AI-Powered Early Detection of Academic Risk Using a Soft Voting Ensemble: A Step Towards Sustainable Learning

Dr. Mazen Al-Zyoud
Department of Computer Science, Al al-Bayt University, Mafraq, Jordan)
alzyoud_mazen@aabu.edu.jo

Ghadeer Sulieman
Department of Computer Science, Al al-Bayt University, Mafraq, Jordan)
2220901006@st.aabu.edu.jo

Haya Al-Hadramy
Department of Computer Science, Al al-Bayt University, Mafraq, Jordan)
2020901008@st.aabu.edu.jo

Dr.Najah Al-shnableh
Department of Computer Science, Al al-Bayt University, Mafraq, Jordan)
najah2746@aabu.edu.jo

Abstract *The increasing complexity of student needs and the ongoing risks of academic failure in higher education demand data-driven strategies to promote sustainable learning. This study presents an AI-based framework designed to personalize academic support and improve predictive performance. After data preprocessing, eight individual classification algorithms—XGBoost, Random Forest, CatBoost, LightGBM, Gradient Boosting, MLP, SVM, KNN, and Naive Bayes—were applied and evaluated. Based on their results, four main algorithms (XGBoost, Random Forest, CatBoost, and LightGBM) were selected to build a soft voting ensemble model that combines their strengths to achieve higher accuracy and stability. The ensemble model demonstrated outstanding performance, achieving an accuracy of 99.12% and a recall of 99.97%, outperforming all individual models. These results validate the efficacy of AI-driven ensemble models for forecasting academic risks early and for promoting more sustainable and efficient learning environments. Ultimately, this study contributes to the creation of more inclusive educational systems by enabling timely interventions, equipping educators to assist students in a proactive manner, and reducing preventable cases of academic failure through intelligent analysis.*

Keywords: *Artificial Intelligence, Academic Performance Prediction, Educational Sustainability, Ensemble Learning, Voting, SMOTE.*

I. INTRODUCTION

Predictive analytics is an essential concept that will enhance education [1–6] in the context of the scientific revolution that majority of schools worldwide are involved in [7, 8]. Predictive analytics is a very effective technique that finds patterns and behaviors in data sets and artificial intelligence algorithms to forecast future results for students in educational settings [9, 10]. Consequently, predictive analytics plays a critical role in assisting educational establishments in knowing the conduct and performance of their students. This is accomplished by examining a variety of statistics, including grades, attendance, and levels of involvement. Additionally, predictive analytics may provide predictions for the future [11– 13]. Additionally, it allows instructors to estimate which students might be at risk of falling behind or dropping out of school. By taking a proactive stance, educational organizations can intervene early and meet the needs of at-risk students before they face these issues [14–17]. For instance, let's say a prediction model indicates that, given a student's present performance or level of interest, they are likely to suffer. Teachers can then offer further assistance in the form of counseling, tutoring, or additional assistance [18-21]. The students' concerns are proactively resolved with the aid of this predictive model. In further than raising individual student outcomes, predictive analytics additionally enhancement the learning environment as a whole and teaches educators how to make strategic decisions that will improve educational processes [21,25]. Predictive analytics, therefore, enhances the strategy of educational organizations, permitting rapid and accurate Potential efforts to efficiently address the requirements of the school and its students, as well as to offer advice and assistance by utilizing data-driven insights.

II. LITERATURE REVIEW

Elhayes [26] suggests using Artificial Neural Network (ANN), Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB) and XGBoost models to forecast student success. The study provides use of a dataset that includes social details, demographic data, student grades, and school-related attributes. The main objective is to evaluate comparing the effectiveness of various models used for machine learning in assigning grades to student performance, for example "Good," "Fair," and "Poor." The results of the study show that when it comes to predicting student performance, XGBoost performs substantially better than ANN, RF, SVM and XGBoost. Based on certain statistics pertaining to students' performance in mathematics and Portuguese, the findings.

Doz et al. [27] examine the adoption of fuzzy logic and machine learning approaches (random forest, gradient boosting regression, and support vector machines) to forecast student performance. A particular dataset involving attendance records, class conduct ratings, and student scores in a variety of disciplines is used in the study. Its primary objective is to create predictive models for student performance, involving grade point average (GPA) prediction, performance categorization, dropout identification, and pass/fail outcomes. Random Forest did a good job of predicting the "Good" category, while SVM's overall accuracy was high. The study nevertheless fails to explicitly address feature selection, outliers, or missing values.

Parkavi et al. [28] analyze methods to predict student performance through exploratory data analysis (EDA) and machine learning methods, particularly multiple linear regression and K-Nearest Neighbors (KNN). The study utilizes the use of specific data set that was gathered from 400 engineering students. The main goal is to forecast student performance and visualize their abilities, with an emphasis on appreciating the effects of changing from traditional classroom instruction to online learning.

Sabbir et al. [29] To enhance the accuracy and effectiveness of student dropout prediction, use and assess the performance of five predictive models: SMOTE, XGBoost, KNN, Decision Tree, and Random Forest. The study used a Kaggle dataset of higher education students that included socioeconomic characteristics, academic pathways, and student information. XGBoost and Random Forest represented the best-performing models. For further improve, the authors recommend investigating hybrid algorithms and inconsistent properties.

Ramirez [30] Establish and evaluate machine learning models (SVM, Random Forest, Gradient Boosting, Neural Networks, and Logistic Regression) to forecast student results and determine essential success operators. Data from 15,000 undergraduate students at a sizable public university in the United States will be gathered for the purpose of the endeavour. Up to 85% of performance predictions and 85.6% of dropout assessments have been accurate.

Chukwuemeka et al. [31] utilize machine learning algorithms (KNN, K-means, Naive Bayes, decision trees, and linear regression) to forecast student academic performance and determine the variables that influence it. The primary concern of the study, which makes use of a dataset gathered from 1000 students in Nigerian schools, is to determine the parameters that affect student rating and create predictive models for classifying students into numerous performance groups. The study found that a student's ethnicity and educational level of parents have an impact on their evaluation.

Albahli [32] provides a strategy that combines Bayesian Optimization for hyperparameter tuning with the Synthetic Minority Oversampling Technique (SMOTE) for controlling irregular data. A dataset of 5,000 student records gathered over five semesters at a variety Saudi Arabian university is utilized in the study. The machine learning methods used in the study for classification are Random Forest and Decision Tree. In contrast to other models, the Decision Tree with SMOTE and Bayesian Optimization performed better.

Gaftandzhieva et al. [33] discusses the utilization of statistical and machine learning techniques to forecast students' final grades based on their involvement in Zoom online lectures and online activities in the Moodle Learning Management System (LMS). 105 students who enrolled in the University of Plovdiv's Object-Oriented Programming program for the 2021–2022 academic year were included in the dataset. For prediction, the study used machine learning algorithms (RF, XGBoost, KNN, SVM). At 78%, the Random Forest algorithm was the highest reliable.

Alsulami et al. [34] provide a model that increases the prediction precision of student performance by combining the ensemble method (Bagging and Boosting) with conventional data mining techniques (Decision Tree, Naive Bayes, and Random Forest). A voting technique is additionally included in the model to further hone predictions. The Kalboard 360 E-Learning system delivered the dataset's 480 records and 17 attributes. The maximum accuracy (77.9%) was obtained by using decision trees for boosting.

Ref	**Year**	**Problem of research**	Models	**Accuracy**
[26] Elhayes	2025	Evaluating machine learning methods for predicting student outcomes.	XGBoost, ANN, RF, SVM, XG Boost. Vaive Bayes	XGBoost, 95%
[27]	2023	Analyzing student performance using machine learning and fuzzy logic.	Random forest, logistic regression, fuzzy logic, SVM.	LR:0.84 RF:0.94, SVM :0.98
[28]	2024	Predicting student performance using EDA and ML techniques.	KNN, multiple regression, EDA.	99% (multiple regression)
[29]	2024	Using predictive analytics to reduce dropout rates.	Random forest, XGBoost, KNN, Decision Tree.	0.99 AUC (Random Forest)
[30]	2023	Predicting student performance and dropout rates in higher education.	Logistic regression, random forests, neural networks.	0.85
[31]	2023	Early prediction of low-performing students.	Linear regression, decision tree, Naïve Bayes, KNN, KMeans.	SVM : 95%
[32]	2023	Improving prediction accuracy using Bayesian optimization for imbalanced data.	Bayesian Optimization, SMOTE, Random Forest, Decision Tree.	Random Search: 95 % Improved with Bayesian Optimization
[33]	2022	Open issue of prediction accuracy in student performance classification.	Stacking ensemble using RF, DT, AdaBoost, SVM, Logistic Regression.	RF.78%
[34]	2023	Challenges in analyzing e-learning data and improving prediction accuracy.	Bagging, boosting, and voting on DT, Naive Bayes, RF.	77% (Boosting)

III. METHODOLOGY

In an effort to address academic performance disparities and mitigate dropout risks in higher education, this study adopted a data-driven artificial intelligence framework designed to personalize learning and foster educational sustainability. The methodology was implemented in two interconnected phases: the first phase focused on organizing and cleaning the data, while the second phase focused on training and evaluating classification models, with an emphasis on creating a robust ensemble model.

The dataset, which was obtained from Kaggle in the first phase under the title Students Performance in Exams, included 1,000 student records and nine features that included academic, socioeconomic, and demographic characteristics. The preprocessing involved creating a binary target variable, performance, indicating whether a student failed any of the three subjects (math, reading, or writing). Additional features, such as has failed and average score, were developed to provide more context to the dataset. After categorical variables were transformed using One-Hot Encoding, SMOTE was applied to equalize the class distribution. Synthetic data was generated using make classification and merged with the existing dataset to enhance the diversity of the sample space. Finally, all features were standardized via Standard Scaler to ensure consistent scaling. In the second phase, a diverse array of individual classification algorithms was trained, including tree-based and advanced models such as XGBoost, Random Forest, CatBoost, and LightGBM, in addition to the Multi-Layer Perceptron (MLP). Additionally included were conventional models like SVM, KNN, and Naive Bayes. The purpose of these algorithms was to guarantee diversity in learning mechanisms, which is essential for boosting the potency of ensemble tactics. The pivotal step in this phase was the development of a Voting Ensemble model using soft voting, whereby each individual model's probabilistic output was weighted according to its prior performance. This method aims to maximize the strengths of individual classifiers while minimizing their biases. The top four models—XGBoost, LightGBM, CatBoost, and Random Forest—were all part of the ensemble. By utilizing a weighted aggregate, this ensemble method enhanced the stability, accuracy, and generalizability of classification, rendering it highly beneficial in real-world educational contexts. Figure 1 shows a description of the methodology employed by the study,

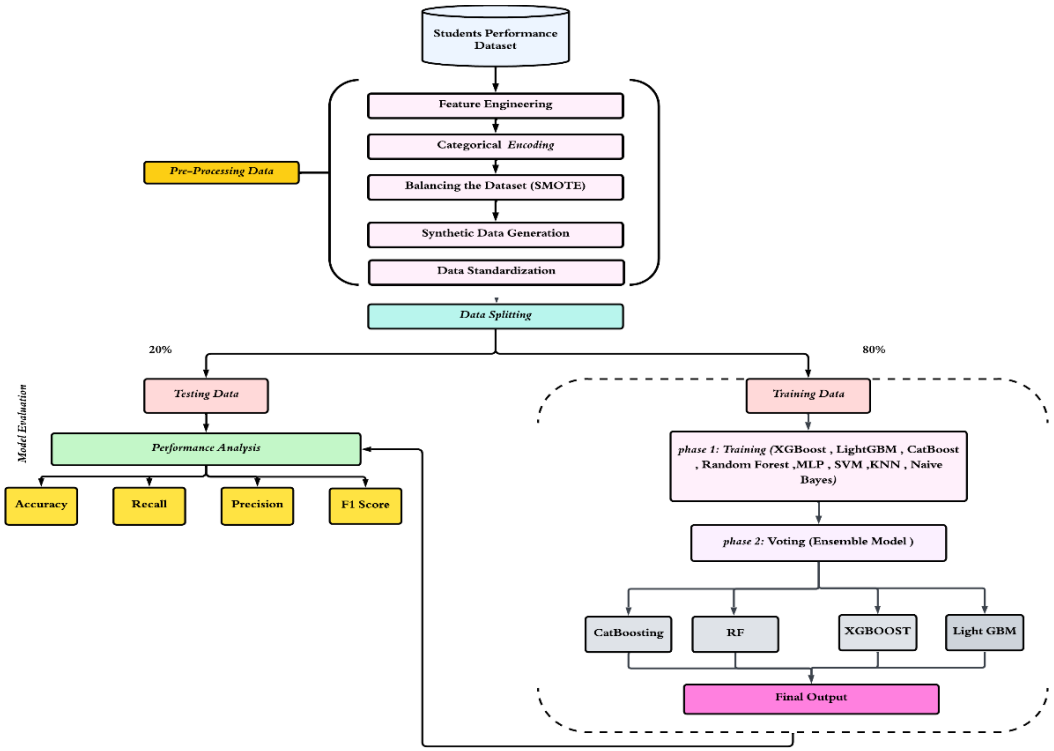

Figure 1 Methodology Flowchart

IV. RESULT

Table 1 offers a comprehensive comparison of various individual classification algorithms, employing metrics such as F1-score, Accuracy, ROC AUC, Precision, and Recall. Despite this, more advanced tree-based models such as XGBoost, Random Forest, and CatBoost surpassed the others, each reaching an accuracy greater than 97%, while traditional models like Naive Bayes and KNN showed subpar performance. Tree-based models were selected for the final ensemble due to their consistently outstanding performance, although SVM and MLP produced results that were also quite satisfactory.

Performance comparison for all models

	Model	Accuracy	ROC AUC	Precision	Recall	F1-score
0	XGBoost	0.980400	0.997300	0.974700	0.991400	0.982900
1	RandomForest	0.979400	0.997100	0.973100	0.991400	0.982100
2	CatBoost	0.978000	0.997200	0.971500	0.990500	0.980800
3	LightGBM	0.976500	0.996900	0.970700	0.988800	0.979500
4	GradientBoosting	0.976000	0.997400	0.972200	0.986200	0.979000
5	MLP	0.972600	0.992900	0.972200	0.980100	0.976000
6	SVM	0.911900	0.938600	0.881000	0.976700	0.926300
7	KNN	0.884000	0.953100	0.905700	0.889400	0.897000
8	NaiveBayes	0.759600	0.845800	0.840800	0.710800	0.769400

Table 2 highlights the final performance of the Voting Ensemble model, which integrated four powerful classifiers: XGBoost, LightGBM, CatBoost, and Random Forest. This ensemble exhibited outstanding performance across all metrics, achieving an almost perfect recall rate of 99.97% and an accuracy of 99.12%. This highlights the model's ability to identify nearly all students who have failed. These findings underscore the efficacy of the probabilistic soft voting method in improving the robustness and applicability of classification, surpassing the capabilities of individual models.

	Accuracy	ROC AUC	Precision	Recall	F1-score
Voting (CatBoost, LightGBM, Random Forest and XGBoost)	0.991200	0.999900	0.988900	0.999700	0.998500

The results of this study demonstrated the high effectiveness of an AI-driven approach in classifying student performance, with clear superiority shown by the ensemble model. The Voting Ensemble, which combined the top five individual models, achieved an outstanding accuracy of 99.12% and an exceptional ROC AUC score of 0.9999—surpassing all standalone models. This superior performance is attributed to the model's ability to merge diverse learning paradigms, enhancing prediction reliability while minimizing individual biases. These findings underscore the value of weighted ensemble models in real-world educational settings aimed at providing smart and proactive interventions to mitigate academic failure risks.

Figure 2 visually demonstrates the accuracy of all individual classification models along with the final Voting Ensemble model used in this study. This bar chart clearly shows the effectiveness of the developed AI framework, especially the strong performance of the Voting Ensemble, in accurately identifying academic risks and revealing student performance gaps.

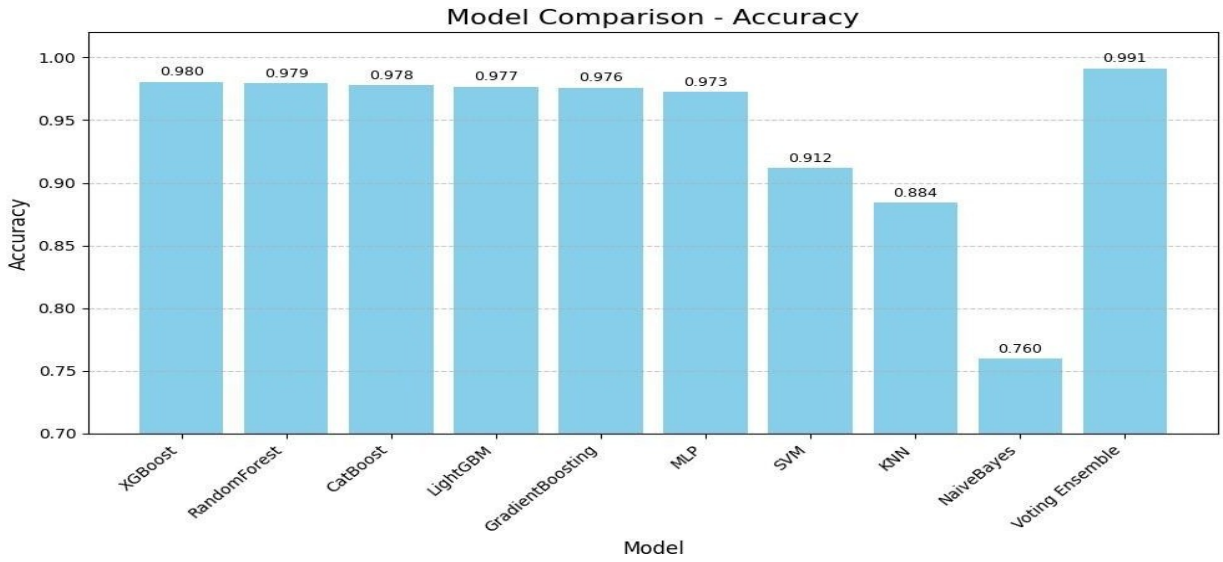

Figure 2 Comparison of Classification Model Accuracy in Predicting Student Academic Performance.

This figure3 clearly illustrates the discriminative power of the various models employed. Notably, the Voting Ensemble achieves an exceptional ROC AUC of 1.000, signifying its near-perfect ability to distinguish between student performance outcomes. High-performing tree-based models (XGBoost, RandomForest, CatBoost, LightGBM, GradientBoosting) also demonstrate robust discrimination with AUC values around 0.997. While conventional models show lower performance, this curve unequivocally underlines the ensemble's superior capability in accurately identifying academic risk, reinforcing its value for proactive intervention.

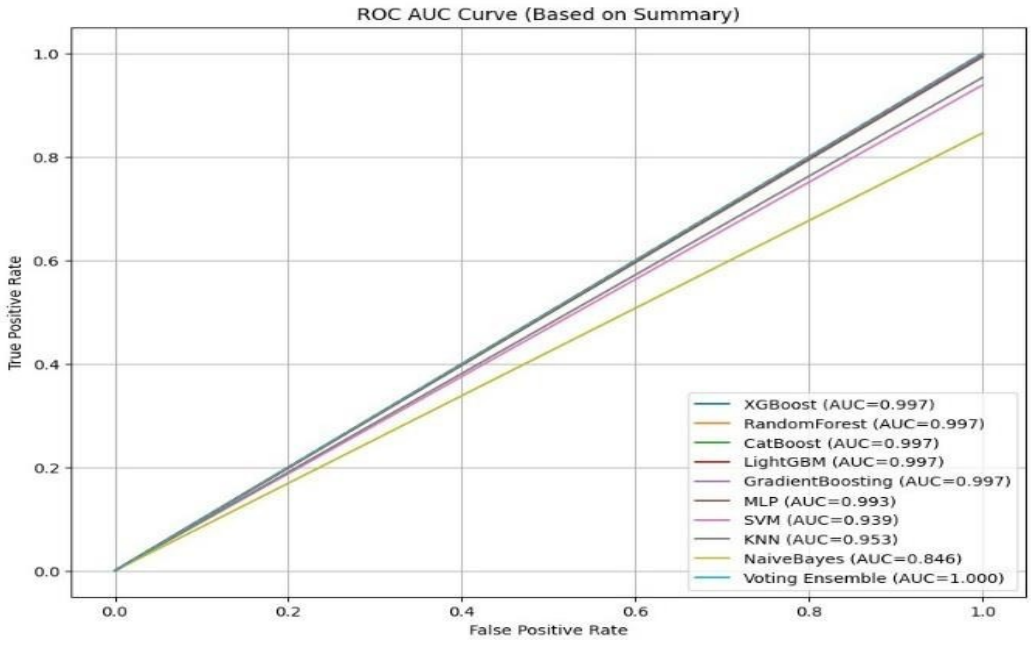

Figure 3 ROC AUC Curve for Classification Models.

V. REFERENCES

1- Eckerson, W. W. (2007). Predictive analytics. *Extending the Value of Your Data Warehousing Investment. TDWI Best Practices Report*, *1*, 1-36.

2- McCarthy, R. V., McCarthy, M. M., Ceccucci, W., Halawi, L., McCarthy, R. V., McCarthy, M. M., ... & Halawi, L. (2022). *Applying predictive analytics* (pp. 89-121). Springer International Publishing.

3- Alsaeedi, A. H., Hadi, S. M., & Alazzawi, Y. (2024). Adaptive Gamma and Color Correction for Enhancing Low-Light Images. *International Journal of Intelligent Engineering & Systems*, *17*(4).

4- Alsaeedi, A. H., Al-Mahmood, H. H. R., Alnaseri, Z. F., Aziz, M. R., Al-Shammary, D., Ibaida, A., & Ahmed, K. (2024). Fractal feature selection model for enhancing high-dimensional biological problems. *BMC bioinformatics*, *25*(1), 12.

5- Basha, S. J., Rao, A. K., & Ammannamma, T. (2025). Transforming Education With Predictive Analytics: A Data-Driven Approach to Student Achievement. In *Driving Quality Education Through AI and Data Science* (pp. 433-456). IGI Global Scientific Publishing.

6- Al-Shammary, D., Radhi, M., Alsaeedi, A. H., Mahdi, A. M., Ibaida, A., & Ahmed, K. (2024). Efficient ECG classification based on the probabilistic Kullback-Leibler divergence. *Informatics in medicine unlocked*, *47*, 101510.

7- Douglass, J. (2007). The global higher education race. *International Higher Education*, (49).

8- Alsaeedi, A. H., Al-Shammary, D., Hadi, S. M., Ahmed, K., Ibaida, A., &AlKhazraji, N. (2024). A proactive grey wolf optimization for improving bioinformatic systems with high dimensional data. *International Journal of Information Technology*, *16*(8), 4797-4814.

9- Y. Huang, O. H. Lu, J. C. Huang, C. Yin, and S. J. Yang, "Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs," Interactive Learning Environments, vol. 28, no. 2, pp. 206-230, 2020.

10- A. H. Alsaeedi, M. A. Al-Sharqi, S. S. Alkafagi, R. R. Nuiaa, A. S. D. Alfoudi, S. Manickam, A. M. Mahdi, and A. M. Otebolaku, "Hybrid extend particle swarm optimization (EPSO) model for enhancing the performance of MANET routing protocols," Journal of Al-Qadisiyah for computer science and mathematics, vol. 15, no. 1, pp. Page 127-136, 2023.

11- Renick, T. M. (2020). Predictive Analytics, Academic Advising, Early. *Big data on campus: Data analytics and decision making in higher education*, 177.

12- R. R. Nuiaa, S. A. A. A. Alsaidi, B. K. Mohammed, A. H. Alsaeedi, Z. A. A. Alyasseri, S. Manickam, and M. A. Hussain, "Enhanced PSO Algorithm for Detecting DRDoS Attacks on LDAP Servers," International Journal of Intelligent Engineering & Systems, vol. 16, no. 5, 2023.

13- Al Ogaili, R., Alsaeedi, A. H., Alkafagi, S. S., &Alfoudi, A. S. D. (2022). A critical review of Optimization MANET routing protocols. *Wasit Journal of Computer and Mathematics Science*, *1*(4), 44-54.

14- Foster, E., & Siddle, R. (2020). The effectiveness of learning analytics for identifying at-risk students in higher education. *Assessment & Evaluation in Higher Education*, *45*(6), 842-854.

15- Nuiaa, R. R., Manickam, S., Alsaeedi, A. H., &Jabor Al-Shammary, D. E. (2022). Evolving Dynamic Fuzzy Clustering (EDFC) to Enhance DRDoS_DNS Attacks Detection Mechnism. *International Journal of Intelligent Engineering & Systems*, *15*(1).

16- Nuiaa, R. R., Manickam, S., &ALsaeedi, A. H. (2022). A Comprehensive Review of DNS-based Distributed Reflection Denial of Service (DRDoS) Attacks: State-of-the-Art. *Int. J. Adv. Sci. Eng. Inf. Technol*, *12*(6), 2452-2461.

17- Hadi, S. M., Alsaeedi, A. H., Nuiaa, R. R., Manickam, S., &Alfoudi, A. S. D. (2022). Dynamic Evolving Cauchy Possibilistic Clustering Based on the Self-Similarity Principle (DECS) for Enhancing Intrusion Detection System. *International Journal of Intelligent Engineering & Systems*, *15*(5).

18- Bai, X., Zhang, F., Li, J., Guo, T., Aziz, A., Jin, A., & Xia, F. (2021). Educational big data: Predictions, applications and challenges. *Big Data Research*, *26*, 100270.
19- Gharkan, H. K., Radif, M. J., & Alsaeedi, A. H. (2025). Analysis of AI-Empower Predictive Models for Predicting Student Performance in Higher Education. *Journal of Al-Qadisiyah for Computer Science and Mathematics*, *17*(1), 103-121.
20- Parivara, S. A. (2025). Leveraging Data Analytics for Enhanced. *Impacts of AI on Students and Teachers in Education 5.0*, 349.
21- Awashreh, R., & Hassiba, A. (2025). Revolutionizing education with AI: personalized learning, predictive analytics, and gamification. In *Insights Into Digital Business, Human Resource Management, and Competitiveness* (pp. 149-170). IGI Global Scientific Publishing.
22- Ashaari, M. A., Singh, K. S. D., Abbasi, G. A., Amran, A., & Liebana-Cabanillas, F. J. (2021). Big data analytics capability for improved performance of higher education institutions in the Era of IR 4.0: A multi-analytical SEM & ANN perspective. *Technological Forecasting and Social Change*, *173*, 121119.
23- Janahi, Y., & Obeidat, M. (2025). Predictive Analytics and AI in Education: A Systematic Literature Review on Identifying and Supporting At-Risk Students. *Available at SSRN 5240920*.
24- Albukhnefis, A. L., Al-Fatlawi, T. T., &Alsaeedi, A. H. (2024). Image Segmentation Techniques: An In-Depth Review and Analysis. *Journal of Al-Qadisiyah for Computer Science and Mathematics*, *16*(2), Page-195.
25- Mohamed Hashim, M. A., Tlemsani, I., & Matthews, R. (2022). Higher education strategy in digital transformation. *Education and information technologies*, *27*(3), 3171-3195.
26- Elhayes, M. A. (2025). Forecasting students Grades based on students' performance using machine learning techniques. *Artificial Intelligence Information Security*, *3*(8), 32-56.
27- Doz, D., Cotič, M., & Felda, D. (2023). Random forest regression in predicting students' achievements and fuzzy grades. *Mathematics*, *11*(19), 4129.
28- Parkavi, R., Karthikeyan, P., Sujitha, S., & Abdullah, A. S. (2024). Enhancing Educational Assessment: Predicting and Visualizing Student Performance using EDA and Machine Learning Techniques. *Journal of Engineering Education Transformations*, 240-245.
29- Sabbir, W., Abdullah-Al-Kafi, M., Afridi, A. S., Rahman, M. S., & Karmakar, M. (2024, February). Improving predictive analytics for student dropout: A comprehensive analysis and model evaluation. In *2024 11th International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 951-956). IEEE.
30- Esomonu, N. P. M. (2025). Utilizing AI and Big Data for Predictive Insights on Institutional Performance and Student Success: A Data-Driven Approach to Quality Assurance. *AI and Ethics, Academic Integrity and the Future of Quality Assurance in Higher Education*, *29*.
31- Ramírez, J. G. C. (2024). Predictive analytics in education: utilizing machine learning to forecast student performance and dropout rates. *Asian American Research Letters Journal*, *1*(5).
32- Albahli, S. (2024). Efficient hyperparameter tuning for predicting student performance with Bayesian optimization. *Multimedia tools and applications*, *83*(17), 52711-52735.29
33- Gaftandzhieva, S., Talukder, A., Gohain, N., Hussain, S., Theodorou, P., Salal, Y. K., & Doneva, R. (2022). Exploring online activities to predict the final grade of student. *Mathematics*, *10*(20), 3758.
34- Yakubu, M. N., & Abubakar, A. M. (2022). Applying machine learning approach to predict students' performance in higher educational institutions. *Kybernetes*, *51*(2), 916-934.