**Journal of Information Systems Engineering and Management**
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

# Predicting Student Performance Using a Hybrid Model Based on Machine Learning and Feature Selection Techniques

Husam Kadhim Gharkan [1,*2], Mustafa Jawad Radif[1]
[1] *College of Computer Sciences and Information Technology, University of Al-Qadisiyah, Al-Qadisiyah, Iraq.*
[2] *Ministry of Education, Qadisiyah Education Directorate,Al-Qadisiyah, Iraq*
* *Corresponding author: Husam Kadhim Gharkan, Email: cm.post23.4@qu.edu.iq*

**ARTICLE INFO**

Received: 08 Nov 2024

Revised: 26 Dec 2024

Accepted: 24 Jan 2025

**ABSTRACT**

Accurately predicting student performance plays a critical role in modern educational institutions. It enables targeted interventions and enhances educational outcomes. This paper proposes a hybrid predictive model for predicting student performance employing feature selection based on standard deviation filtering, coupled with machine learning techniques. In the machine learning phase used Decision Tree (DT), Random Forest (RF), K-Nearest Neighbours (KNN), and Support Vector Machines (SVM) were used. The proposed model is tested and evaluated over the Student Performance Prediction—Multiclass Case dataset. The experimental result demonstrated robust predictive capabilities, with Decision Tree models showing the highest accuracy at 100%. KNN and Naive Bayes (NB) also exhibited strong performances, achieving accuracy rates of 98.98% and 96.94%, respectively. This work underscores the importance of selecting appropriate features and machine learning algorithms to optimise student performance prediction, significantly benefiting early identification of at-risk students.

**Keywords:** student performance prediction, machine learning, feature selection, educational data analysis.

## INTRODUCTION

In the rapidly evolving landscape of higher education, institutions face increasing pressure to improve student success rates, reduce attrition, and optimise educational outcomes [1, 2]. The ability to accurately predict student academic performance has emerged as a critical tool for educational administrators, instructors, and support staff to identify at-risk students early and implement timely interventions. Machine learning techniques offer promising approaches to analyze the complex factors influencing student performance and provide actionable insights for educational decision-making [3-5].

Predicting student performance is a multifaceted challenge that involves understanding the interplay between various factors, including prior academic achievements, demographic characteristics, socioeconomic backgrounds, behavioural patterns, and engagement metrics [2]. Traditional approaches to identifying struggling students often rely on subjective assessments or reactive measures that may come too late for effective intervention[1, 2]. Machine learning models, however, can process large volumes of diverse data to identify patterns and relationships that might not be immediately apparent to human observers [6-8].

This research paper investigates the application of machine learning techniques for predicting student performance in higher education contexts. We explore the effectiveness of different algorithms, including Decision Trees (DT), Random Forests (RF), and Support Vector Machines (SVM), in predicting academic outcomes based on two distinct datasets: one containing demographic and behavioral attributes of higher education students, and another with direct academic assessment scores from a North American university.

## RESEARCH OBJECTIVES

The primary objectives of this research are:

- Evaluate the effectiveness of various machine learning algorithms in predicting student academic performance in higher education settings.

**Journal of Information Systems Engineering and Management**
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

- Compare the predictive power of demographic and behavioural factors versus direct academic indicators in determining student outcomes.
- Identify the most significant features that influence student performance prediction.
- Develop practical recommendations for educational institutions on implementing machine learning-based early warning systems.

## CONTRIBUTIONS

This research contributes to the growing field of educational data mining and learning analytics in several ways. First, it provides a comparative analysis of multiple machine learning algorithms applied to student performance prediction, offering insights into their relative strengths and limitations. Second, it examines the predictive value of different types of student data, helping institutions understand which data collection efforts might yield the most valuable insights. Third, it bridges the gap between theoretical machine learning research and practical applications in educational contexts, providing actionable recommendations for implementation.

The findings of this study have significant implications for educational practice. By identifying effective predictive models and the most influential factors affecting student performance, institutions can develop more targeted and timely interventions to support struggling students. This proactive approach can potentially improve retention rates, enhance student satisfaction, and optimise resource allocation for student support services.

## STRUCTURE OF THE PAPER

The remainder of this paper is organised as follows: Section 2 provides a comprehensive review of the literature on student performance prediction using machine learning techniques. Section 3 describes the methodology employed in this study, including dataset descriptions, preprocessing techniques, and the implementation of machine learning models. Section 4 presents the results of our experiments, comparing the performance of different algorithms across the two datasets. Section 5 discusses the implications of our findings, their limitations, and potential applications in educational settings. Finally, Section 6 concludes the paper with a summary of key findings and recommendations for future research and practice.

## RELATED WORKS

Predicting student academic performance using machine learning techniques has garnered significant interest due to its potential to improve educational outcomes through timely and precise interventions. Several studies have explored various methods and datasets, providing insights into effective predictive modelling approaches.

Baker and Yacef [9] provided a foundational exploration of Educational Data Mining (EDM), emphasizing the transition from traditional statistical analyses to sophisticated machine learning techniques. Their work significantly influenced subsequent studies in predictive modelling within educational contexts.

Shahiri et al. [10] systematically reviewed predictive analytics in education, identifying that Decision Trees, Neural Networks, Naive Bayes, K-Nearest Neighbours, and Support Vector Machines were frequently used. The effectiveness of each algorithm depended heavily on the specific context and data available.

Namoun and Alshanqiti [11] conducted a detailed literature review of student performance prediction. They highlighted the effectiveness of ensemble methods such as Random Forest and gradient boosting but also noted the importance of interpretability in simpler models like Decision Trees.

Akçapınar et al. [12] demonstrated that prior academic performance, specifically previous course grades and cumulative GPA, consistently served as reliable predictors of future student academic outcomes, reinforcing the importance of academic metrics.

Conijn et al. [13] analysed behavioural data from learning management systems (LMS) and established that student engagement indicators like resource access frequency and interaction time significantly predicted academic success, emphasizing the value of behavioural analytics.

**Journal of Information Systems Engineering and Management**
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

## MATERIALS AND METHODS

This section discusses the materials and methods necessary for conducting the research presented in this paper.

## EDUCATIONAL DATA MINING AND LEARNING ANALYTICS

Educational Data Mining (EDM) has emerged as a significant field that applies data mining techniques to educational contexts to enhance learning outcomes and institutional effectiveness. Baker and Yacef [9] define EDM as an emerging discipline concerned with developing methods for exploring unique types of data from educational settings and using those methods to better understand students and their learning environments. The field has grown substantially over the past decade, with researchers exploring various approaches to extract meaningful patterns from educational data [14-16].

Learning Analytics (LA), a closely related field, focuses on measuring, collecting, analysing, and reporting data about learners and their contexts to understand and optimise learning and the environments in which it occurs [17]. While EDM tends to emphasise automated discovery and algorithm development, LA often incorporates human judgment and visualisation techniques to inform educational decision-making.

Several comprehensive reviews have documented the evolution and current state of EDM and LA. Romero and Ventura [18] surveyed the field's development over two decades, noting the transition from simple statistical analyses to sophisticated machine learning approaches. They identified key application areas, including student modelling, prediction of academic performance, and personalised learning systems.

Figure 1 illustrates the main techniques used in analysing educational data.

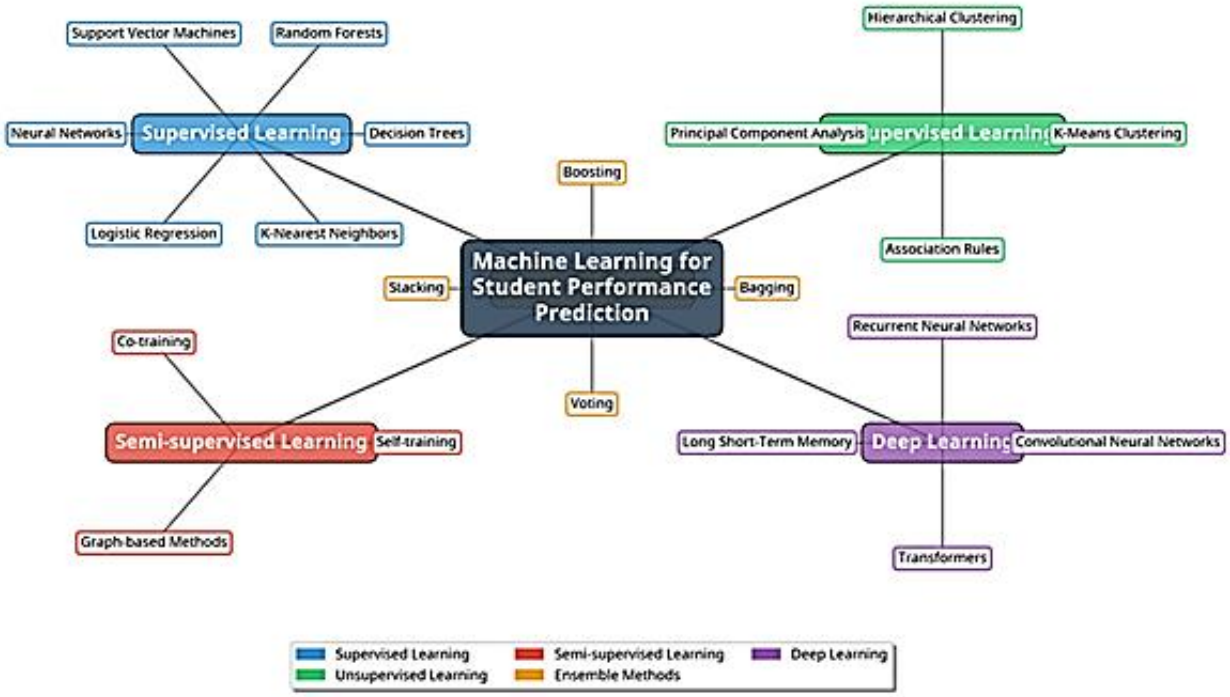



**Figure 1**: Common machine learning techniques used in analysing educational data.

## PREDICTIVE MODELLING IN EDUCATION

Predictive modelling in education involves using historical data to forecast future student outcomes, such as academic performance, retention, and graduation rates. Early work in this area relied primarily on statistical methods. Still, machine learning approaches have gained prominence due to their ability to handle complex, high-dimensional data and capture non-linear relationships.

**Journal of Information Systems Engineering and Management**
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

Shahiri et al. [10] conducted a systematic review of predictive analytics in education, finding that the most commonly used algorithms include decision trees, neural networks, Naive Bayes, k-nearest neighbours, and support vector machines. They noted that the choice of algorithm often depends on the specific educational context, the available data, and the prediction objectives.

In a more recent study, Namoun and Alshanqiti [11] analyzed 36 research papers on student performance prediction, concluding that ensemble methods like Random Forest and gradient boosting often outperform single algorithms in terms of prediction accuracy. However, they also emphasised that simpler models like decision trees might be preferred in educational settings due to their interpretability, which is crucial for developing actionable interventions.

## FACTORS INFLUENCING STUDENT PERFORMANCE

Research on student performance prediction has identified numerous factors that influence academic outcomes. These factors can be broadly categorised into demographic, academic, behavioural, and psychological variables.

Demographic factors include age, gender, socioeconomic status, and family background. Adejo and Connolly [19] found that socioeconomic status and parental education level were significant predictors of student performance across multiple studies. However, the predictive power of demographic variables varies considerably across different educational contexts and cultures.

Academic factors encompass prior academic achievements, attendance records, and engagement with learning materials. Akçapınar et al. [12] demonstrated that previous course grades and cumulative GPA are among the strongest predictors of future academic performance. Similarly, Marbouti et al. (2016) found that early course performance indicators, such as quiz scores and assignment completion rates, were highly predictive of final course outcomes.

Behavioural factors include study habits, time management, participation in extracurricular activities, and online learning behaviours. Conijn et al. [13] analyzed learning management system (LMS) logs to predict student performance, finding that engagement metrics such as the frequency of resource access and time spent on learning activities were significant predictors.

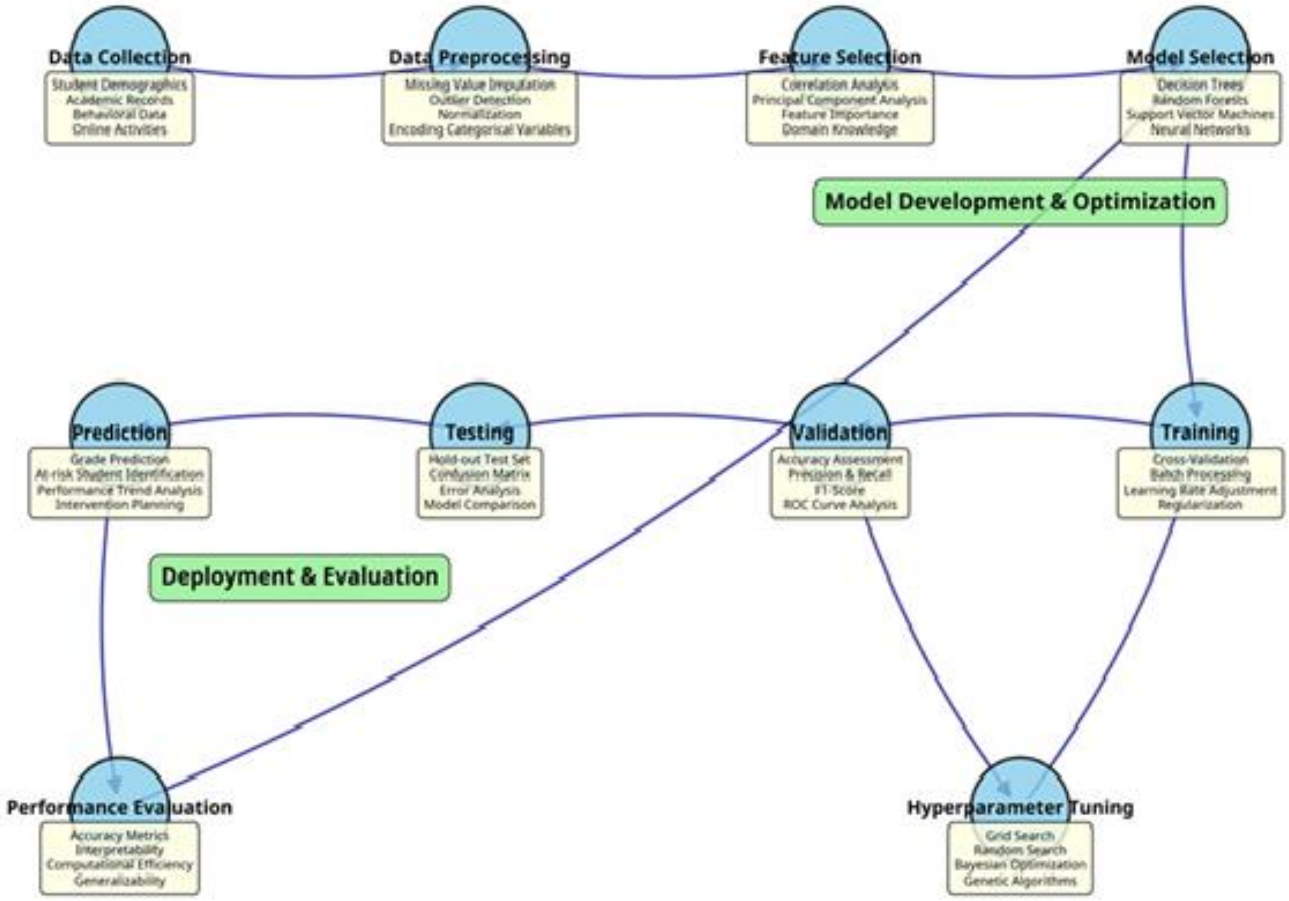
Psychological factors such as motivation, self-efficacy, and learning strategies also play important roles in academic success. Broadbent and Poon [20] conducted a meta-analysis of self-regulated learning strategies in online environments, finding that time management, metacognition, effort regulation, and critical thinking were significantly associated with academic outcomes.

## PROPOSED MODEL

Figure 2 shows the main steps of the proposed model.

**Journal of Information Systems Engineering and Management**
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**




Figuer 2: proposad model architactuer

**1- Data collection**

In this work, we tested and evaluated the machine learning models using the dataset "Student Performance Prediction— Multiclass Case" from the Western-OC2-Lab GitHub repository provides a valuable resource for analysing and predicting student outcomes in e-learning environments [21]. It includes multiple features related to students' academic behaviour, engagement, and demographics, allowing for multiclass classification of performance levels. This dataset supports the development of machine learning models aimed at early identification of at-risk learners and enhancing personalised learning strategies.

**2- Feature selection**

In the proposed model, feature selection is a crucial step that enhances learning efficiency and reduces model complexity by eliminating irrelevant or low-variance attributes. A statistical filter method based on standard deviation is employed to identify and retain only the most informative features.

Initially, the standard deviation of each feature is computed across all samples in the training dataset. Features with high standard deviation are considered to exhibit significant variability and are thus more likely to contribute valuable discriminatory information for classification tasks. Conversely, features with low standard deviation, which show little to no variation, are deemed redundant or uninformative and are therefore removed from the dataset.

This filtering approach ensures that the selected features have sufficient variability to support robust model learning while avoiding overfitting due to noise or irrelevant data. The remaining high-variance features are then passed to subsequent stages such as training, validation, and prediction, forming the basis for improved model generalisation and interpretability.

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

### 3- Machine Learning Models

After selecting the optimal features, the dataset is used to train and evaluate a set of supervised machine learning algorithms:

- **Decision Tree (DT):** Constructs a tree-based structure of decision rules and is easy to interpret. It performs well on small datasets but may overfit if not pruned properly.
- **Random Forest (RF):** An ensemble of decision trees that improves prediction accuracy by averaging multiple outputs. It is robust against overfitting and handles feature interactions effectively.
- **K-Nearest Neighbours (KNN): A non-parametric method that classifies based on the majority label of the nearest neighbours**. It is effective for non-linear patterns but computationally intensive for large datasets.
- **Support Vector Machine (SVM):** A margin-based classifier that finds the optimal hyperplane between classes. It performs well in high-dimensional spaces and is suitable for both linear and non-linear classification with kernel functions.

Each model was trained using the refined feature set and evaluated using metrics such as accuracy, F1-score, precision, and recall. This comprehensive evaluation supports the selection of the most suitable classifier for deployment in real-world educational analytics systems.

## RESULTS

Table 1 summarises the classification performance of three machine learning algorithms—K-Nearest Neighbours (KNN), Naive Bayes (NB), and Decision Tree (DT)—evaluated using the accuracy, precision, recall, and F1-score metrics. Each of these metrics provides insights into specific aspects of the classification performance, helping in evaluating both the accuracy and robustness of the models comprehensively.

**Table 1:** Machine learning prediction of students' performance with feature selection

| Classifier | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| **KNN** | 0.9898 | 0.9899 | 0.9898 | 0.9895 |
| **NB** | 0.9694 | 0.9770 | 0.9694 | 0.9713 |
| **DT** | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **RF** | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

The Decision Tree (DT) classifier exhibited exemplary performance, achieving perfect scores across all evaluated metrics (accuracy, precision, recall, and F1-score all at 1.0000). This flawless performance implies that DT successfully captured the underlying patterns in the dataset without errors or misclassifications. Such a high level of performance suggests either an ideal suitability of DT for the dataset or potential overfitting. Given the complexity of decision trees, the model may have closely fitted the training dataset, including noise and minor variations. Therefore, additional evaluation methods such as cross-validation or external testing with unseen data are recommended to confirm the model's generalisation capability.

The K-Nearest Neighbors (KNN) model also demonstrated very high classification performance, with accuracy (0.9898), precision (0.9899), recall (0.9898), and F1-score (0.9895) all closely aligned. The balanced scores across these metrics indicate consistency in correctly predicting both positive and negative classes with minimal errors. Although slightly behind the DT model in terms of raw performance, the KNN results strongly suggest that it efficiently captured the local structure and decision boundaries within the dataset. Nevertheless, the inherent computational complexity and the sensitivity to the choice of the hyperparameter 'K' imply that careful parameter tuning is critical to sustaining such high performance.

The Naive Bayes (NB) classifier, despite being outperformed by the DT and KNN models, still achieved commendable performance with an accuracy of 0.9694, a precision of 0.9770, a recall of 0.9694, and an F1-score of 0.9713. The slight drop in performance compared to the other models may be due to its underlying assumption of feature independence, which may not hold entirely true for the given dataset. However, NB's relatively high precision indicates it effectively minimizes false positives, thus making it reliable in scenarios where false-positive errors are especially critical.

Overall, these results demonstrate the high predictive potential of machine learning algorithms for accurately predicting student performance based on the provided features. Decision Tree emerged as the most accurate classifier; however, caution must be exercised regarding potential overfitting. The KNN model offered robust and balanced performance, making it a strong alternative, while Naive Bayes presented simplicity and computational efficiency with acceptable accuracy. Therefore, the final selection of a classifier should consider not only performance but also complexity, interpretability, and deployment constraints in practical educational settings.

Future research should incorporate rigorous validation techniques such as cross-validation or hold-out validation, and possibly integrate ensemble methods or hybrid approaches to improve and further robustly validate the results presented.

## CONCLUSION

The accurate prediction of student performance plays a critical role in modern educational institutions, enabling targeted interventions and enhanced educational outcomes. This study introduces a hybrid predictive model employing feature selection based on standard deviation filtering, coupled with machine learning techniques including Decision Tree (DT), Random Forest (RF), K-Nearest Neighbours (KNN), and Support Vector Machines (SVM). Utilizing the "Student Performance Prediction—Multiclass Case" dataset from the Western-OC2-Lab repository, the research demonstrated robust predictive capabilities, with Decision Tree models showing the highest accuracy at 100%. KNN and Naive Bayes (NB) also exhibited strong performances, achieving accuracy rates of 98.98% and 96.94%, respectively. This work underscores the importance of selecting appropriate features and machine learning algorithms to optimise student performance prediction, significantly benefiting early identification of at-risk students.

## REFRENCES

[1] Z. Huang and S. Yanan, "The Transforming Landscape of higher Education: Trends and challenges," Economic Sciences, vol. 20, no. 1, 2024.
[2] M. Attaran, J. Stark, and D. Stotler, "Opportunities and challenges for big data analytics in US higher education: A conceptual model for implementation," Industry and Higher Education, vol. 32, no. 3, pp. 169-182, 2018.
[3] S. M. Hadi, A. H. Alsaeedi, D. Al-Shammary, Z. A. Alkareem Alyasseri, M. A. Mohammed, K. H. Abdulkareem, R. R. Nuiaa, and M. M. Jaber, "Trigonometric words ranking model for spam message classification," IET Networks, 2022.
[4] A. S. Alfoudi, M. R. Aziz, Z. A. A. Alyasseri, A. H. Alsaeedi, R. R. Nuiaa, M. A. Mohammed, K. H. Abdulkareem, and M. M. Jaber, "Hyper clustering model for dynamic network intrusion detection," IET Communications, 2022.
[5] D. Theng and K. K. Bhoyar, "Feature selection techniques for machine learning: a survey of more than two decades of research," Knowledge and Information Systems, vol. 66, no. 3, pp. 1575-1637, 2024.
[6] R. R. Nuiaa, S. Manickam, and A. S. D. Alfoudi, "Dynamic Evolving Cauchy Possibilistic Clustering Based on the Self-Similarity Principle (DECS) for Enhancing Intrusion Detection System," 2022.
[7] H. Dhrir, M. Charfeddine, N. Tarhouni, and H. M. Kammoun, "Machine learning-and deep learning-based anomaly detection in firewalls: a survey," The Journal of Supercomputing, vol. 81, no. 6, p. 761, 2025.
[8] X. Cheng, "A Comprehensive Study of Feature Selection Techniques in Machine Learning Models," Insights in Computer, Signals and Systems, vol. 1, no. 1, p. 10.70088, 2024.
[9] R. S. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," Journal of educational data mining, vol. 1, no. 1, pp. 3-17, 2009.

**Research Article**

[10] A. M. Shahiri, W. Husain, and N. a. A. Rashid, "A review on predicting student's performance using data mining techniques," procedia computer science, vol. 72, pp. 414-422, 2015.
[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, 2020.
[12] G. Akçapınar, A. Altun, and P. Aşkar, "Using learning analytics to develop early-warning system for at-risk students," International Journal of Educational Technology in Higher Education, vol. 16, no. 1, pp. 1-20, 2019.
[13] R. Conijn, A. Van den Beemt, and P. Cuijpers, "Predicting student performance in a blended MOOC," Journal of Computer Assisted Learning, vol. 34, no. 5, pp. 615-628, 2018.
[14] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," Computers & education, vol. 113, pp. 177-194, 2017.
[15] A. H. Alsaeedi, A. M. Al-juboori, H. H. R. Al-Mahmood, S. M. Hadi, H. J. Mohammed, M. R. Aziz, M. Aljibawi, and R. R. Nuiaa, "Dynamic Clustering Strategies Boosting Deep Learning in Olive Leaf Disease Diagnosis," Sustainability, vol. 15, no. 18, p. 13723, 2023.
[16] K. S. Zaman, M. B. I. Reaz, S. H. M. Ali, A. A. A. Bakar, and M. E. H. Chowdhury, "Custom hardware architectures for deep learning on portable devices: A review," IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 11, pp. 6068-6088, 2021.
[17] G. Siemens and P. Long, "Penetrating the fog: Analytics in learning and education," EDUCAUSE review, vol. 46, no. 5, p. 30, 2011.
[18] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," Wiley interdisciplinary reviews: Data mining and knowledge discovery, vol. 10, no. 3, p. e1355, 2020.
[19] O. Adejo and T. Connolly, "An integrated system framework for predicting students' academic performance in higher educational institutions," International Journal of Computer Science and Information Technology, vol. 9, no. 3, pp. 149-157, 2017.
[20] J. Broadbent and W. L. Poon, "Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review," The internet and higher education, vol. 27, pp. 1-13, 2015.
[21] M. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Multi-split optimized bagging ensemble model selection for multi-class educational data mining," Applied Intelligence, vol. 50, no. 12, pp. 4506-4528, 2020.