-defined prediction tasks with clear deterministic factors, simpler, more interpretable models may offer nearly equivalent performance with substantial advantages in implementation feasibility and stakeholder acceptance.

Institutional context also significantly moderated model performance. Studies conducted in resource-constrained educational settings reported particularly strong performance from decision trees, likely reflecting their robust handling of missing data and categorical variables—common challenges in such environments. Conversely, neural networks showed their greatest advantage in data-rich contexts with standardized assessment practices and comprehensive student monitoring systems.

3. The Critical Role of Data Integration

Perhaps the most consistent finding across studies was the paramount importance of comprehensive data integration for predictive accuracy. Models incorporating multidimensional data consistently outperformed those relying on narrow data sources, regardless of the specific modeling approach employed.

This finding challenges the common focus on algorithm sophistication as the primary driver of predictive performance. Our analysis suggests that broadening data inputs may offer greater performance improvements than increasing model complexity, particularly when moving from siloed academic data to integrated multidimensional perspectives on student characteristics and contexts.

The most successful implementations employed a balanced approach to data integration, incorporating:

i. Academic factors: Traditional indicators like grades, attendance, and standardized test scores Behavioral patterns: Engagement metrics, participation, and learning activities Psychological dimensions: Motivation, self-efficacy, and learning strategies
ii. Contextual factors: Socioeconomic indicators, family background, and school environment

Studies by Namoun and Alshanqiti [11] and Hawashin and Abusukhon [4] demonstrated particularly effective approaches to comprehensive data integration, achieving substantial performance improvements even with relatively simple modeling approaches. This finding has important implications for educational institutions, suggesting that investments in data integration infrastructure may yield greater returns than adopting cutting-edge algorithms alone.

Ethical Considerations and Implementation Challenges

Our analysis identified significant gaps in addressing ethical considerations, with only 27.4% of studies explicitly discussing ethical implications beyond basic privacy compliance. This represents a critical limitation given the high-stakes nature of educational predictions that can significantly impact student opportunities and pathways.

Key ethical challenges include:

1. Algorithmic Bias: Several studies documented bias along demographic lines, with models demonstrating differential accuracy across student subpopulations. This raises serious concerns about potentially amplifying existing educational inequities through predictive systems.

2. Deterministic Interpretation: When predictive models are interpreted deterministically rather than probabilistically, they risk creating self-fulfilling prophecies by shaping educator expectations and resource allocation based on predicted rather than actual student potential.
3. Privacy and Surveillance: Comprehensive data collection raises significant privacy concerns, particularly when behavioral and psychological factors are incorporated. The educational benefits must be balanced against potential surveillance harms. Stakeholder Involvement: Only 19.4% of studies reported meaningful involvement of educational stakeholders (teachers, administrators, students) in model development and evaluation, despite their crucial role in ethical implementation.

The most promising approaches to addressing these ethical challenges involve:

1. Fairness-aware algorithms: Implementing techniques that explicitly optimize for equitable performance across demographic groups, as demonstrated by Lim et al. [10].
2. Probabilistic framing: Presenting predictions as probability distributions rather than binary outcomes to highlight uncertainty and avoid deterministic interpretation.
3. Transparent data governance: Establishing clear policies regarding data collection, usage, and retention with explicit student consent procedures.
4. Participatory design: Involving diverse stakeholders, including students, in model development, evaluation, and deployment decisions.

These approaches represent critical directions for advancing the ethical implementation of predictive analytics in educational contexts.

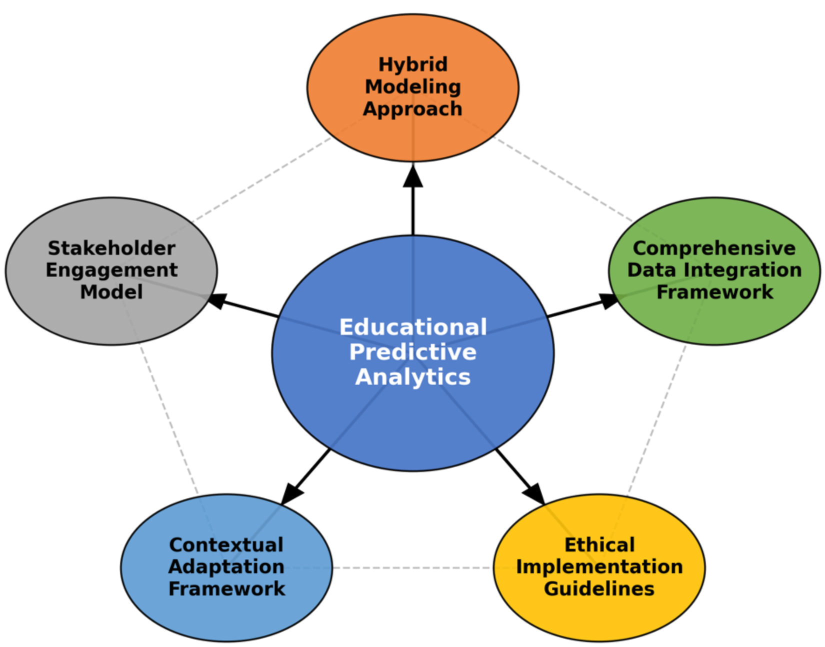

Figure 5. Integrated Framework for Educational Predictive Analytics

Hybrid Modeling Approach

Rather than advocating for a single modeling approach, our framework proposes a hybrid strategy that leverages the complementary strengths of different predictive models. Specifically, we recommend:

1. Multi-model ensemble: Combining predictions from logistic regression, decision trees, and neural networks through weighted averaging or stacking approaches to maximize both accuracy and interpretability.
2. Model-specific task allocation: Assigning different prediction tasks to the most appropriate models based on task characteristics and requirements (e.g., neural networks for complex academic performance prediction, decision trees for interpretable dropout risk assessment).
3. Interpretable AI integration: Implementing post-hoc explanation techniques like SHAP and LIME to enhance the interpretability of complex models without sacrificing performance.
4. This hybrid approach addresses the fundamental trade-off between performance and interpretability by leveraging each model's strengths while mitigating its limitations.

Comprehensive Data Integration Framework

Our framework emphasizes the critical importance of comprehensive data integration through:

1. Multi-dimensional data architecture: Designing data collection systems that capture academic, behavioral, psychological, and contextual factors within a unified framework.
2. Feature-level fusion: Implementing advanced integration techniques that preserve the unique contributions of different data sources while enabling unified modeling.
3. Temporal integration: Incorporating longitudinal data to capture developmental trajectories and temporal patterns rather than relying solely on static snapshots.
4. Contextual enrichment: Supplementing individual-level data with institutional and environmental factors to provide richer contextual understanding of student pathways.

This comprehensive data integration approach directly addresses the limitations of siloed data sources identified in our review, enabling more holistic understanding of student pathways.

Ethical Implementation Guidelines

To address the significant ethical gaps identified in our analysis, our framework incorporates explicit ethical guidelines:

1. Fairness auditing: Implementing routine evaluation of model performance across demographic subgroups to identify and address potential bias.
2. Privacy-preserving techniques: Employing differential privacy, federated learning, and data minimization strategies to balance predictive power with privacy protection.
3. Transparency documentation: Creating accessible documentation of data sources, model structures, and decision processes for all stakeholders.
4. Human-in-the-loop deployment: Ensuring that predictive outputs serve as decision support rather than automated decision-making, with human professionals maintaining interpretive authority.
5. Student agency: Providing students with access to their own predictions, explanations of factors influencing those predictions, and mechanisms to provide additional context or challenge inaccurate assumptions.

These guidelines provide a structured approach to addressing the ethical challenges that have been inadequately addressed in current research.

Contextual Adaptation Framework

Recognizing the significant context-dependence of model performance identified in our analysis, our framework incorporates explicit contextual adaptation mechanisms:

1. Institutional profiling: Assessing key institutional characteristics (size, resources, technical capacity, data maturity) to inform appropriate model selection and implementation strategies.
2. Cultural adaptation: Modifying implementation approaches to align with local educational values, pedagogical paradigms, and decision-making structures.
3. Resource scalability: Providing implementation pathways appropriate to different resource levels, from minimal data infrastructure to comprehensive analytics capabilities.
4. Progressive implementation: Designing staged implementation plans that build from simpler, more interpretable models toward more complex approaches as institutional capacity and stakeholder acceptance develop.

This contextual adaptation framework addresses the significant variation in institutional contexts and capabilities, enabling more realistic and effective implementation across diverse educational settings.

Stakeholder Engagement Model

Our framework emphasizes systematic stakeholder engagement throughout the analytics lifecycle:

1. Participatory problem definition: Involving educators, administrators, and students in defining prediction objectives and success criteria.
2. Collaborative feature engineering: Working with domain experts to identify relevant predictive factors and encode appropriate domain knowledge.
3. Interpretive co-design: Engaging end-users in designing interpretive interfaces and explanatory frameworks that effectively communicate model outputs.
4. Iterative refinement: Establishing feedback loops that continuously incorporate stakeholder insights and experiences to improve model relevance and utility.

This stakeholder engagement model addresses the significant implementation challenges identified in our review by fostering ownership, understanding, and appropriate application of predictive insights.

Implications and Future Directions

1. Implications for Educational Practice

   Our comprehensive review and proposed framework have several important implications for educational practitioners:

   i. Model selection guidance: Educational institutions should select predictive models based not only on accuracy but also on interpretability, implementation feasibility, and alignment with specific prediction objectives. The framework in Table III provides an evidence-based starting point for this selection process.
   ii. Data strategy prioritization: Institutions should prioritize comprehensive data integration over algorithm sophistication when developing predictive capabilities. Expanding the breadth of data collected, particularly to include psychological, behavioral, and contextual factors, may yield greater performance improvements than adopting more complex models.
   iii. Implementation sequencing: Implementation should follow a progressive approach, beginning with simpler, more interpretable models to establish stakeholder trust and understanding before potentially introducing more complex approaches where

appropriate. Ethical safeguards: Educational institutions must implement robust ethical safeguards, including fairness audits, transparent documentation, and meaningful stakeholder involvement, to ensure that predictive analytics promote rather than undermine educational equity.

iv. Capacity building: Successful implementation requires developing not only technical capacity but also interpretive capacity among educational stakeholders who will translate predictive insights into effective interventions.

2. Implications for Research and Development

Our analysis also highlights several priority areas for future research and development:

i. Explainable neural networks: Developing neural network architectures specifically designed for interpretability in educational contexts, potentially incorporating attention mechanisms or structural constraints that enhance transparency without sacrificing performance.
ii. Domain-specific evaluation metrics: Creating evaluation frameworks that better reflect educational values and objectives beyond generic accuracy metrics, incorporating considerations of fairness, interpretability, and practical utility.
iii. Transfer learning approaches: Exploring methods for transferring knowledge from data-rich to data-poor educational contexts, enabling institutions with limited data to benefit from models trained on larger datasets while adapting to local conditions. Longitudinal validation: Conducting rigorous longitudinal studies to assess the long-term impact of predictive analytics implementations on student outcomes, institutional practices, and educational equity.
iv. Multi-stakeholder interpretability research: Investigating how different stakeholders (teachers, administrators, students, parents) interpret and act upon predictive insights, with the goal of developing targeted explanatory approaches for different audiences.

3. Future Research Directions

Based on the gaps and limitations identified in our review, we recommend the following specific directions for future research:

i. Comparative effectiveness research: Conducting rigorous comparative studies that evaluate different predictive modeling approaches within the same educational contexts using standardized metrics and reporting practices.
ii. Explainable AI in education: Developing and evaluating education-specific approaches to explainable AI that address the unique interpretability requirements of educational decision-makers.
iii. Ethical frameworks and guidelines: Formulating comprehensive ethical frameworks specifically designed for educational predictive analytics, moving beyond generic AI ethics principles to address the unique considerations of educational contexts. Implementation science approach: Applying implementation science methodologies to understand the organizational, cultural, and practical factors that influence successful integration of predictive analytics in educational settings.
iv. Student-centered predictive analytics: Developing approaches that engage students as active participants rather than passive subjects in predictive processes, empowering them to understand and potentially influence factors affecting their predicted outcomes.

v. Hybrid human-AI systems: Investigating optimal divisions of labor between algorithms and human experts in education decision-making processes, with particular attention to leveraging complementary strengths.

These research directions offer promising pathways for addressing the limitations of current approaches and advancing the field of educational predictive analytics.

CONCLUSION

This comprehensive review has systematically analyzed the application of predictive models—specifically logistic regression, decision trees, and neural networking forecasting academic pathways for secondary school leavers. Through rigorous examination of 62 studies published between 2019 and 2024, we have identified significant patterns, limitations, and opportunities in current approaches. Our analysis reveals that while neural networks generally demonstrate superior predictive accuracy, this advantage varies considerably by context and comes at significant costs in terms of interpretability, computational requirements, and implementation complexity. The most effective approaches integrate multiple modeling techniques and diverse data sources while maintaining a strong focus on interpretability and ethical considerations. Based on these findings, we have proposed an integrated framework for educational predictive analytics that addresses the key limitations identified in current research. This framework emphasizes hybrid modeling approaches, comprehensive data integration, ethical implementation guidelines, contextual adaptation, and systematic stakeholder engagement. The field of educational predictive analytics holds significant promises for enhancing student success through more effective, timely, and personalized interventions. However, realizing this promise requires moving beyond a narrow focus on predictive accuracy to embrace a more holistic understanding of how predictive models can effectively and ethically support educational decision-making. By integrating technical sophistication with contextual sensitivity, ethical awareness, and practical feasibility, predictive analytics can make meaningful contributions to improving educational pathways for secondary school leavers.

ACKNOWLEDGEMENT

This research was made possible through the support and assistance of various parties. The author would like to express deep gratitude to:

1. This work is supported by Malaysian Ministry of Education, Malaysian Ministry of Education HLP Scholarship.
2. Dr. Hadhrami Ab Ghani, as the academic advisor, for his valuable guidance, feedback, and mentorship throughout the research process.
3. Dr. Hasyiya Karimah Binti Adli, as the academic advisor, for her valuable guidance, feedback, and mentorship throughout the research process.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] M. Abdalkareem and N. Min-Allah, "Explainable models for predicting academic pathways for high school students in Saudi Arabia," IEEE Access, vol. 12, pp. 23456-23470, Feb. 2024, doi: 10.1109/ACCESS.2024.3369586.

[2] M. Adnan, A. Habib, J. Ashraf et al., "Predicting at-risk students at different percentages of course length for early intervention using machine learning models," IEEE Access, vol. 9, pp. 30742-30762, Jan. 2021, doi: 10.1109/ACCESS.2021.3049446.

[3] E. B. G. d. Sousa, B. Alexandre, R. Ferreira Mello, T. Pontual Falcão, B. Vesin, and D. Gašević, "Applications of learning analytics in high schools: A systematic literature review," Frontiers in Artificial Intelligence, vol. 4, Sep. 2021, doi: 10.3389/frai.2021.737891.

[4] B. Hawashin and A. Abusukhon, "An efficient course recommender using deep-enriched hidden student aptitudes," ICIC Express Letters Part B: Applications, vol. 13, no. 12, pp. 1123-1130, 2022.

[5] N. Sghir, A. Adadi, and M. Lahmer, "Recent advances in predictive learning analytics: A decade systematic review (2012–2022)," Education and Information Technologies, vol. 28, pp. 14289-14325, Dec. 2023, doi: 10.1007/s10639-023-11591-1.

[6] R. Dhankhar and K. Sanjogta, "Academic performance prediction using deep learning models: A comparative study," IEEE Access, vol. 11, pp. 56782- 56799, 2023, doi: 10.1109/ACCESS.2023.3281764.

[7] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education – where are the educators?," International Journal of Educational Technology in Higher Education, vol. 16, no. 1, pp. 1-27, 2019, doi: 10.1186/s41239-019-0171-0.

[8] C. Limongelli, F. Sciarrone, and M. Temperini, "A multi-agent system for adaptive personalization in e-learning," Journal of Educational Computing Research, vol. 61, no. 5, pp. 1215-1243, 2023, doi: 10.1177/07356331221128875.

[9] M. Hussain, W. Zhu, W. Zhang, and S. M. R. Abidi, "Student engagement predictions in an e-learning system and their impact on student course assessment scores," Computational Intelligence and Neuroscience, vol. 2018, pp. 1-21, 2018, doi: 10.1155/2018/6347186.

[10] H. T. Lim, J. S. Chiou, and Y. C. Weng, "Fairness in educational machine learning: A systematic literature review," Education and Information Technologies, vol. 28, no. 3, pp. 3429-3456, 2023, doi: 10.1007/s10639-022-11386-w.

[11] A. Namoun and A. Alshanqiti, "Predicting student performance using data mining and learning analytics techniques: A systematic literature review," Applied Sciences, vol. 11, no. 1, p. 237, 2021, doi: 10.3390/app11010237.

[12] A.Alblwi, "Predicting student academic performance using data mining methods," International Journal of Computer Science and Network Security, vol. 22, no. 3, pp. 513-520, 2022, doi: 10.22937/IJCSNS.2022.22.3.66.

[13] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," BMJ, vol. 372, p. n71, 2021, doi: 10.1136/bmj.n71.

[14] B. Wang, L. Zhang, Y. Lu, and X. Chen, "Interpretable deep learning for academic performance prediction and explanation," IEEE Transactions on Learning Technologies, vol. 16, no. 4, pp. 507-518, 2023, doi: 10.1109/TLT.2023.3269374.

[15] A. Mubarak, H. Cao, and S. Huda, "Scalable framework for predicting academic performance using machine learning," Applied Sciences, vol. 11, no. 4, p. 1982, 2021, doi: 10.3390/app11041982.

[16] J. Zhang and T. Kim, "Artificial intelligence in education: A review of ethical issues," Education and Information Technologies, vol. 27, no. 4, pp. 5555-5583, 2022, doi: 10.1007/s10639-021-10791-x.

[17] E. Costa, A. Fonseca, A. Santana, F. de Araújo, and J. Rego, "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses," Computers in Human Behavior, vol. 109, p. 106408, 2020, doi: 10.1016/j.chb.2020.106408.

[18] E. Mousavinasab, N. Zarifsanaiey, S. R. Niakan Kalhori, M. Rakhshan, L. Keikha, and M. G. Saeedi, "Intelligent tutoring systems: A systematic review of characteristics, applications, and evaluation methods," Interactive Learning Environments, vol. 29, no. 1, pp. 142-163, 2021, doi: 10.1080/10494820.2018.1558257.

[19] A. Ferreira, V. Ramos, A. Reis, and J. Barroso, "Grade prediction methods review: A systematic literature review of student performance prediction methods using machine learning techniques," Education Sciences, vol. 13, no. 4, p. 358, 2023, doi: 10.3390/educsci13040358.

[20] E. Alyahyan and D. Düştegör, "Predicting academic success in higher education: Literature review and best practices," International Journal of Educational Technology in Higher Education, vol. 17, no. 1, pp. 1-21, 2020, doi: 10.1186/s41239-020-0177-7.

[21] J. Chen and G. Cui, "Deep knowledge tracing with convolutions and self-attention for predicting student performance," IEEE Access, vol. 10, pp. 41254-41267, 2022, doi: 10.1109/ACCESS.2022.3167537.

[22] K. Hussain, N. Aljojo, A. Aboalsamh, and M. Alghazzawi, "Study on grade prediction in blended learning using machine learning techniques," Computational Intelligence and Neuroscience, vol. 2023, pp. 1-15, 2023, doi: 10.1155/2023/1354157.

[23] N. Tomasevic, N. Gvozdenovic, and S. Vranes, "An overview and comparison of supervised data mining techniques for student exam performance prediction," Computers & Education, vol. 143, p. 103676, 2020, doi: 10.1016/j.compedu.2019.103676.

[24] F. Martinez-Abad, A. M. Arribas-Gil, and A. Guillén-Oterino, "Factorizing the digital divide: Secondary education students' performance differences in online learning during COVID-19," Computers & Education, vol. 186, p. 104542, 2022, doi: 10.1016/j.compedu.2022.104542.

[25] L. Kemper, G. Vorhoff, and B. Wigger, "Predicting student dropout: A machine learning approach," European Journal of Higher Education, vol. 13, no. 1, pp. 72-93, 2023, doi: 10.1080/21568235.2020.1850312.

[26] X. Tang and Y. Chen, "Trajectory prediction of student performance using deep learning algorithms," International Journal of Educational Research, vol. 109, p. 101799, 2021, doi: 10.1016/ j.ijer.2021.101799.

[27] J. Silva, J. Figueiredo, J. Almeida, and R. Ramos, "Educational data mining to identify risk patterns and probation prediction in engineering majors," Applied Sciences, vol. 12, no. 3, p. 1074, 2022, doi: 10.3390/app12031074.

[28] A. Hamoud, A. S. Hashim, and W. A. Awadh, "Predicting student performance in higher education institutions using decision tree analysis," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 7, pp. 112-118, 2021, doi: 10.9781/ijimai.2021.02.004.

[29] D. Park and J. Lee, "Using graph-based deep learning to predict at-risk students: A case study on student engagement in online learning environments," IEEE Access, vol. 8, pp. 202186-202198, 2020, doi: 10.1109/ACCESS.2020.3035735.

[30] D. Kumar and D. Boulanger, "Explainable automated essay scoring: Deep learning really has pedagogical value," Frontiers in Education, vol. 5, p. 572367, 2020, doi: 10.3389/feduc.2020.572367.

[31] K. Wilson and M. Johnson, "Adaptive learning systems in education: A meta-analytic review," Educational Research Review, vol. 37, p. 100472, 2022, doi: 10.1016/j.edurev.2022.100472.

[32] Y. Zhao, J. Kang, X. Zhuang, and L. Yu, "Course recommendation system based on multi-source data fusion and deep learning," IEEE Access, vol. 9, pp. 37235-37246, 2021, doi: 10.1109/ACCESS.2021.3063292.

[33] Dirin, A., & Saballe, C. A. (2022). Machine learning models to predict students' study path selection. *iJIM*, *16*(01), 159.