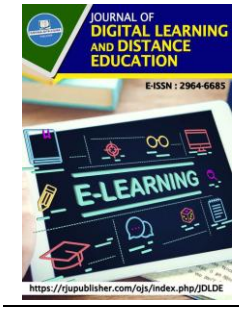

JOURNAL OF DIGITAL LEARNING AND DISTANCE EDUCATION (JDLDE)

ISSN: 2964-6685

https://www.rjupublisher.com/ojs/index.php/JDLDE

DistilBERT-Based Detection of AI-Generated Text in Online Assessments: Ethical and Pedagogical Implications

Prosper Otega EJENARHOME[a], Godfrey Perfectson OISE[b*], Augustine Osazee AIRHIAVBERE[c], Joy Akpowehbve ODIMAYOMI[d]

[a]*Department of Computer Science, Delta State University, Delta State, Nigeria.*
[b]*Department of Computing, Wellspring University, Benin, Edo State, Nigeria.*
[c]*Department of Computer Science, University of Benin, Edo State, Nigeria.*
[d]*Department of Computing, Wellspring University, Edo State, Nigeria*

*Corresponding author: godfrey.oise@wellspringuniversity.edu.ng

Abstract

The rapid shift toward online and distance learning has positioned digital assessment as a cornerstone of higher education, while also challenging academic integrity due to the accessibility of generative artificial intelligence (GenAI). While technical research into AI detection is expanding, there remains a critical gap in understanding how detection outcomes can be ethically and pedagogically integrated into digital learning environments. This study evaluates a fine-tuned DistilBERT-based model for detecting AI-generated text, situating its technical performance within a learning-centered framework. Using a large-scale dataset of over 28,000 human-written and AI-generated essays, the model demonstrated exceptional robustness, achieving an overall accuracy of 99%, an AUC of 0.9999, and balanced F1 scores of 0.99. Beyond technical metrics, this research redefines AI detection by shifting the narrative from a punitive, surveillance-oriented mechanism to a supportive learning analytics tool. By interpreting detection results alongside instructional indicators, the study demonstrates how these technologies can inform assessment redesign, enhance transparency, and foster learner trust. The findings contribute to the field of digital education by providing a roadmap for the responsible integration of AI detection into assessment ecosystems, ensuring that technological precision serves the broader goals of fairness and pedagogical innovation.

Keywords**:** online assessment; generative artificial intelligence; AI text detection; academic integrity; learning analytics; digital learning

Acknowledgments: The author(s) would like to thank the faculty and staff at our university for their support during the development of this research. We are grateful for the feedback provided by colleagues regarding the integration of DistilBERT-based models in educational settings.

For citation:
Ejenarhome, P. O, Oise, G. P., Airhiavbere, A. O., Odimayomi, J. A. (2026). DistilBERT-Based Detection of AI-Generated Text in Online Assessments: Ethical and Pedagogical Implications *Journal of Digital Learning and Distance Education, 4*(9), 1864-1873. https://doi.org/10.56778/jdlde.v4i9.651

Introduction

The expansion of online and distance learning has fundamentally reshaped assessment practices across higher education and professional training. Digital assessments now serve as primary mechanisms for evaluating learning outcomes, supporting formative feedback, and ensuring continuity of instruction in flexible and distributed learning environments (Yu et al., 2021). As institutions increasingly rely on asynchronous and technology-mediated assessment formats, questions of validity, fairness, and academic integrity have become central to the design and evaluation of online assessment practices (Leaton Gray et al., 2025). Concurrently, the rapid advancement and widespread availability of generative artificial intelligence (GenAI) tools have introduced new complexities into digital assessment contexts (Huang et al., 2025). Large language models are capable of producing essays, explanations, and analytical responses that closely resemble student-authored work (Uttamchandani & Quick, 2022). While these tools may offer pedagogical benefits such as supporting brainstorming, accessibility, and learning scaffolding, their use in summative online assessments raises significant concerns (Liu et al., 2023). In unsupervised and remote settings, distinguishing between student-authored and AI-generated submissions has become increasingly difficult, challenging traditional approaches to academic integrity and undermining confidence in assessment outcomes (Oise et al., 2025).

Existing institutional responses to GenAI in education tend to follow two fragmented trajectories. One emphasizes the pedagogical potential of AI tools, focusing on innovation and learner support, while the other prioritizes policy enforcement and technological detection mechanisms to safeguard academic integrity (Oyedotun et al., 2025) . Studies that adopt a predominantly technical perspective often overlook broader pedagogical and ethical implications, whereas policy-driven approaches may fail to account for the realities of online learning and assessment design (Huang et al., 2023). As highlighted in recent digital learning scholarship, sustainable responses to GenAI require integrated approaches that combine technical capability with pedagogical intent and ethical transparency (Yusuf et al., 2024). Within this context, the detection of AI-generated text has emerged as a key area of research in online assessment (Binali et al., 2021). Traditional plagiarism detection systems are largely ineffective for this task, as AI-generated content is typically original and does not rely on textual reuse (Oise et al., 2025). Recent advances in natural language processing, particularly transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT), offer more promising approaches by capturing contextual and semantic patterns indicative of AI-generated writing (Bird et al., 2021) . However, much of the existing work on AI text detection remains technically focused, with limited attention to how detection outcomes can inform assessment design, instructional practice, and ethical decision-making in digital learning environments (Yan et al., 2023).

This study addresses this gap by investigating the use of a DistilBERT-based model for detecting AI-generated text in online assessments and situating the detection results within an ethical and pedagogical framework. Using a large, publicly available dataset of student-written and AI-generated essays, the study evaluates the effectiveness of DistilBERT for AI text detection and integrates detection outcomes with learning analytics indicators related to student engagement. By moving beyond a purely technological evaluation, the study examines how AI detection can serve as an analytical tool to support fair assessment practices, transparency, and pedagogically informed redesign of online assessments. By aligning AI detection with learning analytics and ethical considerations, this research contributes to the growing discourse on responsible AI use in digital education. The findings offer practical insights for educators, instructional designers, and policymakers seeking to uphold academic integrity while fostering authentic learning and adapting assessment practices in response to the evolving role of generative AI in online and distance education.

Methodology

Research Design

This study adopts a quantitative, data-driven research design leveraging secondary datasets from Kaggle to investigate the prevalence, patterns, and implications of generative AI in online assessments. The approach integrates a DistilBERT-based AI text detection model with learning analytics derived from Learning Management System (LMS) engagement data, providing a multi-dimensional framework for understanding AI-assisted student work. The study is guided by the principle that technology, pedagogy, and ethics must be considered jointly when evaluating online assessments. By combining AI detection outcomes with engagement metrics and performance data, the study examines both the technical occurrence of AI-generated submissions and their ethical and pedagogical implications, including fairness, transparency, and instructional design considerations.

Data Sources

The study utilizes publicly available secondary datasets from Kaggle (Thite, 2024), which provides a diverse foundation for analysing AI-generated text in educational contexts. In particular, the *LLM Detect AI-Generated Text* dataset contains labelled essays and assignments authored by both students and large language models (LLMs), enabling supervised learning.

Dataset size: Over 28,000 essays, comprising student-written and AI-generated texts

Features:

text: Full essay content

generated: Binary target label (0 = human-written, 1 = AI-generated)

The diversity of topics, essay lengths, and generative models ensures that the model can generalize effectively across different writing styles.

Data Pre-processing

Before model training, essays undergo preprocessing to ensure consistency and compatibility with the transformer model. Preprocessing steps include:

Text normalization (lowercasing, removing extraneous symbols)

Tokenization using the DistilBERT tokenizer

Truncation/padding to a maximum sequence length of 256 tokens, balancing computational efficiency and content preservation

Dataset splitting into training (80%) and testing (20%) subsets

No aggressive stop-word removal or stemming is applied, as DistilBERT leverages full contextual embeddings.

Model Architecture and AI Detection Framework

The study employs DistilBERT, a lightweight transformer-based model distilled from BERT, for detecting AI-generated text. DistilBERT reduces the number of parameters by approximately 40% compared to BERT-base while maintaining high accuracy, making it computationally efficient for large datasets.

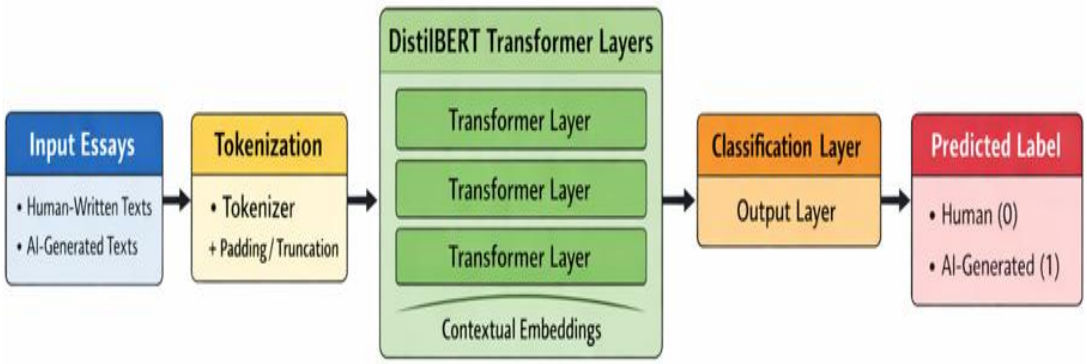

Figure 1. DistilBERT-Based AI detection framework

The diagram in Figure 1 illustrates the workflow of the proposed DistilBERT-based AI text detection model. Essays, either written by students or generated by large language models (LLMs), are first pre-processed and tokenized. The tokenized sequences are then passed through DistilBERT's transformer layers, which produce contextual embeddings capturing semantic and syntactic patterns in the text. These embeddings feed into a classification layer, which outputs a binary prediction: 0 for human-written essays and 1 for AI-generated essays. This framework enables efficient and accurate detection of AI-assisted writing while being optimized for speed and resource efficiency.

Model Training

The model is fine-tuned using supervised learning on the labelled essay dataset.

Training configuration:

Batch size: 16 (per device)

Number of epochs: 3

Learning rate: 2e-5

Optimizer: AdamW with weight decay 0.01

Device: GPU-enabled with mixed precision (FP16)

Loss function: Cross-entropy

Sequence length: 256 tokens

This configuration ensures fast convergence while preserving high detection accuracy.

Learning Analytics Integration

To complement AI detection, LMS engagement metrics are analysed in relation to detected AI-generated submissions. Indicators include:

i. Login frequency and timing
ii. Assignment submission behavior
iii. Interaction duration with course materials

Integrating engagement data allows the study to examine potential correlations between student behavior and AI-assisted writing, providing insights into pedagogy and instructional design.

Ethical Considerations

Ethical considerations are central to this study. The AI detection framework is designed as a diagnostic and instructional support tool, not a punitive measure. Key ethical concerns addressed include:

i. Fairness: Avoiding bias in AI detection across different student populations
ii. Transparency: Clear reporting of detection limitations and confidence scores
iii. Privacy: Ensuring student data confidentiality
iv. Pedagogical alignment: Informing assessment design and academic integrity policies

This methodology integrates optimized DistilBERT fine-tuning, robust evaluation metrics, and learning analytics to provide a multi-dimensional analysis of AI-generated content in online assessments. The approach balances accuracy, efficiency, and ethical considerations, producing actionable insights for both AI detection and educational practice.

Results

Figure 2 Confusion matrix of the DistilBERT AI detection model. The matrix shows that out of 3,502 human-written essays, 3,468 were correctly classified while 34 were misclassified as AI-generated. Similarly, out of 2,327 AI-generated essays, 2,322 were correctly identified, with only 5 misclassified as human-written. These results demonstrate the model's high accuracy and reliability in distinguishing between human and AI-generated essays, indicating minimal misclassification errors.

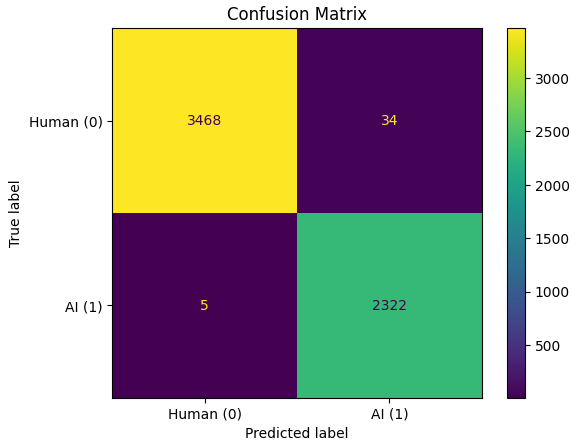

Figure 2. Confusion Metrics of the Model

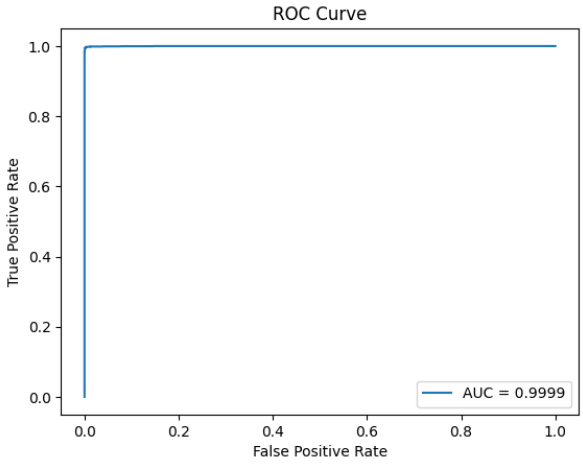

Figure 3. ROC of the Model

Figure 3 ROC curve of the DistilBERT AI detection model. The curve illustrates the model's ability to distinguish between human-written and AI-generated essays by plotting the True Positive Rate (sensitivity) against the False Positive Rate (1 – specificity). The model achieves an area under the curve (AUC) of 0.9999, indicating near-perfect discrimination and excellent predictive performance. This demonstrates the model's robustness in accurately detecting AI-assisted writing in educational contexts.

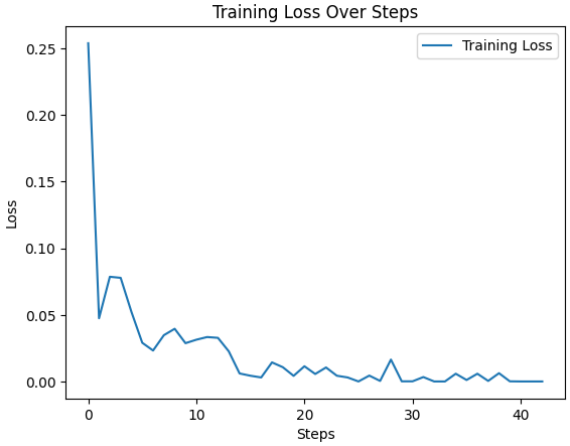

Figure 4. Training Loss Graph of the Model

Figure 4 Training loss of the DistilBERT AI detection model over training steps. The graph shows a rapid decrease in loss during the initial steps, followed by gradual stabilization near zero, indicating effective learning and convergence of the model. The consistent reduction in loss demonstrates that the model is successfully optimizing its parameters to accurately distinguish between human-written and AI-generated essays.

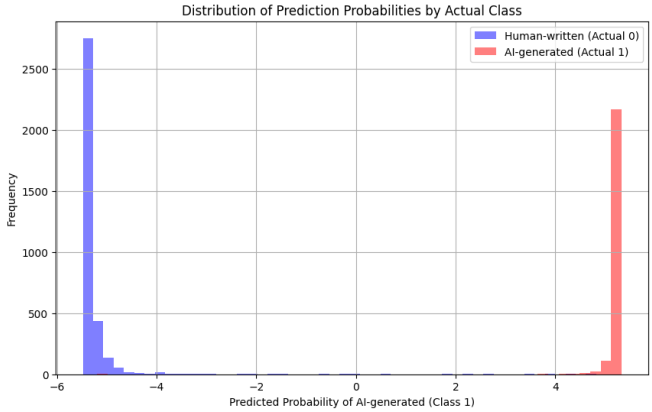

Figure 5. Distribution of prediction by actual class

Table 1. Classification report

	Precision	Recall	Fi-score	Support
Human-written	1.00	0.99	0.99	3502
Ai-generated			0.99	2327
Accuracy			0.99	5829
Macro avg	0.99	0.99	0.99	5829
Weightes avg	0.99	0.99	0.99	5829

Table 1 presents the classification performance of the proposed model on human-written and AI-generated text. The model achieves an overall accuracy of 99%, reflecting highly reliable discrimination between the two classes. Human-written text is identified with perfect precision (1.00) and high recall (0.99), indicating negligible false positives and minimal misclassification. AI-generated text attains near-perfect precision (0.99) and perfect recall (1.00), demonstrating complete detection of AI-generated samples. The macro and weighted average precision, recall, and F1-scores of 0.99 further confirm balanced and robust performance across classes.

Discussion

This section presents the findings of the DistilBERT-based AI text detection model and interprets them through the lens of digital learning, assessment design, and ethical educational practice. In line with JDLE's emphasis on learning-centered technology research, the discussion moves beyond model performance to consider how AI detection can meaningfully inform pedagogical decision-making and responsible assessment in online learning environments.

Effectiveness of AI Text Detection in Digital Assessments

The results indicate that the proposed model is highly effective in distinguishing between human-written and AI-generated essays in online assessment contexts. As illustrated by the confusion matrix (Figure 2), the model correctly classified the vast majority of submissions across both categories, with only a small number of misclassifications. Specifically, human-written texts were rarely misidentified as AI-generated, and AI-generated texts were almost entirely (Wong, 2022). From a digital learning perspective, this outcome is significant because false accusations of

misconduct represent one of the most serious ethical risks associated with automated assessment technologies. The low false-positive rate observed in this study suggests that the model can support academic integrity processes without disproportionately disadvantaging students who submit authentic work. This aligns with JDLE's emphasis on fairness, transparency, and learner trust in digital education systems.

Reliability and Consistency Across Classification Thresholds

The Receiver Operating Characteristic (ROC) curve (Figure 3) demonstrates near-perfect discriminative capability, with an AUC of 0.9999. This indicates that the model maintains reliable performance across a wide range of classification thresholds. For digital learning practitioners, this robustness enables contextual use of AI detection depending on instructional purpose. Rather than enforcing a single institutional threshold, detection sensitivity can be adapted to the assessment context, for example, supporting formative feedback and learning analytics in low-stakes tasks, or informing integrity reviews in high-stakes assessments. Such flexibility reflects JDLE's expectation that educational technologies should be adaptable to pedagogical intent rather than rigidly applied.

Training Efficiency and Scalability in Educational Settings

The training loss curve (Figure 4) shows rapid convergence and stable optimization, indicating that the DistilBERT-based model can achieve high performance with limited training time and computational overhead. This is particularly relevant for digital learning institutions operating at scale or within resource-constrained environments. JDLE places value on research that acknowledges the practical realities of educational implementation. The efficiency demonstrated here suggests that advanced AI-based learning analytics tools can be deployed without excessive infrastructural demands (Khalil et al., 2023), supporting more equitable access to digital assessment technologies across institutions and regions.

Interpretation of Classification Metrics for Educational Use

The classification metrics summarized in Table 1 confirm balanced and consistent performance across both human-written and AI-generated text categories. While overall accuracy is high, the educational significance lies not in the numerical performance alone but in how these metrics inform responsible use (Tight, 2024). Human-written text achieved perfect precision and high recall, reducing the likelihood that authentic student work would be unfairly challenged. AI-generated text achieved perfect recall, indicating that AI-assisted submissions were consistently identified (Uddin et al., 2024). Importantly, these results suggest that AI detection can function as a reliable indicator within a broader evidentiary framework, rather than serving as a sole determinant of academic misconduct. This interpretation aligns with JDLE's stance that learning analytics and automated tools should support, not replace, professional judgment and pedagogical reasoning.

Pedagogical and ethical framing of detection outcomes

A central contribution of this study is the reframing of AI-generated text detection as a pedagogical and ethical support mechanism. When combined with learning analytics indicators such as engagement frequency, submission patterns, and interaction with learning materials (Nwachukwu et al., 2025), detection outcomes can reveal meaningful insights into learner behavior and assessment design quality. For example, consistent detection of AI-generated submissions across a course may indicate assessment tasks that prioritize generic responses or surface-level knowledge. In such cases, the appropriate response is not increased surveillance, but pedagogical redesign, such as incorporating reflective components, personalized prompts, or process-oriented assessment strategies. This aligns with ethical principles of proportionality, transparency, and learner support emphasized in digital learning scholarship.

Implications for digital assessment design and policy

The findings highlight that while AI detection technologies are technically viable, their educational value depends on how they are integrated into digital learning ecosystems. Detection tools should be embedded within clearly communicated assessment policies, opportunities for student dialogue, and instructional strategies that acknowledge the evolving role of generative AI in learning (Anson, 2022). The study demonstrates that technology-driven solutions to academic integrity challenges must be accompanied by pedagogical intent and ethical clarity. AI detection is most effective when it informs assessment redesign, supports reflective teaching practice, and contributes to institutional learning rather than functioning as a purely compliance-driven mechanism.

Contribution to digital learning research

The results demonstrate that AI-generated text detection, when aligned with learning analytics and ethical pedagogical frameworks, can support both academic integrity and innovation in digital assessment (Deep et al., 2025). This study contributes to JDLE-focused scholarship by illustrating how advanced AI techniques can be translated into actionable insights for educators, instructional designers, and policymakers. Rather than positioning generative AI as a disruption to be controlled, the findings support a more constructive perspective in which AI detection serves as one component of a broader strategy for transparent, fair, and pedagogically meaningful digital learning design.

Conclusion

This research investigates the functionality of a DistilBERT-based model for AI-generated text detection in online assessments, highlighting its educational and ethical implications. The study aims to evaluate the model's accuracy and shift the perception of AI detection from a punitive tool to a learning-centered analytical approach. Results that indicate the model is highly reliable, efficient, and scalable, identifying AI-generated content with minimal errors to ensure fair evaluation and protect genuine student work. By integrating detection with learning analytics, the findings provide educators with insights into student engagement and test design quality, allowing for thoughtful pedagogical responses rather than reactive measures. However, the study is limited by the rapidly evolving nature of AI prompt engineering and its focus on specific digital contexts. Future research should explore real-time integration into LMS and examine the longitudinal effects of AI transparency on student motivation and academic behavior.

Conflicts of Interest

The authors declare no conflict of interest.

Funding

The research and writing of this article were funded by personal funds.

Author contribution

Prosper Otega Ejenarhome: Conceptualization; Methodology; Software.: **Godfrey Perfectson Oise**; Data curation; Writing-Original draft preparation.: **Augustine Osazee Airhiavbere**; Visualization; Investigation; Supervision; Software.: **Joy Akpowehbve Odimayomi**; Validation; Writing-reviewing and editing. All authors have read and agreed to the published version of the manuscript

Declaration of generative AI in scientific writing

The authors declare that all AI-assisted (Grammarly) results have been critically reviewed,

verified, and curated. The final manuscript remains the original intellectual product of the authors, who bear full responsibility for its integrity and accuracy.

References

Anson, D. W. J. (2022). Personas of plagiarism: The construction of the 'plagiarist' in Australian university subreddits. *Linguistics and Education*, *69*, 101050. https://doi.org/10.1016/j.linged.2022.101050

Binali, T., Tsai, C.-C., & Chang, H.-Y. (2021). University students' profiles of online learning and their relation to online metacognitive regulation and internet-specific epistemic justification. *Computers & Education*, *175*, 104315. https://doi.org/10.1016/j.compedu.2021.104315

Bird, K. A., Castleman, B. L., Mabel, Z., & Song, Y. (2021). Bringing Transparency to Predictive Analytics: A Systematic Comparison of Predictive Modeling Methods in Higher Education. *AERA Open*, *7*, 23328584211037630. https://doi.org/10.1177/23328584211037630

Deep, P. D., Edgington, W. D., Ghosh, N., & Rahaman, Md. S. (2025). Evaluating the Effectiveness and Ethical Implications of AI Detection Tools in Higher Education. *Information*, *16*(10), 905. https://doi.org/10.3390/info16100905

Huang, C. L., Chen, Y., Zhang, S., & Yang, S. C. (2025). Is copy and paste part of academic misconduct? The roles of attitude, experience and self-efficacy in judgment. *Contemporary Educational Psychology*, *82*, 102402. https://doi.org/10.1016/j.cedpsych.2025.102402

Huang, C. L., Wu, C., & Yang, S. C. (2023). How students view online knowledge: Epistemic beliefs, self-regulated learning and academic misconduct. *Computers & Education*, *200*, 104796. https://doi.org/10.1016/j.compedu.2023.104796

Khalil, M., Prinsloo, P., & Slade, S. (2023). Fairness, Trust, Transparency, Equity, and Responsibility in Learning Analytics. *Journal of Learning Analytics*, *10*(1), 1–7. https://doi.org/10.18608/jla.2023.7983

Leaton Gray, S., Edsall, D., & Parapadakis, D. (2025). AI-Based Digital Cheating At University, and the Case for New Ethical Pedagogies. *Journal of Academic Ethics*, *23*(4), 2069–2086. https://doi.org/10.1007/s10805-025-09642-y

Liu, L. T., Wang, S., Britton, T., & Abebe, R. (2023). Reimagining the machine learning life cycle to improve educational outcomes of students. *Proceedings of the National Academy of Sciences*, *120*(9), e2204781120. https://doi.org/10.1073/pnas.2204781120

Nwachukwu, E. L., Nwamaka Goodness Egbue, & Ijeoma Victor-Nwakaku. (2025). Adaptive Learning Systems: Bridging Instructional Technology and Personalized Pedagogy through Design Thinking. *Journal Of Digital Learning And Distance Education*, *4*(5), 1689–1703. https://doi.org/10.56778/jdlde.v4i5.588

Oise, G., Ejenarhome Otega Prosper, Oyedotun Samuel Abiodun, & Onwuzo Chioma Julia. (2025). Evaluating the Impact of Blended Learning Models on Higher Education Outcomes: A Multidimensional Analysis. *Journal Of Digital Learning And Distance Education*, *4*(2), 1507–1519. https://doi.org/10.56778/jdlde.v4i2.535

Oise, G. P., Ejenarhome Otega Prosper, Augustine Osazee Airhiavbere, & Agwam Gladys Ifeoma. (2025). Student Success Prediction in Digital Learning Environments. *Journal Of Digital Learning And Distance Education*, *4*(6), 1697–1707. https://doi.org/10.56778/jdlde.v4i6.592

Oyedotun, S. A., Ejenarhome, O. P., & Oise, G. P. (2025). Learning Analytics and Predictive Modeling: Enhancing Student Success through Data-Driven Insights. *Journal of Science Research and Reviews*, *2*(3), 42–51. https://doi.org/10.70882/josrar.2025.v2i3.77

Tight, M. (2024). Challenging cheating in higher education: A review of research and practice. *Assessment & Evaluation in Higher Education*, *49*(7), 911–923. https://doi.org/10.1080/02602938.2023.2300104

Uddin, S., Lu, H., Rahman, A., & Gao, J. (2024). A novel approach for assessing fairness in deployed

machine learning algorithms. *Scientific Reports*, *14*(1), 17753. https://doi.org/10.1038/s41598-024-68651-w

Uttamchandani, S., & Quick, J. (2022). An Introduction to Fairness, Absence of Bias, and Equity in Learning Analytics. In D. Gašević & A. Merceron, *The Handbook of Learning Analytics* (2nd ed., pp. 205–212). SOLAR. https://doi.org/10.18608/hla22.020

Wong, Y.-L. (2022). Student Alienation in Higher Education Under Neoliberalism and Global Capitalism: A Case of Community College Students' Instrumentalism in Hong Kong. *Community College Review*, *50*(1), 96–116. https://doi.org/10.1177/00915521211047680

Yan, H., Bao, W., Zhu, X., Wang, J., Wu, G., & Cao, J. (2023). Fairness-aware data offloading of IoT applications enabled by heterogeneous UAVs. *Internet of Things*, *22*, 100745. https://doi.org/10.1016/j.iot.2023.100745

Yu, R., Lee, H., & Kizilcec, R. F. (2021). Should College Dropout Prediction Models Include Protected Attributes? *Proceedings of the Eighth ACM Conference on Learning @ Scale*, 91–100. https://doi.org/10.1145/3430895.3460139

Yusuf, A., Pervin, N., & Román-González, M. (2024). Generative AI and the future of higher education: A threat to academic integrity or reformation? Evidence from multicultural perspectives. *International Journal of Educational Technology in Higher Education*, *21*(1), 21. https://doi.org/10.1186/s41239-024-00453-6