Global Journal of Engineering and Technology Advances
eISSN: 2582-5003
Cross Ref DOI: 10.30574/gjeta
Journal homepage: https://gjeta.com/

(Research Article)

Machine learning based misconduct pattern detection and prevention system in sustainable communities (Campus hostels)

Sunday Oluwadare Oladipo, Omotola Faith Towolawi *, Alli Ademola Akinpelu and Ajibola Abiola Folahan

Department of Computer Science, School of Computing, Babcock University, Ilishan-Remo, Ogun State, Nigeria.

Global Journal of Engineering and Technology Advances, 2026, 27(03), 001-008

Publication history: Received on 22 April 2026; revised on 31 May 2026; accepted on 02 June 2026

Article DOI: https://doi.org/10.30574/gjeta.2026.27.3.0131

Abstract

Campus hostels are environments where students live, interact with people and develop socially. However, campus hostels are also spaces where various forms of misconduct such as theft, drug use and many more can occur and such misconducts affects student safety and disrupt the overall well-being of students and traditional methods of predicting misconduct have proved to be less effective due to their inability to handle complex data patterns. This project focuses on the development of a machine learning-based system for detecting and preventing misconduct in campus hostels. The system made use of three machine learning algorithms, which are Linear Regression, Random Forest and XGBoost, to analyze past misconduct data and identify patterns for predictions. Data preprocessing techniques such as encoding and normalization were carried out to improve model performance and accuracy. The findings showed that XGBoost performed best with an R-Squared Score of over 93%, Mean Squared Error of 1.16 and Root Mean Squared Error of 1.08, indicating strong predictive capability. The system was interfaced into a web-based platform to allow administrators receive predictions. The system provides features like data visualizations, analytical dashboards, and report generation to help administrators in decision making. This project shows that machine learning can improve misconduct detection and promote safer hostel environment for students.

Keywords*:* Campus Safety; Machine Learning; Misconduct Detection; Pattern Detection; Predictive Analysis

1. Introduction

Conventional methods of detecting misconduct are usually manual, surveillance systems or rule-based mechanisms. These methods can only identify glaring misconducts and are usually ineffective in detecting complex or recurring patterns [1]. These limitations have led to the exploration of more advanced technologies, focusing mainly on those based on machine learning and artificial intelligence [2], [3]. As a result, universities struggle to proactively address these misconducts before it becomes a bigger problem. In recent times, machine learning has emerged as a very powerful tool for analyzing large datasets and identifying hidden patterns. Machine learning is a branch of artificial intelligence that allows systems to learn from data and make predictions without being programmed [4]. Its ability to process large volume of data makes it very useful in areas such as anomaly detection and predictive analysis. The application of machine learning in education has gained attention. Researchers have shown that machine learning algorithms can effectively predict student performance, detect anomalies and identify cases of academic misconduct [5]. Studies have shown that the prevalence of academic misconduct has increased greatly, particularly with the rise of digital learning technologies and online assessments [6]-[8]. For example, Decision Trees, Random Forest and Support Vector Machines have been used to detect cheating behaviors and predict the likelihood of misconduct [9]. In addition, the rise of artificial intelligence tools has greatly increased the occurrence of misconduct that is difficult to detect by using traditional methods such as manually scanning through records [10], [11].

* Corresponding author: Towolawi Omotola Faith.

One of the key advantages of machine learning is its ability to predict. Machine learning models can identify trends and predict future behaviors, by learning from historical data. This allows institutions take proactive measures [12], [13]. This is very important in campus hostels where misconduct happen often and can go undetected due to limited supervision and inadequate monitoring systems. Despite the growing body of research on machine learning applications in education, there is limited focus on its use in detecting misconduct within campus hostels. Most existing studies concentrate more on academic dishonesty in classroom or online settings, leaving a gap in addressing misconduct in environments where students live [5], [8], [14]. Also, many institutions still rely on reactive approaches which are after the incident has happened, rather than predictive, data driven systems.

In addition to detection challenges, recent studies have highlighted the effect of behavioral and environmental factors in influencing misconduct among students. Factors such as peer influence, academic pressure, and opportunity have been identified as major factors that contribute to dishonest practices in higher institutions. For instance, studies conducted among Nigerian undergraduates reveal a high prevalence of academic dishonesty linked to environmental and social influences [15], [16]. Also, a research on campus criminality shows that students in campus hostels are more vulnerable to misconduct due to weak supervision and increased interaction among peers [17]. Past studies have also shown that inadequate monitoring and poor management lead to the rise of theft in campus hostels [18].

To address these challenges, this project proposes a machine learning-based misconduct pattern detection and prevention system for campus hostels. The system uses historical data to identify recurring pattens of misconduct and predict future happenings. By integrating predictive analysis with a user-friendly interface, the system aims to support administrators in making informed decisions and implementing timely interventions. This project contributes to the advancement of intelligent systems in educational environments by demonstrating how machine learning can be used to improve safety, enhance monitoring, and promote ethical behavior within campus hostels.

2. Methodology

This project follows a structured, data-driven approach for the development of a machine learning-based system aimed at detecting and preventing misconduct patterns in sustainable communities, particularly campus hostels. The methodology includes data collection, data preprocessing, system design, and implementation.

2.1. Data Collection

Secondary data was used in this work. This was gotten from the Security Unit of the University. The dataset consisted of records of misconduct incidents across all hostels over a period of ten years (2019 to 2025 sessions). Each record contained the type of misconduct, semesters, year of occurrence of misconduct, hostel names, and the number of student(s) involved. These records are contained in separate session files and were merged into a single dataset to enable analysis of trends and patterns over time. The snapshot of the raw data is shown in figure 1 with attributes such as misconduct type and hall names.

	Misconduct_Type	WELCH HALL	WINSLOW HALL	BETHEL HALL	NELSON M HALL	SAMUEL. A HALL	NEAL WILSON HALL	GIDEON TROOPERS HALL	ADELEKE HALL	HAVILAH HALL	FAD HALL	WHITE HALL	NYBERG HALL	OGDEN HALL	MARIGOLD HALL	CRYSTAL HALL
0	Theft/Missing Items	14	2	5	17	3	15	9	1	1	0	2	0	4	2	0
1	Use/Possession of Drugs	18	7	4	28	2	24	4	0	1	1	1	0	3	0	0
2	Lurking/Pair Loitering	10	2	3	8	4	10	7	1	15	2	2	3	0	3	8
3	Bullying/Gangsterism	0	1	2	6	1	1	0	0	0	0	0	0	0	0	1
4	Exeat Violation	9	3	1	3	0	6	0	0	1	1	2	0	1	0	7
5	Sabbath/Worship Violation	12	7	19	12	5	25	1	0	0	0	4	1	0	2	3
6	Other Violations	4	0	7	5	1	8	5	0	6	1	1	0	5	1	1
7	Breach of Contract	2	3	0	3	0	1	1	0	1	0	0	0	0	0	0
8	Disturbance	3	1	2	0	2	0	0	0	0	1	0	0	0	0	0
9	Fighting/Affray	2	0	2	0	0	3	1	0	2	0	0	0	0	0	0
10	Illicit Transaction	1	2	0	1	0	3	3	0	1	1	2	0	0	3	0

Figure 1 Snapshot of Raw Data

2.2. Data Preprocessing

Data preprocessing and cleaning was embarked on to ensure the quality and reliability of the dataset. The collected data contained some missing values, inconsistencies, and duplicate entries which were addressed during this stage. Missing values were handled by using substitution with mean values and estimation based on the data. Numerical data was also normalized to ensure uniformity and this was done by using the Min-Max Scaling technique, which transforms values into a consistent range between 0 and 1. Furthermore, categorical data such as hostel names and type of misconduct were converted into numerical format using One-Hot encoding. The issue of imbalanced data was also addressed, as some types of misconduct occur more frequently than the others using data-sampling and weighting techniques to ensure that the models do not act biased towards more frequent misconducts. Furthermore, the dataset was subjected further analysis and classification which included sorting the data into categories and identifying key patterns such as frequently occurring misconduct types, high risk level hostels, semesters of misconduct.

2.3. Model Development

Following the labelling of the dataset, machine learning algorithms were used which included Linear Regression, Random Forest and Extreme Gradient Boosting (XGBoost). The models were selected due to their ability to handle structured data and capture complex relationships between variables. The dataset was divided into training (80%) and testing sets (20%). The training data is used to enable the models learn while the testing data is used to evaluate their performance on unseen data. The model performance was evaluated using R-Squared Score, Mean Squared Error and Root Mean Squared Error to determine how accurately the models can predict misconducts. R-Squared Score is a statistical measure that indicates the goodness of fit of a regression model. The value lies between 0 and 1, which is 0 to 100%. When the R-Squared Score equals to 1, it means the model perfectly fits the data and there is no difference between the predicted and actual value. However, if the R-Squared Score equals to 0, it means the model does not predict any variability in the model. The Mean Squared Error (MSE) analyzes the accuracy of the model by measuring its average squared difference between predicted values and the actual values in the dataset. A lower MSE indicates that the model's predictions are closer to the actual values, which signifies better accuracy. The Root Mean Squared Error (RMSE) is the measure of how well a regression line fits the data points, and a low RMSE indicates that the model fits the data well and has more precise predictions.

2.4. System Design

The system as depicted in figure 2 was divided into three layers (the data layer, machine layer and application layer). The data layer is responsible for data collection and preprocessing, the machine layer handles model training and prediction and the application layer provides an interface for administrators to interact with the system. The system follows the System Development Life Cycle, which includes requirement gathering, system analysis, design, coding, testing implementation and maintenance. This ensures the system is designed in an efficient matter.

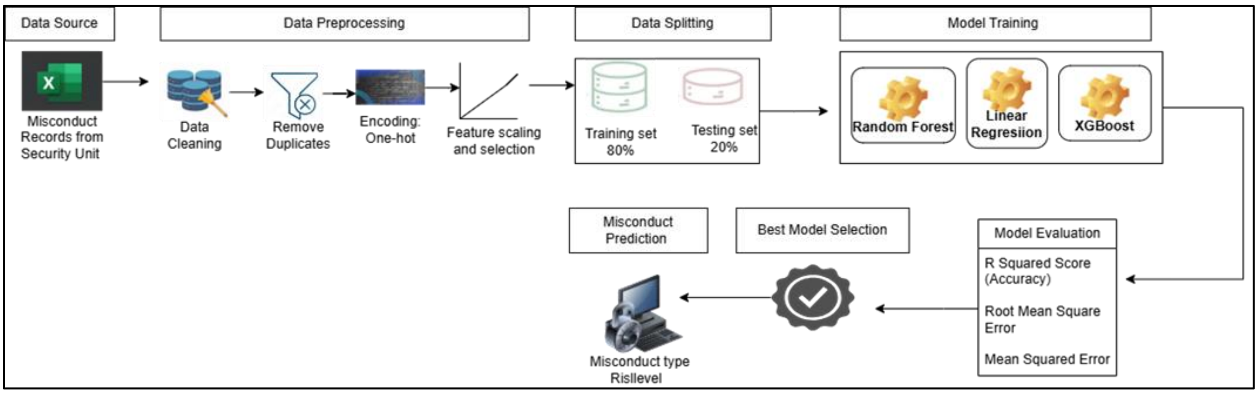

Figure 2 System Architecture Diagram

2.5. Implementation

During the implementation stage, the most efficient machine learning model (XGBoost) was integrated into a software-based platform as shown in figure 3. The system was designed to include features such as incident logging, automated pattern recognition, prediction generation, and report generation. These features enabled administrators to monitor misconduct trends and make informed decisions. Finally, the system operates as a continuous learning cycle, where new data can be added to improve the accuracy and performance of the models. Based on the identified patterns, the system highlights high risk halls and misconduct count among other things implemented.

3. Results and discussion

Rigorous evaluation was essential to verify that the trained model could properly predict misconduct patterns rather than simply memorizing the training data which could result in overfitting.

Table 4.1 shows a negative R Squared value (-0.0002) for Linear regression which indicates that the model performed poorly and was unable to capture the relationship between variables effectively. It also has a high root mean squared error and mean squared error which indicates that the model has high prediction errors and is not accurately capturing the relationship between variables.

Table 1 Linear Regression performance evaluation

Metrics	Values
R Squared Score	-0.0002
Root Mean Squared Error	4.12
Mean Squared Error	16.93

Table 4.2 shows a R Squared Score of 7.88%, an improvement in the Random Forest model. It also has a better root mean squared error and mean squared error, showing that the model captured the relationship between variables but can only make 7.88% of accurate predictions. The model showed slight improvement over linear regression but still had low predictive performance.

Table 2 Random Forest performance evaluation

Metrics	Values
R Squared Score	0.0788
Root Mean Squared Error	3.95
Mean Squared Error	15.60

Table 4.3 shows a R Squared Score of 93.16%, and low root mean squared error and mean squared error using XGBoost, making it a very reliable model. XGBoost outperformed other models with showing strong predictive accuracy and minimal error.

Table 3 XGBoost performance evaluation

Metrics	Values
R Squared Score	0.9316
Root Mean Squared Error	1.08
Mean Squared Error	1.16

This comparative analysis confirmed that XGBoost was the right choice for this application, and selected as the optimal algorithm. Table 4.4 show that XGBoost performed the best, with the highest R Squared Score (93%) and the lowest error values (RMSE = 1.08, MSE = 1.16) which indicate accurate predictions. Random Forest had a low R Squared Score (7%) and higher error values (RMSE = 3.95, MSE = 15.60) which means it is less effective in the prediction. However, Linear Regression performed the worst, with a negative R Squared Score (-0.0002) and the highest errors (RMSE = 4.12, MSE = 16.93) showing that it could not properly model the relationship in the dataset.

Table 4 Comparative Performance of all the models

Model	R Squared Score	Root Mean Squared Error	Mean Squared Error
Linear Regression	-0.0002	4.12	16.93
Random Forest	0.0788	3.95	15.60
XGBoost	0.9316	1.08	1.16

The best performing model was deployed using a web-based system with the following interfaces. Figure 3 depicts the interface for the admin or security officer to login based on their roles.

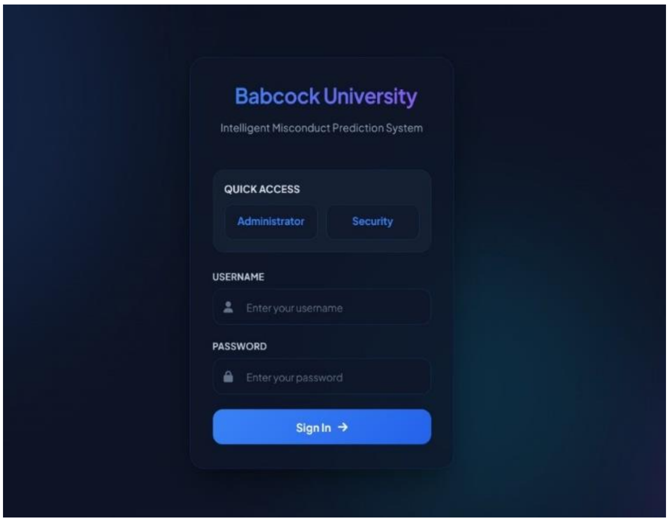

Figure 3 Login Page (Admin or Security)

The admin is brought to the dashboard as seen in figure 4. It shows the total number of halls, predictions and high-risk halls after predictions have been made. It basically gives an overview of the system.

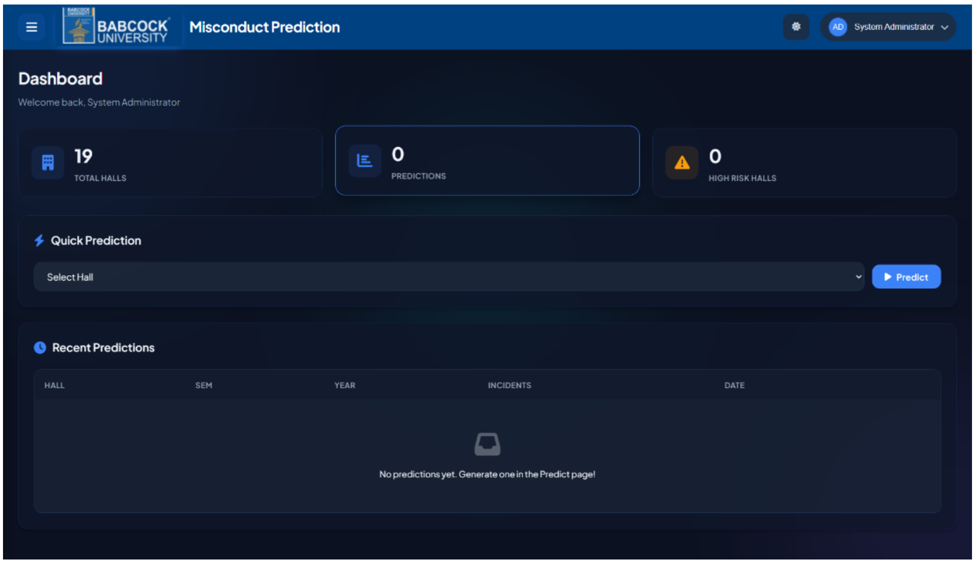

Figure 4 The Admin dashboard

Figure 5 is the make prediction page that is visible to the security officers or any other assigned stakeholder. It allows the stakeholders to make predictions by selecting the hall, semester and year, then click on generate to view the results.

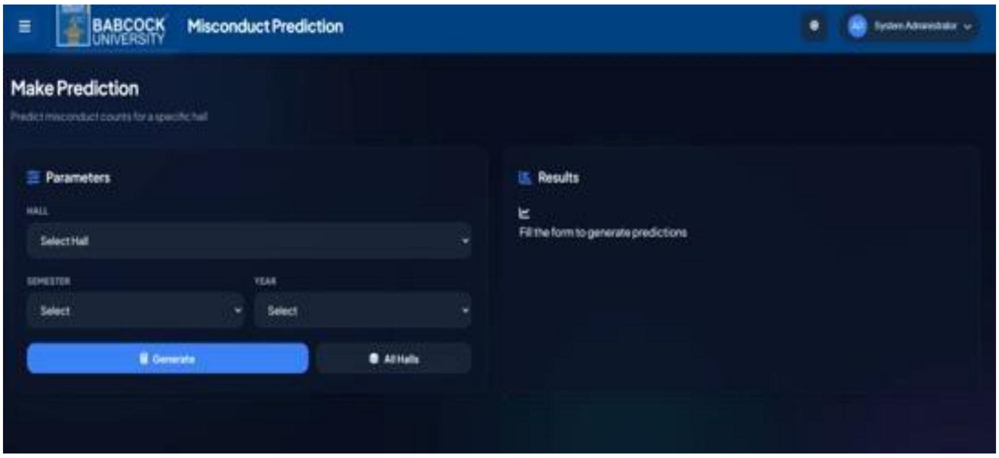

Figure 5 The prediction Page

Following the use of the prediction platform, the result is usually displayed in the form seen in figure 6. It shows the selected hall and the total predicted incidents with a chart that breakdowns the different misconduct types and count. It also shows the risk level of each misconduct type.

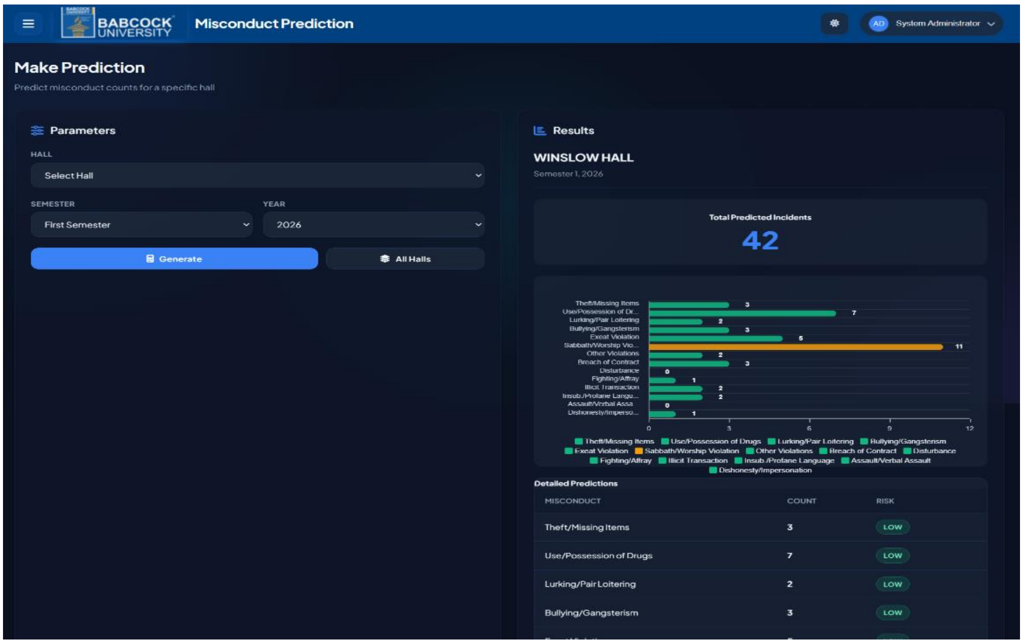

Figure 6 Sample result of prediction

Figure 7 depicts the history of the previous predictions that have been made. It shows the hall name, semester, year, number of incidents and the date. Administrators can also view or delete records.

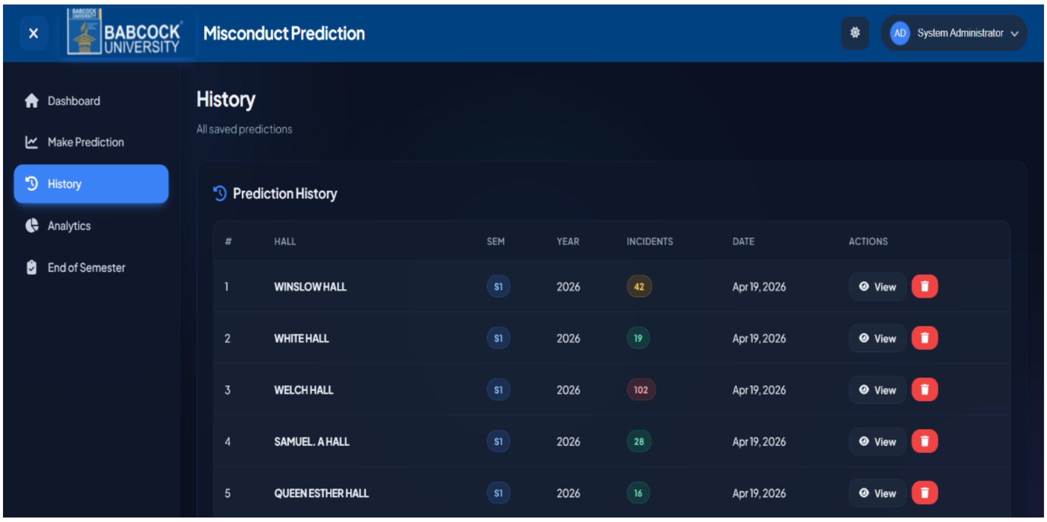

Figure 7 History of previous predictions

The results showed that machine learning can be a practical and effective tool for addressing misconducts on campus, turning years of incident records into something useful for administrators. The system addressed the limitation of traditional methods by moving from reactive to proactive approaches. The model's performance depends on the quality of historical data and requires continuous update to remain effective. Overall, the project demonstrates that predictive, data-driven approaches can enhance misconduct detection, improve monitoring, and support proactive decision-making in educational institutions.

4. Conclusion

It is clear that machine learning is amenable to playing practical role managing misconduct in university hostels. It makes incident pattern detection possible based on years of recorded incident data made. Bringing machine learning into the way hostel misconduct is managed is a practical step towards smarter and more effective campus administration.

Of the three models that were tested, XGBoost outperformed Linear Regression and Random Forest as it handled the complexity of the data better and had better prediction metrics across different hostels and time periods.

Furthermore, the developed system effective in that stakeholders such as governments, security operatives, hostel administrators and the community can use it to get ahead of potential problems rather than always reacting after incident. This shift in approach can make a difference, whether it is in how security is planned, how staff and resources are deployed, or simply in keeping the campus environment safer for students.

Compliance with ethical standards

Acknowledgements

The authors express their sincere gratitude to the Department of Computer Science, School of Computing, Babcock University, Ilishan-Remo, Ogun State, Nigeria and all individuals who provided support and valuable contributions throughout the course of this research.

Disclosure of conflict of interest

No conflict of interest to be disclosed.

Statement of ethical approval

This research was conducted in accordance with accepted ethical principles for academic research. The study focused on the development of a machine learning based misconduct pattern detection and prevention system for campus hostels. All data used for analysis were handled responsibly and solely for research purposes. No procedures involving physical intervention, experimentation on human subjects, or animal testing were conducted. Confidentiality, privacy and data protection considerations were observed throughout the study, and the research was carried out in a way that promotes fairness, integrity, and responsible use of artificial intelligence technologies.

Statement of informed consent

The study did not involve direct interaction with human participants, clinical investigations or the collection of personally identifiable number. Also, formal informed consent was not required. Where data were obtained from institutional records, dataset from the security unit, such data were anonymized and used strictly for academic and research purposes. The authors ensured that no individual could be identified from the data used in this study.

References

[1] D. J. Hand, "Classifier technology and the illusion of progress," *Statistical Science*, vol. 21, no. 1, pp. 1–14, 2006.

[2] W. Alsabhan, "Student cheating detection in online exams using machine learning," *Sensors*, vol. 23, no. 8, pp. 1–17, 2023.

[3] A. Maguya, "Machine learning approach for quantifying academic misconduct," *Acta Infologica*, vol. 8, no. 2, pp. 145–158, 2024

[4] T. M. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.

[5] 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.

[6] F. M. Comas-Forgas and J. Sureda-Negre, "Academic plagiarism: Explanatory factors from students' perspective," *Journal of Academic Ethics*, vol. 10, no. 3, pp. 217–232, 2012.

[7] A. Yusuf, M. Walters, and K. Khan, "Generative artificial intelligence and the future of higher education," *International Journal of Educational Technology in Higher Education*, vol. 21, no. 1, pp. 1–18, 2024.

[8] T. Paustian and B. Slinger, "Students' use of large language models in higher education," *Frontiers in Education*, vol. 9, pp. 1–12, 2024.

[9] F. Kamalov, H. Sulieman, and D. Calonge, "Machine learning based approach to exam cheating detection," *PLOS ONE*, vol. 16, no. 8, 2021.

[10] A. Amirzhanov, R. Mustafin, and D. Kydyraliev, "Plagiarism detection methods: A systematic review," *Frontiers in Computer Science*, vol. 7, pp. 1–20, 2025.

[11] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2021.

[12] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[13] A. Holmes, J. Bialik, and C. Fadel, "Artificial intelligence in education: Promise and implications," *Education Sciences*, vol. 13, no. 3, pp. 1–16, 2023.

[14] L. Tiong and H. Lee, "E-cheating detection using deep learning techniques," *Computers & Education: Artificial Intelligence*, vol. 4, pp. 1–10, 2023.

[15] O. A. Jacob, O. Ayinde, and A. J. Jacob, "Prevalence of academic dishonesty as a form of corrupt practices among university undergraduates in Kwara State, Nigeria," *Nigerian Journal of Social Studies*, vol. XXI, no. 1, 2018.

[16] A. O. Balogun, E. O. Uyanne, and L. O. Badamas, "Prevalence and causes of misconduct among university undergraduates in Ilorin, Nigeria," *Interdisciplinary Journal of Education*, vol. 4, no. 2, 2021.

[17] O. I. Stephen, A. I. Ayomide, and V. A. Esan, "Campus criminality and vulnerability of victimisation among undergraduate students in Ekiti State urban universities," *IJRISS*, vol. 8, no. 7, 2024.

[18] Project Clue, "An appraisal of increase in property crimes among students of village hostel, University of Jos."