Predicting Student Academic Performance with Machine Learning: A Systematic Literature Review

M. Z. A. Chek[1], I. L. Ismail[2], Jamal. N[3], Z. H. Zulkifli[4], Rinda Nariswari[5], M. S. Asrulsani[6]

[1,6]Actuarial Science Department, UiTM Perak Branch

[2,3]Department of Statistics and Decision Science, UiTM Perak Branch

[4]Actuarial Partners Consulting, Malaysia

[5]Department of Computer Science, BINUS Indonesia

DOI: https://dx.doi.org/10.47772/IJRISS.2026.100300005

Received: 06 March 2026; Accepted: 11 March 2026; Published: 24 March 2026

ABSTRACT

Predicting student academic performance has become an essential research focus in higher education as institutions seek to improve retention rates, academic success, and educational quality. The increasing availability of educational datasets through student information systems and learning management systems provides opportunities for applying machine learning techniques to predict academic outcomes and identify at-risk students.

This study presents a systematic literature review (SLR) of machine learning approaches used for predicting student performance in higher education. The review follows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework to ensure transparency and replicability.

Peer-reviewed studies published between 2015 and 2025 were collected from major academic databases including Scopus, Web of Science, IEEE Xplore, ScienceDirect, SpringerLink, ACM Digital Library, and Google Scholar. The screening process resulted in a final selection of relevant studies examining predictive models in educational data mining and learning analytics.

The results indicate that Random Forest, Support Vector Machines (SVM), Decision Trees, Logistic Regression, and Artificial Neural Networks are the most frequently used algorithms for student performance prediction. Several studies demonstrate predictive accuracy ranging between 70% and 95%, indicating the effectiveness of machine learning models for identifying students at risk of academic failure.

The most influential predictive features include previous academic performance, attendance records, LMS engagement, assignment submissions, and demographic characteristics. The review also identifies several research gaps, including limited use of explainable artificial intelligence, insufficient cross-institution datasets, ethical concerns related to student data, and underutilization of deep learning methods.

The findings highlight the importance of integrating predictive analytics into educational decision-making systems and developing interpretable models that support early intervention strategies in higher education.

Keywords - Student Academic Performance Prediction, Machine Learning in Education, Educational Data Mining, Learning Analytics, SLR

INTRODUCTION

Higher education institutions worldwide face growing challenges related to student retention, academic achievement, and institutional performance. Increasing student dropout rates and academic failure have motivated universities to explore innovative strategies to improve learning outcomes and support student success. In recent years, the rapid digitalization of educational environments has generated large volumes of data through multiple institutional systems.

These systems include Student Information Systems (SIS), Learning Management Systems (LMS), online assessment platforms, and academic record databases. Student Information Systems typically store institutional data such as enrollment information, grades, and attendance records.

Learning Management Systems collect behavioral data related to students’ interactions with online learning environments, including login frequency, participation in discussion forums, and assignment submissions.

Online assessment platforms capture evaluation data such as quiz performance and examination results, while academic record databases maintain historical academic information. These datasets provide valuable insights into student learning behaviors, academic progress, and engagement patterns [1]–[5].

The emergence of Educational Data Mining (EDM) and Learning Analytics (LA) has enabled researchers and institutions to analyze educational data using advanced computational methods. Machine learning algorithms are particularly well suited for identifying hidden patterns and predicting academic outcomes based on historical data.

Machine learning techniques can analyze large and complex datasets to identify patterns that influence student success or failure. These predictive models support academic decision-making and enable institutions to develop early intervention programs.

Predictive analytics has therefore become a critical component of modern higher education systems, enabling institutions to detect students at risk of academic failure and implement targeted support strategies [6]–[10].

Despite technological advancements, many universities still struggle to identify students at risk of academic failure in a timely manner. Traditional statistical analysis methods are limited in their ability to capture complex relationships among multiple academic and behavioral variables.

Student performance is influenced by numerous factors including academic preparation, socioeconomic background, attendance patterns, engagement with online learning systems, and psychological or motivational factors.

Machine learning approaches offer significant advantages by modeling nonlinear relationships and extracting meaningful patterns from large educational datasets. Predictive analytics models can classify students into performance categories and support early detection of academic risks [11], [12].

However, existing research on student performance prediction remains fragmented across different methodologies, datasets, and evaluation metrics.

Therefore, a comprehensive systematic literature review is necessary to synthesize current research findings and identify emerging research directions. This systematic literature review aims to identify machine learning algorithms used to predict student academic performance in higher education.

The review also analyzes datasets and predictive features commonly used in prediction models. In addition, the study evaluates performance metrics applied to assess predictive accuracy and identifies existing research gaps and future research opportunities within the domain of educational data mining.

This study addresses four primary research questions. The first research question investigates which machine learning algorithms are most commonly used for predicting student academic performance. The second research question examines the types of datasets and predictive features used in prediction models.

The third research question analyzes the evaluation metrics used to measure model performance. The final research question identifies existing research gaps within the literature [3], [13]–[16].

BACKGROUND AND CONCEPTS

Educational Data Mining

Educational Data Mining (EDM) refers to the process of extracting meaningful information from educational datasets to support learning analytics and academic decision-making. EDM techniques allow researchers to analyze large volumes of educational data and identify patterns that influence academic outcomes.

EDM applications in higher education include predicting student academic performance, detecting students at risk of dropout, recommending courses or learning pathways, and analyzing student learning behaviors. These applications help institutions improve educational quality and student support services [17], [18].

Learning Analytics

Learning analytics focuses on analyzing data generated from digital learning environments to understand student learning behaviors and improve learning outcomes.

Data used in learning analytics commonly originates from Learning Management Systems and includes interaction logs, assignment submission records, quiz results, and discussion forum participation.

These behavioral indicators provide valuable insights into student engagement and learning patterns and can significantly influence predictive models of academic success [15].

Machine Learning in Education

Machine learning algorithms applied in educational data mining can generally be categorized into supervised learning and deep learning approaches. Supervised learning algorithms are the most commonly used methods for predicting student performance.

These algorithms learn from labeled datasets where the outcome variable, such as final grades or course completion status, is known. Common supervised algorithms used in educational prediction studies include Random Forest, Support Vector Machine, Decision Tree, Logistic Regression, Naïve Bayes, and Gradient Boosting.

Among these algorithms, Random Forest and Support Vector Machine models frequently outperform other methods due to their ability to handle complex datasets and nonlinear relationships. Deep learning models have also gained increasing attention in recent years.

Artificial Neural Networks and Long Short-Term Memory (LSTM) networks have demonstrated strong predictive capabilities for modeling complex patterns in educational datasets. Several studies report high predictive accuracy using deep neural network architectures.

Advanced LSTM-based models have achieved predictive accuracy exceeding 90% in some experimental settings [1].

RESEARCH METHODOLOGY

Systematic Literature Review Approach

This study adopts a systematic literature review methodology following the PRISMA 2020 framework to ensure a transparent and reproducible research process.

The PRISMA methodology includes four main stages: identification, screening, eligibility assessment, and final inclusion of relevant studies [6]–[8].

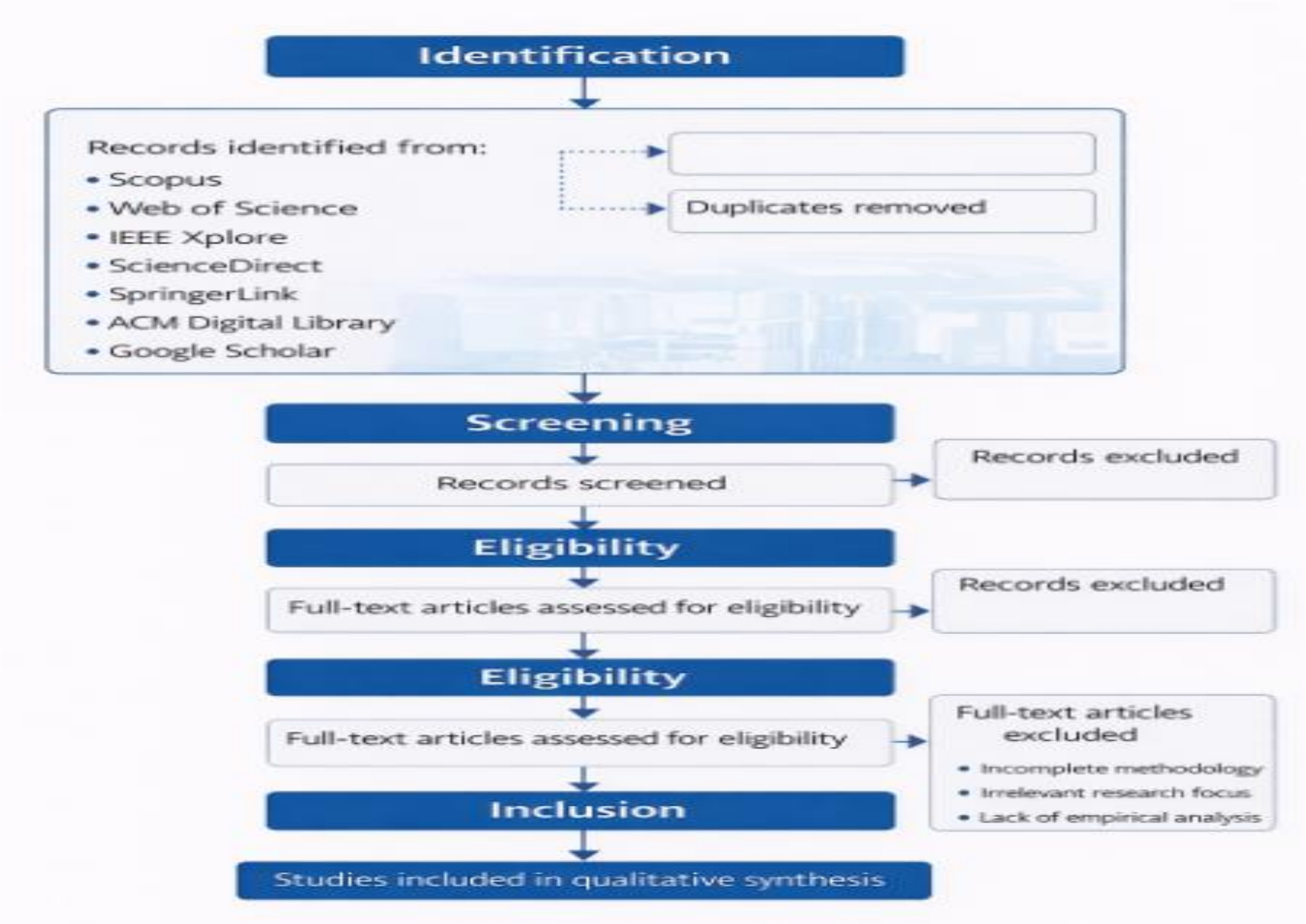

Fig 1 The PRISMA flow diagram illustrating the study selection process.

Figure I. PRISMA 2020 flow diagram illustrating the study selection process used in the systematic literature review. The process includes four stages: identification of records from multiple academic databases, screening of titles and abstracts, eligibility assessment through full-text review, and final inclusion of relevant studies for qualitative synthesis [2]–[5].

Literature Search Strategy

The literature search was conducted across multiple academic databases including Scopus, Web of Science, IEEE Xplore, ScienceDirect, SpringerLink, ACM Digital Library, and Google Scholar. These databases were selected because they contain high-quality peer-reviewed research publications in computer science, educational technology, and data mining.

The search strategy employed a combination of keywords related to student performance prediction and machine learning applications in education. The primary keywords used during the search process included student performance prediction, machine learning in education, educational data mining, learning analytics, and academic performance prediction.

PRISMA Screening Process

The study selection process followed the PRISMA guidelines to ensure systematic identification and evaluation of relevant research articles. During the identification stage, a total of 1,245 records were retrieved from the selected databases.

After removing 245 duplicate records, 1,000 unique studies remained for the screening stage. During the screening process, titles and abstracts were evaluated to determine the relevance of each study to the research objectives. As a result, 720 studies were excluded due to irrelevance to the topic, lack of machine learning methods, or focus on educational levels other than higher education.

Following the screening phase, 280 full-text articles were assessed for eligibility. These articles were evaluated based on predefined inclusion and exclusion criteria. During the eligibility assessment, 195 articles were excluded due to incomplete methodological descriptions, lack of empirical results, or non-peer-reviewed publication status.

Finally, 85 studies were included in the qualitative synthesis of the systematic literature review.

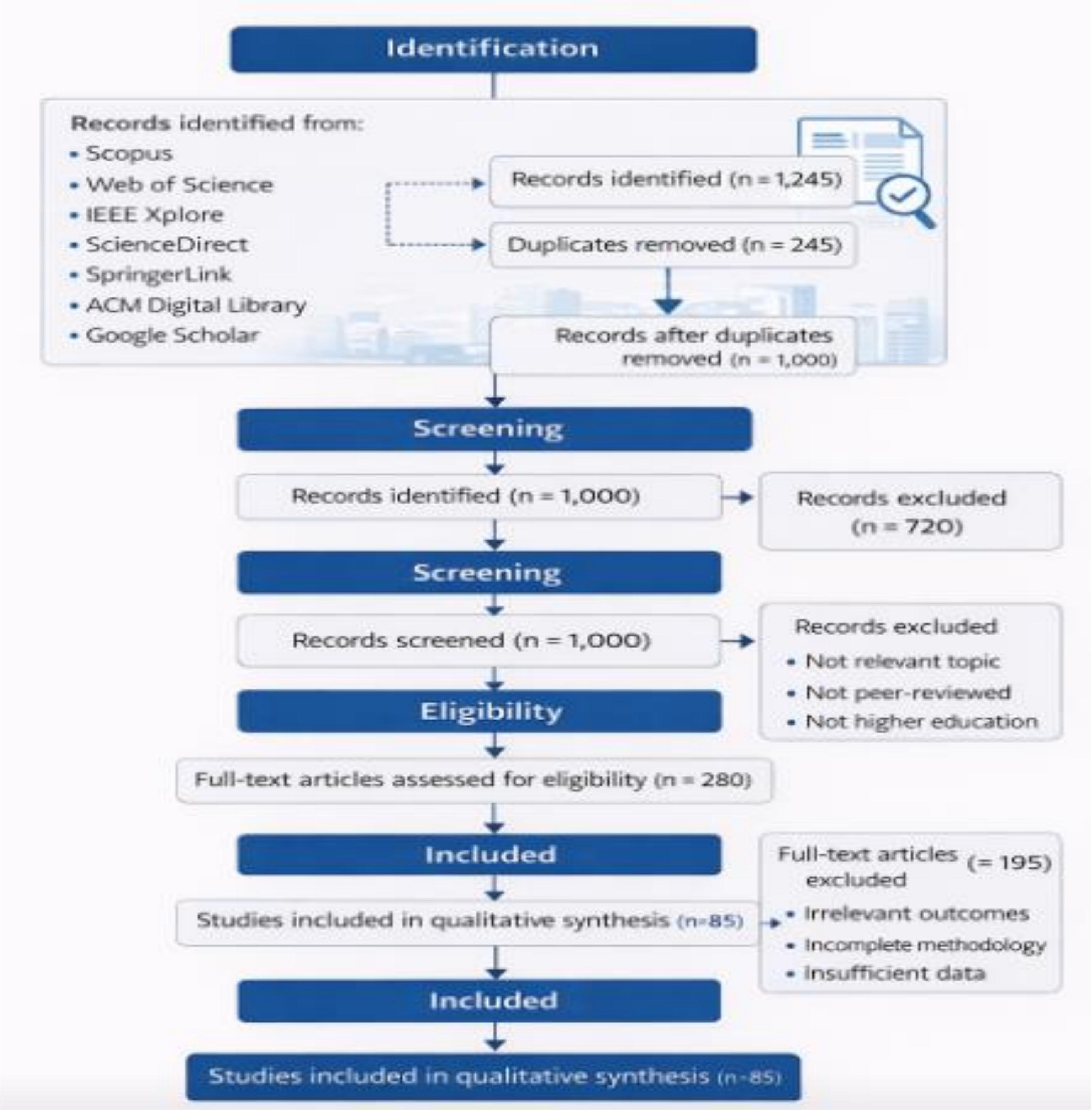
Identification
Records identified from:
• Scopus
• Web of Science
• IEEE Xplore
• ScienceDirect
• SpringerLink
• ACM Digital Library
• Google Scholar
Records identified (n = 1,245)
Duplicates removed (n = 245)
Records after duplicates removed (n = 1,000)
Screening
Records identified (n = 1,000)
Records excluded (n = 720)
Screening
Records screened (n = 1,000)
Records excluded
• Not relevant topic
• Not peer-reviewed
• Not higher education
Eligibility
Full-text articles assessed for eligibility (n = 280)
Included
Studies included in qualitative synthesis (n-85)
Full-text articles excluded (= 195)
• Irrelevant outcomes
• Incomplete methodology
• Insufficient data
Included
Studies included in qualitative synthesis (n-85)

Fig 2 PRISMA 2020 Flow Diagram for Study Selection

RESULTS AND ANALYSIS

Publication Trends

The analysis of selected studies indicates that research on machine learning techniques for student performance prediction has increased significantly since 2018. This trend reflects the growing interest in educational data mining and learning analytics as institutions increasingly adopt data-driven approaches for improving student outcomes [2], [9], [10], [16].

Machine Learning Algorithms Used

Random Forest algorithms consistently demonstrate strong performance across multiple datasets due to their ensemble learning structure [7].

TABLE 1 The Most Frequently Used Machine Learning Algorithms Identified in The Reviewed Studies.

Algorithm	Usage Frequency
Random Forest	High
Support Vector Machine	High
Decision Tree	High
Neural Networks	Moderate
Logistic Regression	Moderate
Naïve Bayes	Low

Data Sources Used

Educational datasets used for student performance prediction typically originate from multiple institutional data sources. These sources include Student Information Systems, Learning Management Systems, demographic datasets, and academic transcripts. Academic performance indicators and engagement metrics are strong predictors of student success [8].

Important Predictive Features

Several predictive features have been identified as influential variables in student performance prediction models. Previous academic grades represent one of the strongest predictors because they reflect students' prior learning achievements. Attendance rate is another important variable, as consistent class participation often correlates with higher academic success.

Learning Management System activity provides valuable behavioral data that reflects student engagement with course materials. Assignment submission patterns can also indicate students' study habits and commitment to coursework. In addition, socioeconomic background variables may influence academic performance through access to educational resources and support systems. Academic assessments and midterm scores have been shown to significantly influence final academic performance predictions [1].

Model Performance Evaluation

Predictive models in student performance research are typically evaluated using several performance metrics. These evaluation metrics include accuracy, precision, recall, F1-score, and ROC-AUC. Accuracy measures the proportion of correct predictions made by the model, while precision and recall evaluate the model's ability to correctly classify positive and negative cases.

F1-score represents the harmonic mean of precision and recall, and ROC-AUC measures the model's classification performance across different threshold levels. Several studies report predictive accuracy ranging between 70% and 95%, depending on the machine learning algorithm used and the characteristics of the dataset [2].

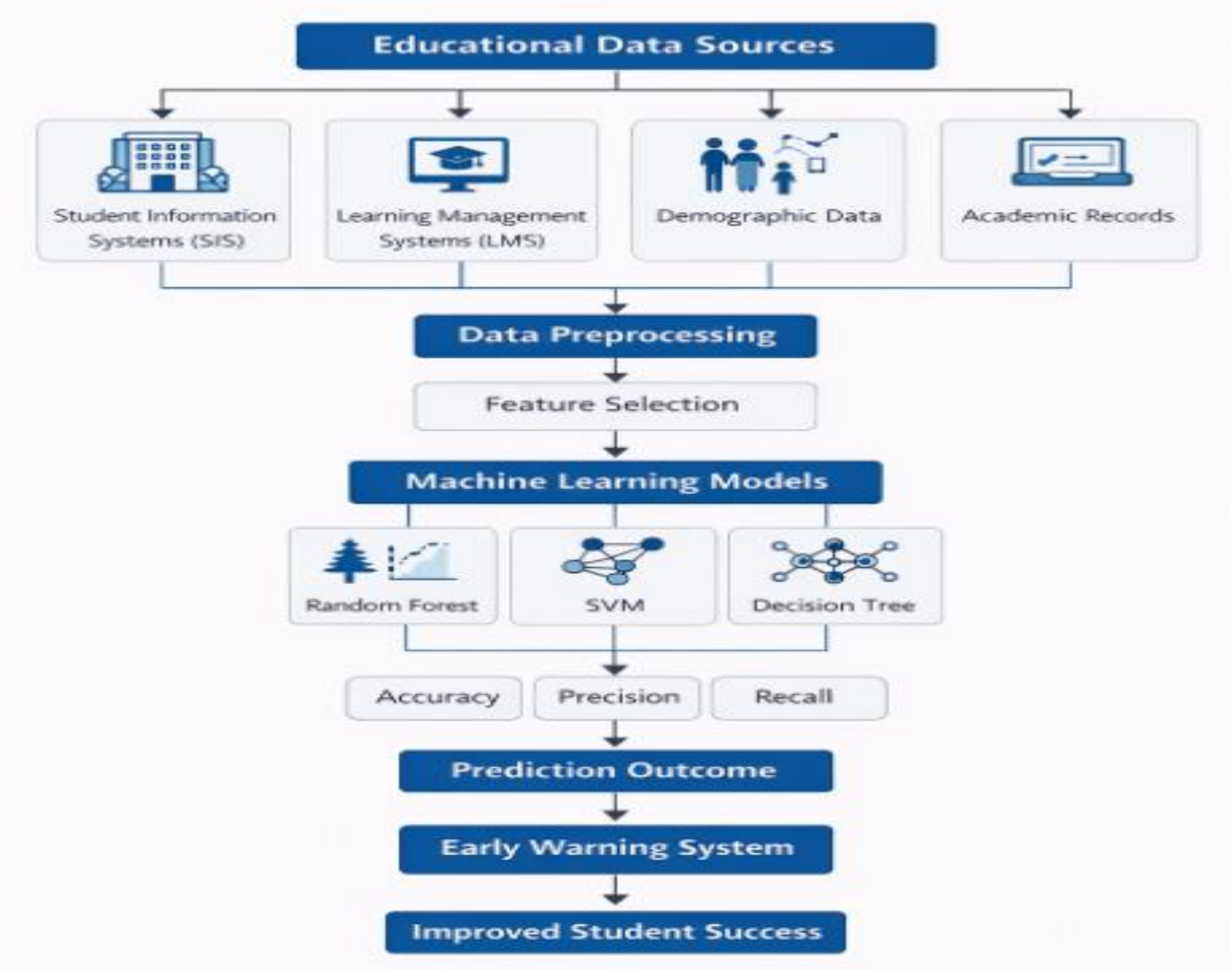

Fig 3. Conceptual Framework for Machine Learning-Based Student Performance Prediction

The findings of this systematic literature review highlight several important insights regarding the application of machine learning techniques in predicting student academic performance. First, Random Forest and Support Vector Machine algorithms are the most widely used predictive models in educational data mining studies. These algorithms demonstrate strong predictive performance due to their ability to capture complex relationships within educational datasets. Second, ensemble learning approaches often outperform single classification models because they combine multiple learning algorithms to improve predictive accuracy and reduce overfitting. Third, academic performance indicators and behavioral engagement metrics represent the most influential predictors of student success. These variables provide valuable insights into both cognitive and behavioral aspects of student learning. Finally, deep learning approaches such as neural networks are emerging as powerful tools for modeling complex educational data patterns [4], [16], [18].

CONCLUSION AND RECOMMENDATIONS

This systematic literature review has several limitations. First, the review relied on a limited set of academic databases, which may have excluded relevant studies published in other repositories [19]–[22]. Second, only English-language publications were included in the analysis, which may introduce language bias. Finally, publication bias may exist because studies reporting positive results are more likely to be published than those reporting negative findings [7].

Machine learning techniques play a crucial role in predicting student academic performance and identifying at-risk learners in higher education. By integrating educational data mining and learning analytics, universities can develop data-driven decision systems that support early intervention strategies and improve student retention [8].

Future research should focus on developing interpretable, scalable, and ethically responsible machine learning models that support effective educational decision-making [2], [12], [18].

Future research should focus on integrating explainable artificial intelligence techniques into student performance prediction models in order to improve model interpretability and transparency. Researchers should also develop cross-institution datasets that allow machine learning models to generalize across different educational contexts. Real-time learning analytics systems represent another promising research direction, as they enable continuous monitoring of student engagement and performance. Furthermore, ethical frameworks for educational artificial intelligence should be developed to ensure responsible use of student data and protect learner privacy [5].

ACKNOWLEDGEMENT

The authors would like to express their sincere appreciation to Universiti Teknologi MARA (UiTM) Perak Branch for providing the academic environment and institutional support that enabled the completion of this research. Special gratitude is extended to the Actuarial Science Department and the Department of Statistics and Decision Science, UiTM Perak Branch, for their encouragement, resources, and continuous support throughout the development of this study.

The authors also acknowledge the valuable contributions from Actuarial Partners Consulting, Malaysia, for providing professional insights related to analytical and predictive modeling approaches, which contributed to the conceptual understanding of machine learning applications in decision-making contexts.

Furthermore, the authors would like to thank the Department of Computer Science, BINUS University, Indonesia, for their academic collaboration and support in advancing interdisciplinary research in machine learning and educational data analytics.

The authors are also grateful to the researchers and scholars whose published works formed the foundation of this systematic literature review. Their contributions in the fields of educational data mining, learning analytics, and machine learning in education have significantly advanced the understanding of predictive analytics for improving student academic outcomes.

Finally, the authors appreciate the constructive feedback and support received from colleagues, reviewers, and the academic community, which have helped improve the quality and clarity of this research.

REFERENCES

1. D. Alboaneen, M. Almelihi, R. AlSubaie, R. Alghamdi, L. Alshehri, and R. Alharthi, “Development of a Web-Based Prediction System for Students’ Academic Performance,” Data, vol. 7, no. 2, p. 21, 2022, doi: 10.3390/data7020021.
2. E. J. Anagu and R. Wella, “Web-Based Machine Learning Model for Predicting Student Academic Performance in Tertiary Institutions,” J. Adv. Comput. Technol. Appl., vol. 7, no. 1, pp. 1–10, 2025.
3. R. S. Baker and K. Yacef, “The State of Educational Data Mining in 2009: A Review and Future Visions,” J. Educ. Data Min., vol. 1, no. 1, pp. 3–17, 2009.
4. P. Cortez and A. Silva, “Using Data Mining to Predict Secondary School Student Performance,” in Proceedings of the 5th Future Business Technology Conference, 2008, pp. 5–12.
5. R. Ferguson, “Learning Analytics: Drivers, Developments and Challenges,” Int. J. Technol. Enhanc. Learn., vol. 4, no. 5--6, pp. 304–317, 2012.
6. Ş. Kocakoyun-Aydoğan, T. Pura, and F. Bingül, “Predicting Students’ Academic Performances Using Machine Learning Algorithms in Educational Data Mining,” Malaysian Online J. Educ. Technol., vol. 12, no. 4, pp. 45–60, 2024.
7. D. Khairy, N. Alharbi, M. A. Amasha, M. F. Areed, S. Alkhalaf, and R. A. Abougalala, “Prediction of

Student Exam Performance Using Data Mining Classification Algorithms," Educ. Inf. Technol., vol. 29, pp. 21621–21645, 2024, doi: 10.1007/s10639-024-12619-w.

8. A. Nabil, M. Seyam, and A. Abou-elfetouh, "Prediction of Students' Academic Performance Based on Courses' Grades Using Deep Neural Networks," IEEE Access, vol. 9, pp. 140731–140746, 2021, doi: 10.1109/ACCESS.2021.3119596.
9. C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," IEEE Trans. Syst. Man. Cybern., vol. 40, no. 6, pp. 601–618, 2010.
10. G. Siemens and R. Baker, "Learning Analytics and Educational Data Mining: Towards Communication and Collaboration," in Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, 2012, pp. 252–254.
11. M. T. Sathe and A. Adamuthe, "Comparative Study of Supervised Algorithms for Prediction of Students' Performance," Int. J. Mod. Educ. Comput. Sci., vol. 13, no. 1, pp. 1–12, 2021.
12. . Jacob and R. Henriques, "Educational Data Mining to Predict Bachelors Students' Success," Emerg. Sci. J., vol. 7, no. 2, pp. 345–357, 2023.
13. G. G. Dongre, "Predicting Student Dropout Rates in Higher Education: A Comparative Study of Machine Learning Algorithms," Int. J. Sci. Res. Eng. Manag., vol. 8, no. 2, pp. 1–10, 2024.
14. F. Adamu-Fika, D. B. Madaki, A. E. Baba-Onoja, A. T. Ramalan, A. T. Mohammed, and K. S. Bature, "Modelled Machine Learning Algorithms to Predict Students' Academic Performance in Tertiary Institutions," in Advances in Multidisciplinary Scientific Research Conference Proceedings, 2023, pp. 418–427.
15. Namraiza, K. Abid, N. Aslam, M. Fuzail, M. S. Maqbool, and K. Sajid, "An Efficient Deep Learning Approach for Prediction of Student Performance Using Neural Networks," VFAST Trans. Softw. Eng., vol. 11, no. 4, pp. 45–56, 2023.
16. L. Vives et al., "Prediction of Students' Academic Performance in the Programming Fundamentals Course Using Long Short-Term Memory Neural Networks," IEEE Access, vol. 12, pp. 5882–5898, 2024, doi: 10.1109/ACCESS.2024.3350169.
17. M. Yağcı, "Educational Data Mining: Prediction of Students' Academic Performance Using Machine Learning Algorithms," Smart Learn. Environ., vol. 9, p. 11, 2022, doi: 10.1186/s40561-022-00192-z.
18. D. T. Tempelaar, B. Rienties, and B. Giesbers, "In Search for the Most Informative Data for Feedback Generation: Learning Analytics in a Data-Rich Context," Comput. Human Behav., vol. 47, pp. 157–167, 2015.
19. M. Z. Awang Chek and I. L. Ismail, "Maximizing Retirement Savings: Strategic Forecasting of Employees' Provident Fund (EPF) Dividends," Int. J. Res. Innov. Soc. Sci., 2024.
20. M. Syakir, M. Z. A. Chek, and I. L. Ismail, "Understanding A Long-Term Care towards Ageing Population in Malaysia," Int. J. Acad. Res. Bus. Soc. Sci., vol. 13, no. 12, pp. 4744–4754, 2023, doi: 10.6007/ijarbss/v13-i12/20328.
21. I. L. Ismail, N. F. Jamal, M. Z. Awang Chek, and M. S. Baharuddin, "Learning Basic Statistics and Probability Through MOOC," Int. J. Mod. Trends Soc. Sci., vol. 2, no. 8, pp. 99–107, 2019, doi: 10.35631/ijmtss.280010.
22. A. N. A. Ahmad Ridzuan, M. Z. Awang Chek, N. M. Abdul Ghafar, and A. B. Ahmad, "Developing an Introduction to Actuarial Science MOOC," Int. J. Acad. Res. Bus. Soc. Sci., vol. 8, no. 1, pp. 600–605, 2018, doi: 10.6007/ijarbss/v8-i1/3833.