A Data-driven Framework for Analyzing Student Dropout in Malaysia

Ahmad Akif Ibrahim, Muhammad Azrin Ahmad, and Nor Azuana Ramli

Abstract **– Student dropout continues to pose a problem in Malaysia's education system. This issue affects secondary schools, technical and vocational education and training (TVET) institutions, and higher education institutions. Drawing on evidence from several peer-reviewed studies, this paper proposes a data-driven framework for analyzing and predicting student dropout in Malaysia. Prior work has identified five consistent categories in the literature: academic performance, socioeconomic conditions, institutional factors, psychological dimensions, and demographic attributes. This framework's novelty lies in its systematic integration of these variables with ensemble machine learning methods and explainable artificial intelligence for interpretability. Machine learning approaches, especially ensemble methods such as Random Forest, XGBoost, and Gradient Boosting, have achieved 87% to 99% accuracy in dropout prediction tasks based on previous studies. The results of this framework are expected to provide evidence-based insights to policymakers and educators in formulating more targeted, relevant, and effective prevention strategies and intervention programs.**

Keywords **– Student dropout, data-driven, framework, machine learning, predictive model.**

I. INTRODUCTION

Student dropout in Malaysia is a pressing issue with serious consequences for individuals, institutions, and national priorities. Dropout rates differ across educational levels and types, especially in secondary schools, technical and vocational education and training (TVET) institutions, and higher education. Student attrition reflects both individual academic challenges and structural issues like educational equity, resource distribution, and support systems (Mokshein et al., 2016).

In Malaysia, various contextual factors influence student dropout rates, including diverse socioeconomic conditions across urban and rural areas, multicultural and multilingual student populations, varying quality of educational infrastructure, and unequal access to support services (Carlo et al., 2024). Recent studies highlight those financial constraints, geographical challenges, and sociocultural norm, particularly in states such as Selangor and Sarawak, significantly affect dropout rates (Sariman et al., 2024).

In TVET institutions, the primary reasons for dropouts are lack of interest (38%), job offers (25%), personal issues (16%), study offers (14%), and health issues (7%) (Rodzi et al., 2024). High dropout rates matter for individuals and institutions alike, leading to economic losses, limited opportunities, and weakened institutional effectiveness. Dropout can limit students' long-term prospects and affect schooling missions and reputations.

Recent advances in data analytics and machine learning have enabled a shift in dropout prevention, moving from reactive interventions to proactive, predictive systems. Algorithms now analyze complex student data to identify at-risk individuals early, enabling timely, focused support (Kok et al., 2024). Some models even achieve near-perfect classification performance with comprehensive data (Mun et al., 2023).

This paper reviews student dropout risk factors in Malaysia and the use of machine learning techniques to predict them. The scope includes identifying Malaysia-specific dropout risk factors and systematically reviewing relevant machine learning methods and their performance. Then, a tailored, data-driven framework for Malaysian institutions is proposed to offer strategic recommendations. The framework can guide the effective application of evidence-based early warning systems to reduce dropout rates and improve student success across Malaysia's education sector.

II. PROBLEM STATEMENT

Student attrition across the education continuum remains a persistent challenge in Malaysia, affecting secondary schools, technical and vocational education and training (TVET) institutions, and higher education institutions. Recognizing its impact on national human capital development, the Malaysian government has introduced various initiatives and policy frameworks to improve student retention and completion rates, most notably the Malaysia Education Blueprint 2013–2025, alongside targeted financial assistance, alternative education pathways, and efforts to strengthen TVET and higher education accessibility.

Despite these interventions, dropout and attrition continue to occur at multiple stages of the education pipeline. At the secondary level, failure to complete schooling or sit for key examinations such as Sijil Pelajaran Malaysia (SPM) restricts further academic and career progression, while in TVET and higher education, attrition disrupts skills development and workforce readiness. Existing studies have identified a wide spectrum of contributing factors, including socio-economic disadvantage, academic underperformance, disengagement, institutional constraints, and limited family or community support. However, much of the current research relies on descriptive or cross-sectional approaches that do not adequately capture the complex, dynamic, and multi-level nature of student attrition across different educational pathways, thereby limiting the ability to identify early

Ahmad Akif Ibrahim, Centre for Mathematical Sciences, Universiti Malaysia Pahang Al-Sultan Abdullah, Malaysia (Email address: akif.ibrahim@gmail.com).
Muhammad Azrin Ahmad, Centre for Mathematical Sciences, Universiti Malaysia Pahang Al-Sultan Abdullah, Malaysia (Email address: azrin@umpsa.edu.my).
Nor Azuana Ramli, Centre for Mathematical Sciences, Universiti Malaysia Pahang Al-Sultan Abdullah, Malaysia (Email address: azuana@umpsa.edu.my).

warning signals and implement timely, targeted interventions.

III. LITERATURE REVIEW ON STUDENT DROPOUT RISK FACTORS IN MALAYSIA

To address dropout effectively, it is necessary to consider the risk factors associated with dropout. This is important for developing effective prevention strategies. Research in the Malaysian context has identified five major categories of risk factors. These factors interact in complicated ways to influence student persistence and attrition decisions.

Academic Risk Factor

Academic performance ranks among the strongest predictors of student dropout in Malaysia. Poor academic achievement and a low cumulative grade point average (CGPA) are strongly associated with an increased risk of dropout (Mokshein et al., 2016; Tarmizi et al., 2019). Students who struggle to maintain satisfactory academic performance face compounding challenges. These include lowered self-confidence, increased stress, and diminished motivation to continue their studies.

The inability to complete studies within the stipulated periods is a significant form of academic attrition, especially in science, technology, engineering, and mathematics (STEM) fields (Tarmizi et al., 2019). Students who fail to meet required academic standards or adapt to course demands, particularly in higher education, where the transition from secondary school requires independent learning and time management, face an increased risk of dropping out (Tarmizi et al., 2019).

Lack of interest in schooling is another important academic risk factor. This is especially seen in secondary education and TVET institutions (Mokshein et al., 2016; Rodzi et al., 2024). When students perceive education as irrelevant to their goals or find the curriculum unengaging, their motivation to continue drops. This disengagement often results in poor attendance, declining performance, and eventually withdrawal from education programs.

Socioeconomic Risk Factor

Socioeconomic pressures are among the most persistent drivers of dropout risk in Malaysia. Economic constraints and poverty consistently emerge as leading drivers of student attrition (Sariman et al., 2024; Carlo et al., 2024; Subra et al., 2019). Families in financial hardship often cannot afford educational expenses. These include tuition fees, transportation, learning materials, and other costs. This financial pressure often forces students to choose between pursuing their education and contributing to their family's income through employment.

Household income and family financial status directly influence dropout decisions (Hashim et al., 2024; Tarmizi et al., 2019; Rahman et al., 2025). Students from low-income families face many disadvantages. These include reduced access to schooling resources, inadequate study environments at home, and pressure to work part-time or full-time to support the family (Subra et al., 2019). The financial burden is especially acute in higher education, where tuition fees and living expenses are substantial.

Parental education level and socioeconomic status affect educational attainment across generations. Low parental education correlates with reduced emphasis on children's education, limited ability to provide academic support, and lower expectations for achievement (Subra et al., 2019). Parents with limited education may not fully understand the long-term benefits of staying in school. They may also lack the knowledge to navigate institutional systems and advocate for their children's needs.

Family structure and functioning also have important roles. Marital instability among parents, weak family relationships, and poor family functioning contribute to social problems that increase the risk of dropout. When parents work long hours or multiple jobs to meet financial needs, they have limited time for parental attention and guidance. Children may then seek attention from friends and others, which can lead to negative social influences (Subra et al., 2019).

Institutional Risk Factor

Institutional conditions shape student persistence patterns. Inadequate facilities and infrastructure are major institutional barriers to student success (Carlo et al., 2024; Raghavan et al., 2015). A lack of computers, poor conditions in the library and computer lab, and inadequate classrooms create poor learning environments. These conditions do not support learner participation and achievement (Raghavan et al., 2015).

The quality of education and effectiveness of instruction directly affect students' outcomes. Teacher teaching methods, lecturer preparedness, and instructional quality all influence student motivation and learning (Zakari et al., 2022; Sabtu et al., 2016). When tutors or facilitators lack knowledge, are unprepared, respond slowly, are unapproachable, or are inexperienced, students receive inadequate academic support, undermining their ability to succeed (Raghavan et al., 2015).

Support services and administrative systems constitute an additional critical institutional dimension. Poor administrative and academic support, lack of information from learning centers, slow responses to inquiries, and difficulty following institutional procedures create barriers that frustrate students and reduce their sense of institutional belonging (Raghavan et al., 2015). Insufficient positive interactions between students and faculty or staff further erode the social integration that promotes persistence (Sabtu et al., 2016).

Assessment and program design elements also contribute to the risk of dropout. Constant changes in assessment formats, delays in releasing examination results, unsuitable examination timetables, and programs that are too demanding or tough all increase student stress and attrition (Raghavan et al., 2015). In open and distance learning settings, problems accessing virtual learning environments and difficulties with online engagement create additional barriers (Tarmizi et al., 2019; Mohamad et al., 2018).

Psychological Risk Factor

Psychological factors are often overlooked, yet they strongly influence dropout risk. Motivation is a central psychological construct that influences persistence decisions (Rahman et al., 2025). Students with low motivation, poor self-esteem, and fear of failure find it difficult to maintain the effort and resilience required to overcome educational obstacles (Raghavan et al., 2015). Depression, dissatisfaction with courses or institutions, and experience of isolation further undermine psychological health and academic engagement (Tarmizi et al., 2019).

Psychosocial problems identified in qualitative research reveal the complicated emotional and social challenges students face (Reza et al., 2025). These include difficulty coping with academic pressure, difficulty maintaining self-confidence, and challenges in developing productive coping strategies. Health problems, both physical and mental, can precipitate dropout when students lack adequate support systems to cope with these challenges as they continue their studies (Tarmizi et al., 2019; Rodzi et al., 2024).

The psychological dimension of dropout is especially clear in students' capacity to study and adapt to educational demands. Inability to keep up with peers, difficulty coping with the number of courses, and poor time management reflect both ability deficits and psychological barriers to competent learning (Raghavan et al., 2015). When students lack social support and feel isolated from peers and institutional communities, these psychological challenges intensify (Tarmizi et al., 2019).

Demographic Risk Factor

Demographic characteristics interact with other risk factors to shape dropout patterns in complex ways. Gender influences dropout risk, though the direction and magnitude of effects vary across contexts (Hashim et al., 2024). Ethnicity and cultural background constitute significant demographic factors in Malaysia's multicultural society (Tray et al., 1984). Indigenous youth face particular challenges related to cultural issues, geographical isolation, and differential access to educational facilities (Hassan et al., 2023). Sociocultural norms in different community's shape attitudes toward education, gender roles, and the perceived value of formal schooling (Sariman et al., 2024).

Age and educational stage influence dropout risk through developmental and contextual mechanisms. First-year students face special challenges related to transition and assimilation, with motivation, academic factors, financial issues, social integration, and socio-demographic factors all contributing to early attrition (Rahman et al., 2025). In open and distance learning settings, adult learners face characteristic challenges related to balancing education with work and family responsibilities (Raghavan et al., 2015).

Geographical location creates differential access to educational opportunities and support services. Transportation problems and inconvenient learning center locations create practical barriers to attendance and engagement (Sariman et al., 2024; Raghavan et al., 2015). Rural students face compounded disadvantages related to infrastructure, resource availability, and distance from educational institutions (Carlo et al., 2024).

IV. REVIEW OF THE MACHINE LEARNING MODELS

Over the past decade, machine learning has been increasingly applied, and researchers have developed progressively more advanced models that achieve remarkable predictive accuracy. This section reviews the major algorithmic approaches, their performance characteristics, and their application for dropout prediction in educational contexts.

Logistic Regression and Decision Trees

Logistic regression and decision trees represent foundational machine learning approaches that offer important advantages in interpretability and computational efficiency. These methods provide transparent decision-making processes that can be easily explained to institutional stakeholders, making them valuable for contexts where model explainability is prioritized.

Logistic regression is a standard baseline for binary classification tasks in educational research. Its probabilistic output provides intuitive risk scores, and its coefficients offer clear insights into how different factors influence the probability of dropout. In an early prediction study using minimal data, logistic regression achieved an F1-score of 0.695 and an area under the Receiver Operating Characteristic (AUC) of 0.690, serving as a baseline against which more complex models were compared (Hamsiah et al., 2025). Research on university student dropout found that logistic regression consistently yielded results closest to optimal performance, even if it did not always achieve the highest accuracy (Segura et al., 2022).

Decision trees offer highly interpretable rule-based classification by hierarchically splitting the data based on feature values. Single decision trees, such as CART (Classification and Regression Trees) and C4.5 (J48), provide transparent decision paths that can be visualized and understood by non-technical stakeholders. In early warning systems using eBook data, the J48 algorithm achieved the best performance on transformed data with an accuracy of 0.833 and a Kappa of 0.665 (Akçapınar et al., 2019).

However, single decision trees are prone to overfitting and generally achieve lower accuracy than ensemble methods. In comparative analyses, the decision trees achieved an accuracy of 91.24%, substantially lower than Random Forest (96.01%) and gradient boosting methods (Utami et al., 2025). The CART algorithm was outperformed by ensemble models in graduation prediction tasks (Alfahid, 2024).

Despite lower predictive accuracy, these interpretable methods remain valuable for initial model development, stakeholder communication, and regulatory compliance. They can serve as components in ensemble approaches or as baseline models for performance comparison. In production systems, C5.0 decision tree models have been successfully deployed for dropout prediction, achieving reasonably good class separability, comparable to Random Forest

performance, while maintaining explainability (Ortigosa et al., 2019).

Neural Networks

Neural network models offer powerful capabilities for modelling complex and non-linear relationships in student data. These methods can automatically learn hierarchical feature representations, potentially capturing subtle patterns that traditional algorithm might miss. However, their opaque nature, difficulty in interpretation, and high computational demands pose challenges for institutional adoption.

Performance benchmarks for neural networks vary considerably depending on architecture and data characteristics. A study using multilayer perceptron (MLP) neural networks achieved 84% correct prediction for regular students, 81% for irregular students, and 76% for dropout students when using a delta threshold higher than 0.6. The optimal network architecture identified was MLP 49-34-1, with 49 input neurons, 34 hidden neurons, and 1 output neuron, trained over 5,000 epochs (Siri, 2015).

Neural network approaches using the TensorFlow and Keras frameworks have shown varying performance depending on the threshold selection. At a 10% dropout-probability threshold, an MLP achieved 77% overall accuracy, with 53% sensitivity for predicting dropout students and 80% specificity for predicting current students. When the threshold was adjusted to 50%, accuracy increased to 90%, but sensitivity dropped to 16% while specificity reached 99% (Tsai et al., 2020). This trade-off between sensitivity and specificity underscores the importance of threshold calibration aligned with institutional priorities.

Neural networks have been evaluated alongside other algorithms in comparative studies. In a modular pipeline assessment, an MLP was included among nine algorithms, though XGBoost ultimately achieved superior performance (Bettahi et al., 2025). The computational requirements and training complexity of neural network approaches can be substantial. Feed-forward neural networks require careful architecture design, hyperparameter tuning, and sufficient training data to avoid overfitting. Despite these challenges, neural networks remain valuable for institutions with large datasets and computational resources, particularly when interpretability is less critical than predictive accuracy (Prasanth et al., 2023).

Ensemble Methods

Ensemble methods combine multiple machine learning algorithms to achieve superior performance compared to individual models. By leveraging the strengths of different algorithms and reducing individual model weaknesses, ensembles consistently demonstrate the highest predictive accuracy in student dropout prediction tasks. Three main ensemble strategies are commonly employed: bagging, boosting, and stacking.

Bagging approaches, such as Random Forest (RF) and Bagged Trees, train multiple models on different subsets of training data and aggregate their predictions. RF has emerged as one of the most widely adopted and effective algorithms for predicting student dropout. As an ensemble method that combines multiple decision trees, RF demonstrates effective performance across diverse datasets and institutional contexts. The algorithm's ability to handle complex, non-linear relationships and its resistance to overfitting make it particularly well-suited for educational data with multiple interacting risk factors.

Performance benchmarks for RF consistently show high accuracy. Research on student retention management found that RF outperformed individual models, with ensemble approaches yielding superior predictive performance (Delen, 2010). The algorithm's feature-importance capabilities provide valuable insights for intervention design. RF can identify which variables contribute most strongly to dropout risk, enabling institutions to prioritize support services and allocate resources effectively. In comparative analyses, RF has demonstrated strength in handling imbalanced datasets when combined with techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) (Utami et al., 2025).

RF's interpretability, while not as transparent as single decision trees, offers reasonable explainability through feature importance rankings. This characteristic makes it more acceptable for institutional adoption compared to pure black-box approaches. The algorithm's computational efficiency and ability to handle missing data further enhance its practical applicability in real-world educational settings (Heredia-Jiménez et al., 2020).

Boosting methods, including AdaBoost, Gradient Boosting, and Extreme Gradient Boosting (XGBoost), sequentially train models with each new model focusing on correcting errors from previous models. Boosted decision tree classifiers have demonstrated optimal performance, with one study reporting an AUC of 0.898 using precision-recall curve analysis (Lee et al., 2019). Gradient Boosting methods, particularly XGBoost, have demonstrated exceptional performance in student dropout prediction tasks. These algorithms build models sequentially, with each new model correcting errors made by previous models, resulting in highly accurate predictions. XGBoost's optimized implementation includes regularization techniques that prevent overfitting and enhance generalization to new data.

XGBoost's performance benchmarks are particularly impressive. In a study on university dropout prevention, XGBoost achieved 96.9% accuracy and 94.4% recall, leading to its selection as the final model (Mun et al., 2023). Research on AI-driven forecasting found that XGBoost achieved the highest accuracy, 87%, among the tested algorithms (Kirange et al., 2025). In early prediction with minimal data, CatBoost (a variant of gradient boosting) achieved the best performance, with an F1 score of 0.770 and an ROC AUC of 0.750 (Hamsiah et al., 2025).

XGBoost is particularly effective for early prediction. A study using pre-enrolment data found that an XGBoost classifier achieved an AUC-ROC of 0.6902 and an F1-score of 0.6946 for the dropout class, with a sensitivity of 88% (Mendívil et al., 2025). This early predictive capability is crucial for implementing proactive interventions before students experience academic failure. One reason for this is its ability to handle feature interactions and non-linear relationships, which makes it effective for capturing the multidimensional nature of dropout risk. In a modular machine learning pipeline, XGBoost achieved an AUC-ROC of 0.993, a F1-score of 0.911, and a recall of 0.944 on

test data, with stable performance across different credit load bins. When tested on external datasets, XGBoost maintained top-ranked performance, achieving an F1-score of 0.792 and an AUC-ROC of 0.922, demonstrating strong generalization capabilities (Bettahi et al., 2025).

Gradient Boosting methods have also shown effectiveness in cascaded model architecture. Research on real-world machine learning applications found that Gradient Boosting models, when cascaded with outputs serving as inputs to subsequent models, achieved improved accuracy by leveraging knowledge from earlier prediction stages. The Gradient Boosting model was preferred for its higher precision (62.195), despite a slightly lower recall than ensemble alternatives (Fernández-García et al., 2021).

Stacking and hybrid ensemble approaches combine predictions from multiple diverse algorithms. Research on real-world machine learning applications found that ensemble models combining Gradient Boosting, Random Forests, and Support Vector Machines significantly improved results, increasing recall accuracy by 10 percentage points and precision by 12 percentage points compared to individual models. At enrolment, the ensemble model identified almost 75% of students who would drop out (Fernández-García et al., 2021).

The effectiveness of ensemble approaches is particularly evident in handling class imbalance. Studies using SMOTE combined with RF have achieved high accuracy while maintaining strong performance on minority-class dropout (Utami et al., 2025). Ensemble methods also demonstrate superior generalization, performing well across different institutional contexts and student populations (Bettahi et al., 2025).

V. PROPOSED DATA-DRIVEN FRAMEWORK

Building on the identified risk factors and advances in machine learning, this section outlines a structured framework designed to support early identification and prevention of student dropout in Malaysian educational institutions. Rather than operating as a purely technical system, the framework connects data processes with institutional decision-making, enabling timely and targeted interventions.

Framework Architecture

At its core, the framework comprises five interconnected components that collectively transform raw institutional data into actionable insights, as shown in Figure 1. The process begins with data collection, which integrates multi-source data from secondary schools, TVET institutions, and higher education institutions in Malaysia. These data sources include student information systems (e.g., academic performance, enrolment history), attendance and behaviour al records, financial and socio-economic information, and institutional support services such as counselling and health center records. In addition, where available, psychometric and well-being datasets, such as the Depression Anxiety Stress Scale (DASS), are incorporated to capture students' psychological conditions. Furthermore, digital learning interaction data (e.g., LMS usage, assignment submissions) may be utilized to reflect student engagement. The integration of these heterogeneous data sources enables a comprehensive, multidimensional representation of student profiles for subsequent analysis. Bringing these sources together ensures a more complete view of each student, capturing not only academic performance but also engagement patterns and socioeconomic context.

Once collected, the data undergoes cleaning, preprocessing, and integration, where inconsistencies are corrected, duplicate records are resolved, and data from multiple systems and timeframes are harmonized into a unified structure. Ensuring high data quality at this stage is critical, as errors can propagate and degrade model reliability. The process then advances to feature engineering, in which raw inputs are transformed into meaningful indicators capturing temporal dynamics, intervariable relationships, and behavioural patterns. Rather than relying on static variables, this stage derives longitudinal features such as trends in attendance, fluctuations in academic performance, and changes in engagement levels, all of which are known to signal early risk of student attrition.

Following this, predictive modeling is performed using machine learning techniques, including ensemble approaches such as Random Forests and XGBoost, which are well-suited for capturing complex, nonlinear relationships and have demonstrated strong performance in educational prediction tasks (Mun et al., 2023; Utami et al., 2025). To address the limitation of advanced models, an Explainable AI (XAI) layer is integrated alongside the modelling process. This layer provides both global and local interpretability through techniques such as feature importance analysis and SHAP (SHapley Additive exPlanations) values, enabling stakeholders to understand not only which factors influence dropout risk overall, but also why specific students are classified as high risk. Such explainability enhances trust, supports fairness assessment, and facilitates actionable insights.

Finally, the model outputs and their explanations are operationalized through decision-support mechanisms. Students are categorized into risk levels, early warning alerts are generated, and recommended interventions are suggested based on key contributing factors identified by the XAI layer. Interactive dashboards enable educators and policymakers to visualize risk patterns, explore underlying drivers, monitor intervention outcomes, and continuously refine strategies. In return, feedback from these interventions can be used to refine models and improve future predictions (Paul et al., 2025). This establishes a continuous learning loop, ensuring that the framework evolves over time and remains responsive to changing student behaviour s and institutional contexts.

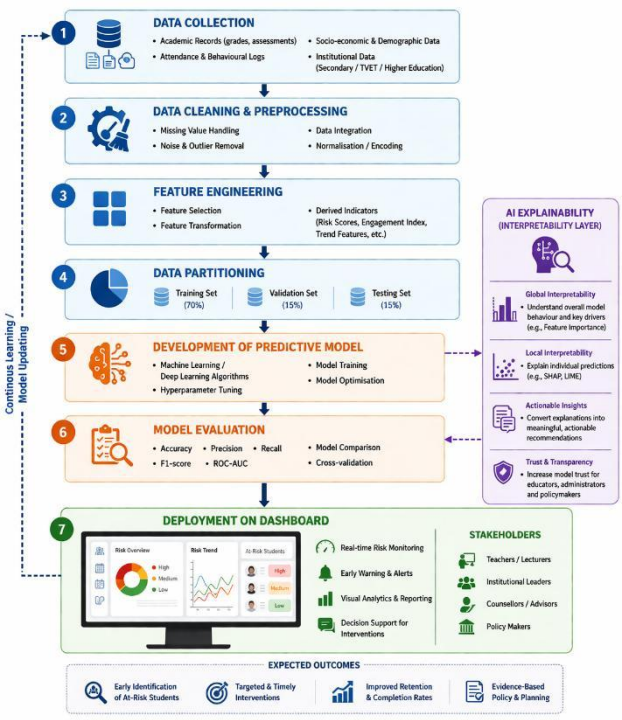


Figure 1. Proposed Framework

VI. CONCLUSION

Student dropout in Malaysia poses a complex challenge shaped by interactions among academic, socioeconomic, institutional, psychological, and demographic factors. This report synthesizes evidence from 60 peer-reviewed studies to develop a comprehensive, data-driven framework for analyzing and predicting student dropout in Malaysian educational contexts. The framework integrates insights from Malaysian dropout research with advances in machine learning to enable proactive, evidence-based dropout prevention.

The review of dropout risk factors reveals the multidimensional nature of attrition in Malaysia. Academic performance, particularly low CGPA and poor achievement, consistently emerges as a primary predictor (Mokshein et al., 2016; Tarmizi et al., 2019; Hashim et al., 2024). Socioeconomic factors, including poverty, family income, and parental education, create pervasive barriers to educational persistence (Subra et al., 2019; Sariman et al., 2024; Carlo et al., 2024). Institutional factors such as inadequate facilities, poor support services, and teaching quality significantly influence student outcomes (Raghavan et al., 2015; Sabtu et al., 2016). Psychological dimensions, including motivation, self-esteem, and mental health, affect students' capacity to persist in the face of challenges (Rahman et al., 2025; Reza et al., 2025). Demographic characteristics, including gender, ethnicity, and geographic location, interact with other risk factors in complex ways (Hashim et al., 2024; Hassan et al., 2023).

Machine learning approaches have demonstrated exceptional capabilities for dropout prediction. Ensemble methods, including Random Forest, XGBoost, and Gradient Boosting, consistently achieve the highest accuracy, typically ranging from 87% to 99% depending on data quality and prediction timing (Mun et al., 2023; Kirange et al., 2025; Kok et al., 2024; Utami et al., 2025). These algorithms can identify at-risk students early in their academic careers, enabling timely interventions before students experience failure or disengage completely. The predictive power of these models increases substantially as more data accumulate throughout the semester, with week 3 identified as an optimal time to establish support interventions (Hamsiah et al., 2025; Akçapınar et al., 2019).

The proposed data-driven framework provides a comprehensive architecture for implementing dropout prediction systems in Malaysian institutions. The framework encompasses data collection from multiple institutional sources; data cleaning and preprocessing; feature engineering to capture relevant risk factors; predictive modeling using ensemble machine learning algorithms with explainable AI; and decision support systems that translate predictions into actionable interventions. Integration with existing institutional systems ensures that predictive insights flow seamlessly into support service workflows.

Strategic recommendations emphasize the need for supportive policy environments at national and institutional levels. National policies should establish dropout-prevention priorities, develop data infrastructure standards, create ethical frameworks for AI in education, and support research and interinstitutional collaboration. Institutional practices should build analytical capacity, implement ethical review processes, develop intervention capacity, and engage students and faculty as partners in retention efforts. Adequate funding and resource allocation are essential for sustainable implementation. Comprehensive data governance frameworks must protect student privacy, ensure algorithmic fairness, maintain transparency, and establish accountability mechanisms.

The transition from reactive responses to proactive, data-driven dropout prevention represents a fundamental shift in how educational institutions support student success. By identifying at-risk students early and providing timely, targeted interventions, institutions can substantially reduce dropout rates and improve educational outcomes. The framework presented in this paper provides Malaysian educational institutions with a roadmap for implementing evidence-based early warning systems that leverage machine learning while upholding ethical standards and protecting student rights.

Future research should continue to refine predictive models for Malaysian contexts, evaluate intervention effectiveness, and address emerging challenges related to algorithmic fairness and student privacy. As institutions gain experience with predictive analytics, knowledge sharing, and collaborative learning will accelerate adoption and improve practices across the Malaysian education system. With sustained commitment from policy makers, institutional leaders, and educational professionals, data-driven dropout prevention can contribute significantly to Malaysia's goals of educational equity, human capital development, and national prosperity.

ACKNOWLEDGEMENT

This work was supported by Centre for Academic Innovation and Competitiveness (PIDA), Universiti

Malaysia Pahang Al-Sultan Abdullah under the grant Pengajaran dan Pembelajaran (PPU240109).

REFERENCES

Akçapınar, G., Altun, A., & Aşkar, P. (2019). Using learning analytics to develop an early-warning system for at-risk students. International Journal of Educational Technology in Higher Education, 16(1), Article 40.

Alfahid, A. (2024). Algorithmic prediction of students' on-time graduation from the university. TEM Journal, 13(1), 572-582.

Bettahi, Y., Benabbou, F., & Sael, N. (2025). A modular and explainable machine learning pipeline for student dropout prediction in higher education. Algorithms, 18(10), Article 662.

Carlo, S., Ting, H., & Thurasamy, R. (2024). Underlying school dropout issues in Sarawak. International Journal of Business Society, 25(2), 1009-1022.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. Decision Support Systems, 49(4), 498-506.

Fernández-García, A. J., Preciado, J. C., Melchor, F., Confalonieri, R., Nieto, A. G., & Crespo, R. G. (2021). A real-life machine learning experience for predicting university dropout at different stages using academic data. IEEE Access, 9, 133076-133090.

Goren, G., Oster, M., Nachmias, R., & Tuvi-Arad, I. (2024). Early prediction of student dropout in higher education using machine learning models. Zenodo.

Hamsiah, A., Sulaiman, S., & Mustaffa, N. (2025). Early prediction of at-risk students using minimal data: A machine learning framework for higher education. Digitus, 3(2), 1-15.

Hashim, N., Ismail, R., & Hassan, N. (2024). Statistical identification of predictors of dropout in secondary education: Evidence from Malaysia. Journal of The Asia Pacific Economy, 29(4), 1847-1869.

Hassan, N., Embong, R., & Yusof, Y. (2023). Incidences and causal factors of school dropouts among the Indigenous youth in Malaysia. International Seminar on Vocational Studies and Educational Journal, 10(9), 1-12.

Heredia-Jiménez, A., Orellana, M., & Baculima, F. (2020). An early warning dropout model in higher education degree programs: A case study in Ecuador. Maskana, 11(2), 43-54.

Kirange, D., Deshmukh, R., & Kirange, M. (2025). AI in academia: Forecasting student dropouts. International Journal Of Scientific Research In Engineering & Technology, 5(4), 1-8.

Kok, S. K., Ng, K. W., Ong, S. L., & Tee, C. K. (2024). A novel predictive modeling for student attrition utilizing machine learning and sustainable big data analytics. Applied Sciences, 14(21), Article 9633.

Lee, W., Choi, H., & Lee, S. (2019). The machine learning-based dropout early warning system for improving the performance of dropout prediction. Applied Sciences, 9(15), Article 3093.

Mendívil, E. G., Eraña-Díaz, M. L., Cruz-Ramírez, M., & Acosta-Mesa, H. G. (2025). Predicting student dropout from day one: XGBoost-based early warning system using pre-enrollment data. Applied Sciences, 15(16), Article 9202.

Mohamad, S. N. M., Sazali, N. S. S., & Salleh, M. A. M. (2018). Malaysia MOOC: Improving low student retention with predictive analytics. International Journal of Engineering and Technology, 7(2.29), 72-76.

Mokshein, S. E., Ahmad, H. H., & Vongalis-Macrow, A. (2016). Trends and factors for dropout among secondary school students in Perak. The Malaysian Online Journal of Educational Science, 4(4), 11-28.

Multi-Class Phased Prediction of Academic Performance and Dropout in Higher Education. (2023). Applied Sciences, 13(8), Article 4702.

Mun, S., Park, M., & Whang, M. (2023). Applying machine learning-based models to prevent university student dropouts. Korean Journal of Educational Research, 36(2), 289-312.

Opazo, D., Moreno, S., Álvarez-Miranda, E., & Pereira, J. (2021). Analysis of first-year university student dropout through machine learning models: A comparison between universities. Mathematics, 9(20), Article 2599.

Ortigosa, A., Carro, R. M., Bravo-Agapito, J., & Lizcano, D. (2019). From lab to production: Lessons learnt and real-life challenges of an early student-dropout prevention system. IEEE Transactions on Learning Technologies, 12(2), 264-277.

Paul, S., Chatterjee, S., & Basu, D. (2025). AI-driven early warning systems for dropout risk and passed-out prediction in higher education. In Advances in Educational Technologies and Instructional Design (pp. 67-89). IGI Global.

Prasanth, P., Subrahmanyam, K., & Srinivasa Rao, P. (2023). Predictive modeling of student behaviour for early dropout detection in universities using machine learning techniques. In Proceedings of the 2023 International Conference on Emerging Trends in Advances in Science and Technology (pp. 1-6). IEEE.

Raghavan, S., Hogan, R., & Tay, L. Y. (2015). Access to higher education via ODL: Addressing attrition to maximize university sustainability. American Journal of Economics, 5(2), 129-135.

Rahman, N. A., Yusof, R., & Ismail, S. (2025). Factors affecting the intention to drop out among first-year students: Evidence from UiTM Kedah. International Journal of Education, Psychology and Counselling, 10(58), 1-15.

Reza, F., Yasin, M. A. S. M., & Yaacob, N. A. (2025). Psychological factors associated with students' withdrawal from Sijil Pelajaran Malaysia: A qualitative study. Journal of Research, Policy & Practice of Teachers and Teacher Education, 15(1), 1-15.

Rodzi, S. N. H. M., Ismail, I. M., & Yaakob, M. F. M. (2024). The cause of student dropout from TVET at the Malaysia Industrial Training Institute. International Journal of Academic Research in Business & Social Sciences, 14(9), 1-12.

Sabtu, M. Y., Yusoff, Y. M., & Omar, M. Z. (2016). Student attrition in technical and vocational education and training (TVET) institutions: The case of XCel Technical College in Malaysia. Asian Social Science,

12(12), 197-205.

Sariman, N. S., Hamzah, M. I., & Bakar, K. A. (2024). Prediction model of school dropout factors using classification techniques in Selangor. Malaysian Journal of Social Sciences and Humanities, 9(6), e02867.

Segura, M., Mello-Román, J., Pardo, P. J., & Cabezas, J. (2022). Machine learning prediction of university student dropout: Does preference play a key role? Mathematics, 10(18), Article 3359.

Siri, A. (2015). Predicting students' dropout at university using artificial neural networks. Italian Journal of Sociology of Education, 7(1), 225-247.

Subra, M., Meerah, T. S. M., & Halim, L. (2019). Family's socio-economic influence on the dropout of Indian students: A case study in the district of Kuala Muda, Kedah. International Journal of Modern Trends in Social Sciences, 2(9), 67-79.

Tarmizi, S. A. A., Mutalib, S., Abdul-Rahman, S., Hamid, N. H. A., & Yusoff, Z. M. (2019). A case study on student attrition prediction in higher education using data mining techniques. In Advances in Intelligent Systems and Computing (Vol. 978, pp. 127-136). Springer.

Tarmizi, S. A. A., Mutalib, S., Abdul-Rahman, S., Hamid, N. H. A., & Yusoff, Z. M. (2019). A review of student attrition in higher education using big data analytics and data mining techniques. International Journal of Modern Education and Computer Science, 11(8), 1-14.

Tray, L. D., Fernandez, J. P., & Loh, P. F. S. (1984). Schooling in Malaysia. Comparative Education Review, 28(2), 280-302.

Tsai, Y. R., Ouyang, C. S., & Chang, Y. (2020). Precision education with statistical learning and deep learning: A case study in Taiwan. International Journal of Educational Technology in Higher Education, 17(1), Article 12.

Utami, Y. T., Rahmawati, D., & Mahmudy, W. F. (2025). Comparative analysis of machine learning algorithms with RFE-CV for student dropout prediction. Jurnal Teknik Informatika, 6(3), 1-10.

Zakari, N. A., Hamzah, M. I., & Razak, K. A. (2022). Keciciran murid sekolah di Malaysia: Suatu pemerhatian awal [School dropout in Malaysia: A preliminary observation]. Malaysian Journal of Social Sciences and Humanities, 7(2), e001288.