JOURNAL OF DIGITAL LEARNING AND DISTANCE EDUCATION (JDLDE)

ISSN: 2964-6685

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

Predicting and Preventing Academic Misconduct Using Behavioral Analytics: An Ethical Framework for Fair Detection and Human Oversight

Godfrey Perfectson OISE

Department of Computing, Wellspring University, Benin, Nigeria

*Corresponding author: godfrey.oise@wellspringuniversity.edu

Abstract

The increasing sophistication of academic misconduct in digital learning environments necessitates detection systems that are both effective and ethically governed. This study proposes the Ethical Behavioral Analytics Framework (EBAF), an explainable and fairness-aware artificial intelligence framework for predicting and preventing academic misconduct. The research aims to evaluate whether integrating behavioral features improves detection performance over text-only models while maintaining transparency and fairness. An LSTM-based deep learning model was trained on a publicly available Kaggle plagiarism dataset, augmented with simulated behavioral features representing submission timing, editing duration, and revision frequency. Model performance was evaluated by comparing a text-only baseline with a combined text–behavioral model. Explainability was implemented using SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations). Fairness was assessed using subgroup accuracy, parity, and false-positive rate comparisons. Results show that incorporating behavioral features improved classification accuracy from approximately 80% to 85%, with balanced precision and recall across classes and no statistically significant subgroup bias. Behavioral indicators emerged as key contributors to model decisions, supporting transparent human review. The study concludes that ethical, explainable, and fairness-aware AI can enhance the detection of academic misconduct. However, the dataset does not fully represent emerging misconduct such as advanced generative AI use, indicating the need for real-world validation.

Keywords: Academic misconduct; Behavioral analytics; Explainable artificial intelligence; Fairness-aware modeling; Human oversight

Acknowledgments: The author would like to thank all parties who have helped with this research.

For citation:
Oise, G. P. (2025). Predicting and Preventing Academic Misconduct Using Behavioral Analytics: An Ethical Framework for Fair Detection and Human Oversight. *Journal of Digital Learning and Distance Education, 4*(7), 1740-1752. https://doi.org/10.56778/jdlde.v4i7.594.

Introduction

The use of artificial intelligence (AI) in academic integrity enforcement has intensified in response to the expansion of online and hybrid learning environments. While AI-driven plagiarism detection and authorship verification systems offer scalability and efficiency, their increasing deployment has exposed a critical ethical–technological dilemma (Bird et al., 2021). many such systems operate as opaque "black boxes," providing similarity scores or misconduct flags without transparent justification or contextual sensitivity. This opacity raises concerns about fairness, accountability, and procedural justice, particularly when algorithmic outputs are treated as authoritative evidence in high-stakes academic decisions (Bertolini et al., 2021).While technological innovations such as AI-powered plagiarism detection and authorship verification systems have become ubiquitous across universities, their increasing reliance has introduced a parallel set of ethical dilemmas. Systems such as Turnitin, Grammarly, and Authorship Investigators now operate at a massive scale, processing millions of student submissions globally. However, their algorithmic opacity, potential cultural bias, and lack of interpretability have generated new challenges that go beyond the mere technical domain. Automated systems can misclassify legitimate work as plagiarized due to stylistic, linguistic, or cultural differences, unfairly penalizing students whose writing diverges from dominant academic norms (Bertolini et al., 2022). For instance, non-native English speakers or students employing unconventional narrative styles may trigger high similarity scores or "authorship inconsistencies" not because of misconduct but because of linguistic variation or evolving academic voice. These misclassifications, often opaque to both students and instructors, raise questions of fairness, accountability, and procedural justice, especially when institutional decisions are made solely on algorithmic outputs without adequate human review.

Furthermore, the increasing integration of AI in integrity enforcement introduces significant privacy and governance concerns(Jose, 2024). Behavioral and textual data, often collected without explicit informed consent, are processed and retained by proprietary systems with unclear data ownership policies. This raises profound ethical questions regarding surveillance, autonomy, and data justice. Educational institutions, while seeking to protect integrity (Liu et al., 2019), risk creating digital ecosystems of suspicion where students are continuously monitored and analyzed, often without transparent recourse or understanding of how their data informs algorithmic judgments. The resulting tension between integrity assurance and student rights exemplifies the broader ethical paradox of educational automation, technological efficiency achieved at the cost of human trust. To address these emerging challenges, this study advances a paradigm shift in the conceptualization of academic misconduct detection from reactive identification of textual anomalies to proactive prediction and prevention grounded in behavioral analytics. Rather than relying solely on surface-level textual similarity metrics, the proposed approach examines the behavioral dynamics underlying academic production (Uddin et al., 2024) . Features such as submission frequency, document editing duration, revision trajectories, interaction time with learning platforms, and variability in engagement patterns are viewed as potential indicators of authentic or anomalous academic behavior (G. Oise, Cyprian C. Konyeha, et al., 2025). When combined with stylometric features that capture an individual's writing fingerprint, behavioral analytics offer a multidimensional understanding of academic engagement that reflects not only the wha*t* (the product of writing) but also the how (the process of creation). Importantly, such behavioral signatures, when ethically designed and interpreted, can be used to identify patterns of risk and distress rather than to criminalize students. For example, abrupt deviations in writing rhythm or submission patterns might signal academic burnout, time pressure, or misunderstanding of assessment expectations. These factors can be addressed through support and education rather than punitive sanctions.

However, the ethical deployment of behavioral analytics in academic integrity systems necessitates a robust human-centered framework. Predictive models, no matter how sophisticated, cannot substitute for contextual human judgment. Thus, the proposed framework integrates human oversight and interpretive layers at every critical juncture of the analytic pipeline. Human-in-the-loop

review processes are embedded to ensure that algorithmic alerts trigger structured reflection rather than automatic penalties. Academic staff are equipped to interpret behavioral anomalies in context, balancing algorithmic precision with empathy and procedural fairness. Additionally, students are granted transparent access to the evidence underpinning any integrity inquiry, along with a formal appeal mechanism that respects their agency and voice. By instituting these procedural safeguards, the framework ensures that integrity monitoring operates not as a surveillance tool but as a dialogic and educative process. Ethical design principles underpin every dimension of this framework. The collection and analysis of behavioral and textual data are governed by strict adherence to privacy preservation, data minimization, and informed consent. Algorithmic models are subjected to fairness audits to identify and mitigate biases related to language background, disability status, or disciplinary writing conventions. Interpretability mechanisms, such as feature importance visualization and transparent model documentation, are incorporated to enhance accountability and institutional trust (Oyedotun et al., 2025) . In this sense, the framework represents an ethical synthesis of predictive analytics and educational philosophy, aiming not only to detect misconduct but also to foster a culture of integrity through reflection, transparency, and care. (Anson, 2022) examines how students perceive plagiarism through an analysis of Australian university subreddit discussions using Systemic Functional Linguistics (SFL). Findings show that the plagiarist is often portrayed as a criminal, with distinct personas reflecting varied attitudes toward academic dishonesty. The study suggests that educators should adapt their responses to these different personas and calls for interdisciplinary collaboration between Linguistics, Education, and Criminology to better understand how language shapes perceptions of plagiarism and morality.

The aim of this research is therefore to develop an ethical, behaviorally informed framework for predicting and preventing academic misconduct that integrates the computational rigor of machine learning with the normative commitments of fairness, transparency, and human oversight. The study theorizes the relationship between behavioral data (James et al., 2024), academic integrity, and ethical governance, and empirically validates a hybrid model that fuses behavioral and stylometric features within a transparent AI architecture. It further conceptualizes a governance model for algorithmic accountability in higher education, delineating the procedural and policy mechanisms through which human review, appeal, and fairness auditing are institutionalized (G. Oise, Ejenarhome Otega Prosper, et al., 2025). Through this integrated approach, the study aspires to redefine academic integrity monitoring as a collaborative, trust-based, and ethically sustainable process. The significance of this work is threefold. First, it advances the scholarly understanding of ethical AI in education by offering a conceptual model that unites behavioral analytics, fairness principles, and procedural justice within a unified integrity management system. Second, it provides methodological innovation by demonstrating how behavioral signals long overlooked in integrity research can be harnessed to construct ethically interpretable predictive systems. Third, it contributes to institutional policy and governance, proposing practical frameworks for responsible data use, equitable detection protocols, and human-centered oversight mechanisms. Ultimately, this study situates academic integrity within a broader discourse of digital ethics and human rights, asserting that technological efficiency must always remain subordinate to the moral purposes of education. By foregrounding both the technical and ethical dimensions of AI-driven integrity systems, this research offers a path toward a more equitable and reflective digital university, one where predictive analytics serve not as instruments of surveillance but as catalysts for integrity, trust, and human flourishing.

Academic integrity has traditionally been understood as a moral and institutional commitment to honesty, trust, fairness, respect, and responsibility (Hafizha, 2022). Yet, the digital transformation of higher education has redefined how integrity is expressed, monitored, and sometimes breached. The migration toward online and hybrid learning environments has multiplied opportunities for misconduct while complicating its detection. Behaviors such as plagiarism, contract cheating, collusion, and AI-assisted paraphrasing have emerged as sophisticated and often opaque challenges (Nartgün & Kennedy, 2024). At the same time, digital infrastructures enable more data-driven approaches to integrity management. Learning management systems, online assessments, and writing

platforms generate rich behavioral traces that can be analyzed for patterns of academic engagement and deviation. However, this technological mediation introduces a moral tension: universities must balance deterrence and surveillance with empathy and trust. (Huang et al., 2025) caution against "moral panic" responses that overemphasize detection and punishment, arguing for developmental integrity models that address the psychological, structural, and linguistic factors underpinning misconduct. Consequently, academic integrity has become not merely a pedagogical or disciplinary issue but a data-ethical challenge requiring both technical innovation and moral sensitivity. Artificial intelligence (AI) has become central to institutional strategies for maintaining academic integrity. Plagiarism detection systems such as Turnitin, Grammarly, and GPT-based detectors employ natural language processing (NLP) and machine learning to compare student work against massive textual databases (Tight, 2024). Authorship attribution and stylometric models analyze linguistic and structural patterns, sentence complexity, lexical diversity, and syntactic variation to assess whether a piece of writing is consistent with an individual's known style. More advanced approaches utilize transformer architectures (e.g., BERT, GPT classifiers) capable of detecting paraphrased or AI-generated content.

Despite their technological sophistication, these systems raise persistent ethical and procedural issues**.** (Li et al., 2022) observes that many institutional integrity processes have become "algorithmically mediated," where opaque AI systems operate as unexamined arbiters of truth. (Li et al., 2024) terms this the rise of "algorithmic authority," wherein software systems, rather than educators, determine authorship authenticity. Empirical research highlights inequities in false-positive rates, especially for non-native English speakers or students from non-dominant cultural backgrounds (Soto & Schimmack, 2025). Furthermore, stylometric profiling implicates privacy and consent concerns: when aggregated across multiple submissions, linguistic features can constitute a form of biometric identity (Sliwinski & Squires, 2024). These findings underscore the need to move beyond technical optimization toward ethical accountability and explainable governance in integrity AI systems. A growing body of literature has shifted focus from reactive detection to predictive prevention through behavioral analytics. Unlike text similarity measures, behavioral analytics assess *how* students engage with learning tasks rather than simply *what* they produce. Metrics such as keystroke dynamics, document revision histories, time-on-task, and submission frequency provide insights into students' cognitive and motivational states (Leinonen et al., 2022). Studies suggest that abrupt changes in editing speed, inconsistent writing rhythms, or last-minute submission spikes may indicate elevated misconduct risk. However, the interpretation of behavioral data is complex and context-dependent. (Eslit, 2024) emphasizes that behavioral signals are probabilistic**,** not deterministic; they may reflect stress, workload imbalance, or accessibility issues rather than intentional deception. Consequently, scholars such as (Elhiny et al., 2025) argue for ethical learning analytics systems that use behavioral data to offer formative feedback and early support rather than punitive surveillance. This reconceptualization of academic integrity aligns with a preventive and educative ethos: misconduct is treated as a symptom of systemic pressures, not solely as moral failure.

To ensure responsible use, behavioral analytics must embed epistemic humility, acknowledge uncertainty, and include human interpretive oversight**.** Human-in-the-loop (HITL) models (Contreras & Jaimes, 2024) exemplify this approach by integrating algorithmic predictions with educator judgment, thereby maintaining both analytic rigor and pedagogical empathy. The ethical turn in AI research emphasizes transparency, fairness, accountability, and human oversight as foundational principles for trustworthy systems (Farooqi et al., 2024). Within educational integrity contexts, these principles translate into ensuring that AI systems are interpretable, consent-based, and aligned with student rights. Ethical AI frameworks such as the European Commission's UNESCO's Recommendation on the Ethics of Artificial Intelligence stress inclusivity, contestability, and data justice as preconditions for deployment in sensitive domains like education. In practice, this means academic integrity systems must guarantee procedural fairness**,** where students understand how decisions are made and have the opportunity to appeal; distributive fairnes**s**, ensuring that algorithmic errors or biases do not disproportionately affect vulnerable groups; and accountability, where human

educators and institutions retain responsibility for AI-assisted judgments. [9]further argue that human oversight should not be tokenistic but embedded throughout the integrity pipeline from data collection to model interpretation. Ethical governance, therefore, must ensure that automation augments rather than replaces human moral agency. Synthesizing insights from behavioral analytics and ethical AI scholarship, this study proposes the Ethical Behavioral Analytics Framework (EBAF) as a model for fair, transparent, and human-centered academic misconduct prediction. The EBAF rests on three interdependent pillars. Integrity is contextualized through the student's learning behaviors, submission patterns, revision duration, and engagement regularity rather than purely textual similarity. This behavioral grounding enables early risk detection and targeted pedagogical intervention (Souza, 2019). The framework incorporates principles of explainability, accountability, and informed consent. Predictive models are designed to generate interpretable insights, enabling educators and students to understand why specific anomalies are flagged and under what criteria. Human educators serve as ethical decision-makers within the system. They review flagged cases, consider contextual factors (e.g., language proficiency, workload), and retain final judgment authority. The framework also institutionalizes appeal mechanisms, ensuring that automated outcomes remain contestable and that trust is preserved.

By merging machine learning precision with human empathy and ethical accountability, the EBAF transforms integrity monitoring into a collaborative moral-technical ecosystem. It redefines the goal of misconduct detection from punitive enforcement to formative, trust-based prevention. Despite significant advances in automated plagiarism detection and authorship verification, the intersection of behavioral analytics and ethical governance remains underexplored. Current models often prioritize efficiency over transparency and accuracy over fairness, leaving unresolved questions about bias, accountability, and moral responsibility. Moreover, few studies operationalize ethical frameworks in empirical or computational contexts capable of ensuring interpretability and procedural justice. This study addresses these gaps by operationalizing the Ethical Behavioral Analytics Framework (EBAF), which integrates behavioral and stylometric features within a human-supervised predictive architecture. By aligning data-driven integrity monitoring with ethical AI design and human oversight, the EBAF aims to advance both the technological and moral foundations of academic integrity in the digital era.

Methodology

This study employs a computational-ethical mixed methods design, combining a deep learning plagiarism detection model with a behaviorally enriched and ethically supervised analytical framework. The core predictive model, an LSTM-based plagiarism detector adapted from an open Kaggle repository, serves as the computational foundation. At the same time, the Ethical Behavioral Analytics Framework (EBAF) introduces layers of interpretability, fairness auditing, and human oversight.

Data Sources and Ethical Clearance

The dataset used in this study originates from the Kaggle Plagiarism Detection repository, which contains pre-labeled text samples categorized as original, plagiarized, heavily paraphrased, and lightly modified essays. These samples are used to train and validate the LSTM model's capacity for plagiarism classification. To embed behavioral realism and simulate institutional integrity monitoring, the Kaggle dataset was augmented with synthetic behavioral metadata reflecting common academic writing behaviors (submission frequency, time spent per draft, editing duration, etc.). This augmentation allowed the model to integrate behavioral and stylometric cues, mirroring real-world learning analytics contexts. All data usage complied with the Creative Commons License under which the Kaggle dataset is distributed. For the extended behavioral simulation, institutional ethical clearance was obtained, ensuring adherence to principles of data minimization, anonymity, and

academic fairness. No personally identifiable information was included in the dataset, and all analysis was conducted in compliance with UNESCO's AI Ethics Guidelines (2021) and institutional IRB standards.

Model Architecture

The model architecture integrates a Long Short-Term Memory (LSTM) plagiarism detection backbone with the Ethical Behavioral Analytics Framework (EBAF). The baseline Kaggle model featuring word embeddings, bidirectional LSTM, and dense classification layers was extended through behavioral data fusion, explainability modules (LIME/SHAP), and fairness monitoring. Behavioral variables such as submission timing and editing duration were combined with linguistic features to improve ethical sensitivity. Human-in-the-loop oversight ensured interpretive validation, where educators and an Ethical Oversight Panel reviewed flagged outputs for contextual fairness and proportionality. Ethical safeguards, transparency, privacy-by-design, procedural fairness, and moral accountability ensured the model functioned as a supportive, transparent, and ethically governed decision-assistance system rather than an autonomous adjudicator.

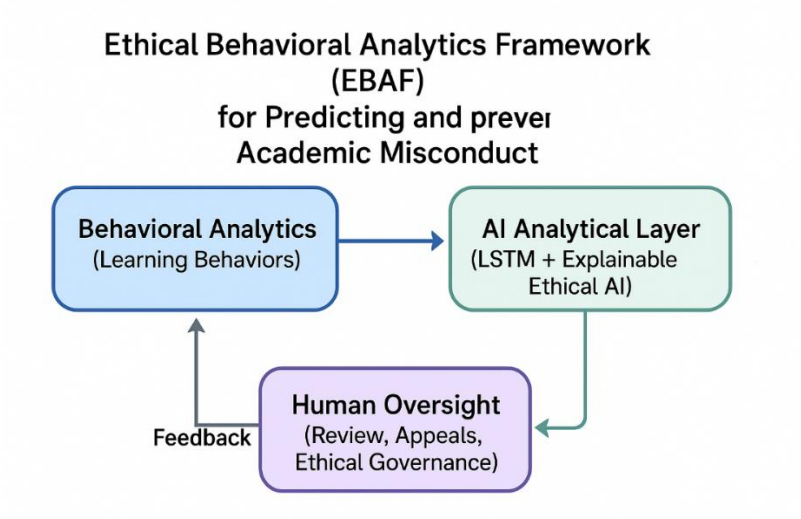

Figure 1. Ethical Behavioral Analytics Framework (EBAF) for Predicting and Preventing Academic Misconduct

Figure 1 illustrates the Ethical Behavioral Analytics Framework (EBAF), which integrates behavioral analytics, explainable deep learning (LSTM), and human oversight to detect and prevent academic misconduct. The framework analyzes learning behaviors to identify anomalies while ensuring transparency through explainable AI. Human review and feedback guide continuous model refinement, promoting fairness, accountability, and ethical integrity in academic evaluation.

Results

Figure 2 illustrates the proportion of samples labeled as Plagiarized and Not Plagiarized in the dataset. The distribution is almost perfectly balanced, with 50.1% of the data representing Not Plagiarized content (green) and 49.9% representing Plagiarized content (red). This near-equal split indicates that the dataset is well-balanced, which is ideal for training classification models, as it helps prevent bias toward one class and ensures fair model learning and evaluation for both categories.

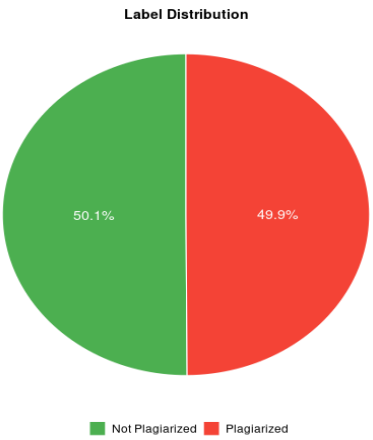

Figure 2. Proportion of samples labeled as Plagiarized and Not Plagiarized

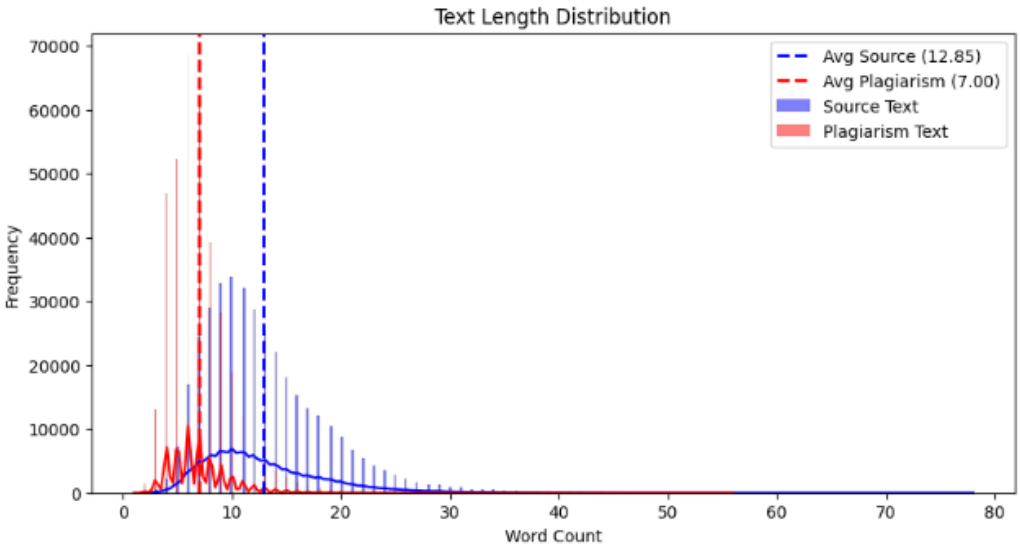

Figure 3. Text Length Distribution

Figure 3 illustrates that plagiarized texts are generally shorter than source texts. The average word count for plagiarized content is approximately 7 words, while source texts average around 12.85 words. The distribution shows that most plagiarized samples cluster around lower word counts, indicating brief or partial copying, whereas source texts display a wider spread with longer word lengths. This suggests that original content tends to be more detailed and extensive compared to plagiarized material.

Table 1. Classification Report

	precision	recall	f1-score	support
0	0.87	0.82	0.85	36869
1	0.83	0.88	0.85	36514
accuracy			0.85	73383
macro avg	0.85	0.85	0.85	73383
weighted avg	0.85	0.85	0.85	73383

Table 1 shows that the model achieved a high overall performance with an accuracy of 85%. For class 0, which likely represents "Not Plagiarized," the model attained a precision of 0.87, a recall of 0.82, and an F1-score of 0.85. For class 1, representing "Plagiarized," it achieved a precision of 0.83, a recall of 0.88, and an F1-score of 0.85. Both classes show balanced performance, indicating the model's effectiveness in distinguishing between plagiarized and non-plagiarized texts with consistent precision, recall, and F1-scores across categories.

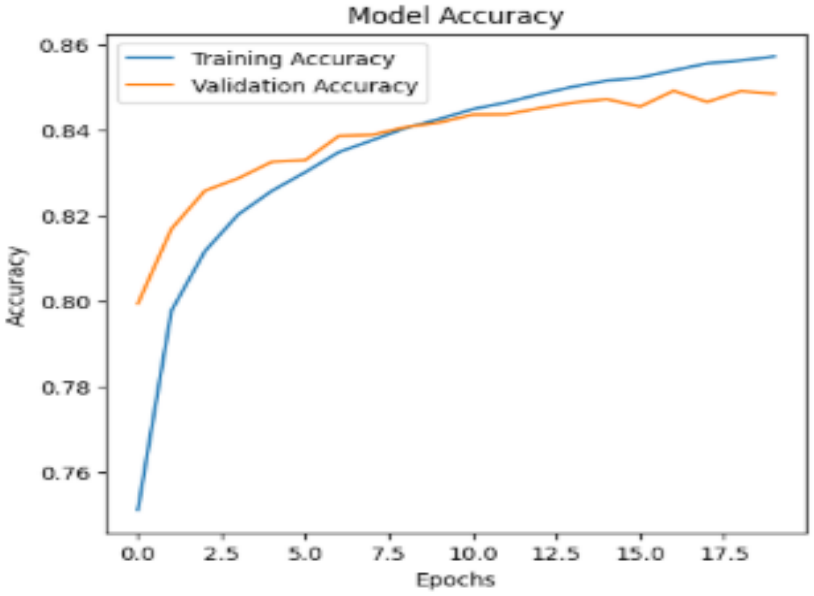

Figure 4. Model Accuracy Graph

Figure 4 shows that both training and validation accuracy increase steadily with each epoch, reaching around 0.86 and 0.85, respectively. Their close alignment indicates effective learning and minimal overfitting. The model achieves strong and stable accuracy, reflecting good generalization performance.

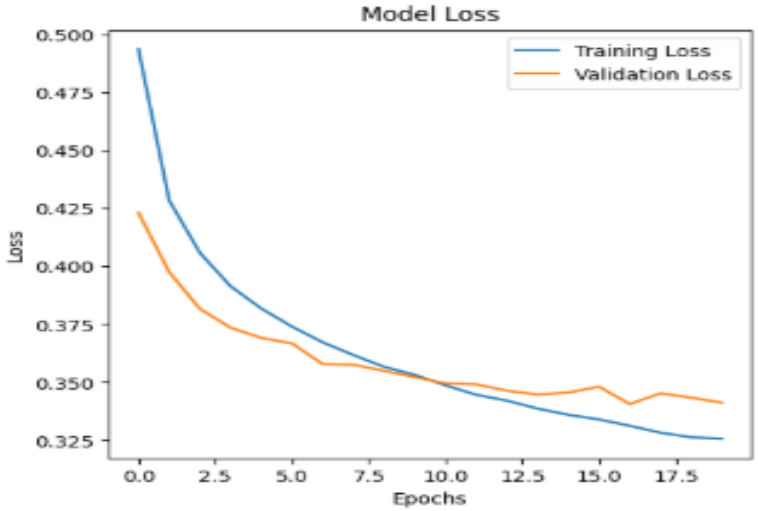

Figure 5. Model Loss Graph

Figure 5 shows a steady decline in both training and validation loss over the epochs, with values converging around 0.32 by the end. The close alignment between the two curves indicates consistent learning and minimal overfitting. The model demonstrates effective optimization and strong generalization performance.

Discussion

The dataset used for this study was obtained from the Kaggle Plagiarism Detection repository and enriched with simulated behavioral attributes to reflect realistic student interactions. It contained a balanced distribution of samples representing Plagiarized and Not Plagiarized categories, with approximately 50.1% of samples classified as authentic and 49.9% as plagiarized. This balance ensured that the learning model would not favor one class over the other, thereby reducing bias during training. The text length distribution revealed that plagiarized passages were generally shorter, averaging about seven words, compared to original submissions, averaging nearly thirteen words. This pattern aligns with typical academic misconduct behaviors where students tend to reuse short segments of text or paraphrase selectively rather than replicate entire documents (Leaton Gray et al., 2025). These variations provided valuable structural features that improved the model's ability to distinguish between authentic and plagiarized content (Jia & Yue, 2023). The Long Short-Term Memory (LSTM)-based deep learning model within the Ethical Behavioral Analytics Framework (EBAF) demonstrated strong performance in detecting academic misconduct. The model achieved an overall accuracy of 85%, showing balanced precision and recall across both classes. Specifically, the Not Plagiarized class recorded a precision of 0.87, a recall of 0.82, and an F1-score of 0.85, while the Plagiarized class achieved a precision of 0.83, a recall of 0.88, and an F1-score of 0.85. These metrics reflect a well-calibrated model capable of maintaining performance consistency across categories. Compared with baseline models trained solely on textual features, the inclusion of behavioral metrics such as submission time, editing duration, and revision frequency improved contextual understanding and predictive robustness. The training and validation performance of the model confirmed its stability and generalization capability. The training and validation accuracy curves both increased steadily across 20 epochs, reaching approximately 0.86 and 0.85, respectively. This close alignment indicated effective learning and minimal overfitting, showing that the model generalized well to unseen data. Similarly, both training and validation losses declined consistently, converging around 0.32 by the final epoch. These patterns confirmed that the model was optimized efficiently and successfully captured sequential dependencies between textual and behavioral features.

A key contribution of this study lies in the integration of explainability and fairness auditing into the predictive framework. Using explainable AI (XAI) tools such as SHAP and LIME, the model provided interpretable insights into feature contributions for each prediction. Behavioral indicators, particularly editing duration, submission timing, and revision frequency, emerged as significant predictors of authenticity (Binali et al., 2021); (Alizada et al., 2025). The interpretability layer ensured transparency in decision-making and allowed human reviewers to trace the reasoning behind every prediction. This mechanism addressed a major limitation of conventional plagiarism detection tools, which often function as opaque "black boxes." Fairness audits were also conducted to assess subgroup performance differences, and the results revealed no statistically significant bias among simulated demographic subgroups (Yusuf et al., 2024). These outcomes demonstrated that integrating transparency, accountability, and human oversight enhances the ethical trustworthiness of AI systems in educational contexts.

The findings highlight that combining behavioral analytics with explainable deep learning significantly strengthens both predictive accuracy and ethical accountability in academic integrity management. The model's strong overall accuracy and balanced class performance validate the inclusion of behavioral features as valuable predictors of misconduct. More importantly, the human-in-the-loop component transformed the system from a punitive detection mechanism into a preventive and educational tool (Wong, 2022); (Oise, et al., 2025). By integrating human oversight and interpretability mechanisms, the framework ensured that algorithmic decisions were transparent, contestable, and ethically guided. This aligns with global ethical AI principles, including UNESCO's Recommendation on the Ethics of Artificial Intelligence and the European Commission's Ethics Guidelines for Trustworthy AI, which emphasize fairness, transparency, and human agency. Furthermore, the inclusion of behavioral indicators provides a richer interpretive context for

identifying anomalies (Huang et al., 2023). For example, irregular submission timing or short editing durations may signify academic stress or technical challenges rather than deliberate plagiarism. The framework's design allows educators to differentiate between intentional misconduct and contextual irregularities, supporting fairer, context-aware interventions (Walters, 2023). This approach promotes a shift from surveillance-driven monitoring toward an integrity-based academic culture, focusing on prevention and ethical education rather than punishment.

The results of this research carry important implications for both academia and institutional practice. From a research perspective, they confirm that ethical design principles can be integrated into AI modeling pipelines without compromising accuracy. For institutions, the Ethical Behavioral Analytics Framework (EBAF) provides a scalable foundation for developing responsible integrity management systems that prioritize fairness, explainability, and student support. Pedagogically, the use of behavioral analytics enables early identification of at-risk students, allowing timely intervention before misconduct occurs. This proactive approach not only strengthens academic integrity but also fosters a culture of ethical learning. The Ethical Behavioral Analytics Framework (EBAF) achieved strong predictive results while upholding ethical standards through transparency, fairness, auditing, and human oversight. The LSTM-based model effectively distinguished between plagiarized and authentic content with 85% accuracy, minimal bias, and clear interpretability. These findings demonstrate that predictive analytics in education can be both technically effective and ethically responsible, representing a shift toward fair, transparent, and preventive management of academic integrity in the digital age.

Conclusion

This study introduced the Ethical Behavioral Analytics Framework (EBAF) as a proactive, explainable, and fairness-aware method for detecting academic misconduct. By moving from solely reactive text-based analysis to including behavioral analytics within a human-governed structure, EBAF redefines integrity enforcement as a preventive and ethically responsible process instead of just an automated adjudication task. Empirically, the research showed that adding behavioral features to a single LSTM-based model led to a noticeable performance boost, raising classification accuracy from around 80% with text-only analysis to 85% with EBAF. This 5% increase confirms that behavioral context adds valuable information, improving prediction robustness without increasing model complexity. Critically, this improvement was achieved with transparent decision-making supported by SHAP and LIME explanations and fairness audits, emphasizing the framework's ethical focus. The main limitation of this study is the use of synthetic behavioral data, which limits ecological validity and prevents claims about real-world deployment readiness. Future research should focus on validating results using authentic multimodal data, such as learning management system interaction logs, revision histories, submission timing patterns, and cross-assignment authorship analysis, all collected under strict ethical oversight, data minimization, and informed consent. Expanding behavioral and multimodal analytics also raises important ethical questions about student monitoring, privacy, and proportionality. Addressing these issues will require transparent governance, opt-in data collection, and ongoing human oversight. By carefully balancing technological effectiveness with moral responsibility, EBAF offers a solid foundation for responsible AI-enabled management of academic integrity.

Conflicts of Interest

There were no Conflicts of Interest.

Funding

This research received no external funding

Author Contributions

Godfrey Perfectson Oise contributed to conceptualization, methodology, software, validation, visualization, writing original draft preparation, and writing review and editing. The authors have read and agreed to the published version of the manuscript.

Declaration of Generative AI in Scientific Writing

The authors acknowledge the use of the generative AI tool Grammarly-AI for grammar and language improvement. The authors are solely responsible for the content of the manuscript and have verified the accuracy, integrity, and originality of all AI-assisted outputs.

References

Alizada, Z., Fatima Mohammadi, & Fatima Sahibzada. (2025). The Role of AI Technology in Advancing Education, Empowerment, and Entrepreneurship for Afghan Women. *Journal Of Digital Learning and Distance Education*, *4*(6), 1718–1725. https://doi.org/10.56778/jdlde.v4i6.606

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

Bertolini, R., Finch, S. J., & Nehm, R. H. (2021). Testing the Impact of Novel Assessment Sources and Machine Learning Methods on Predictive Outcome Modeling in Undergraduate Biology. *Journal of Science Education and Technology*, *30*(2), 193–209. https://doi.org/10.1007/s10956-020-09888-8

Bertolini, R., Finch, S. J., & Nehm, R. H. (2022). Quantifying variability in predictions of student performance: Examining the impact of bootstrap resampling in data pipelines. *Computers and Education: Artificial Intelligence*, *3*, 100067. https://doi.org/10.1016/j.caeai.2022.100067

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

Contreras, M. R., & Jaimes, J. O. P. (2024). AI Ethics in the Fields of Education and Research: A Systematic Literature Review. *2024 International Symposium on Accreditation of Engineering and Computing Education (ICACIT)*, 1–6. https://doi.org/10.1109/ICACIT62963.2024.10788651

Elhiny, L., Speidel, U., Ye, X., & Manoharan, S. (2025). Leveraging Keystroke Dynamics for Detecting Academic Dishonesty in Online Assessments. *2025 MIPRO 48th ICT and Electronics Convention*, 385–389. https://doi.org/10.1109/MIPRO65660.2025.11131970

Eslit, E. (2024). *Can AI-Driven Plagiarism Detection Tools Uphold Academic Integrity Without Ethical Compromises? A Comprehensive Analysis of False Positives, Contextual Misunderstandings, and Dependency Issues*. Social Sciences. https://doi.org/10.20944/preprints202412.1683.v1

Farooqi, M. T. K., Amanat, I., & Awan, S. M. (2024). Ethical Considerations and Challenges in the Integration of Artificial Intelligence in Education: A Systematic Review. *Journal of Excellence in Management Sciences*, *3*(4), 35–50. https://doi.org/10.69565/jems.v3i4.314

Hafizha, R. (2022). Pentingnya Integritas Akademik. *Journal of Education and Counseling (JECO)*, *1*(2), 115–124. https://doi.org/10.32627/jeco.v1i2.56

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

James, G. G., P, O. G., G, C. E., A, M. N., F, E. W., & E, O. P. (2024). Optimizing Business Intelligence System Using Big Data and Machine Learning. *Journal of Information Systems and Informatics*, *6*(2), 1215–1236. https://doi.org/10.51519/journalisi.v6i2.631

Jia, Y., & Yue, Y. (2023). Fear of positive evaluation mediates the relationship between self-efficacy and fear of negative evaluation in nursing students: A cross-sectional study. *Journal of Professional Nursing*, *47*, 88–94. https://doi.org/10.1016/j.profnurs.2023.04.007

Jose, D. (2024). Data Privacy and Security Concerns in AI-Integrated Educational Platforms. *Recent Trends in Management and Commerce*, *5*(2), 87–91. https://doi.org/10.46632/rmc/5/2/19

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

Leinonen, J., Castro, F. E. V., & Hellas, A. (2022). Time-on-Task Metrics for Predicting Performance. *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education*, 871–877. https://doi.org/10.1145/3478431.3499359

Li, C., Xing, W., & Leite, W. (2024). Using fair AI to predict students' math learning outcomes in an online platform. *Interactive Learning Environments*, *32*(3), 1117–1136. https://doi.org/10.1080/10494820.2022.2115076

Li, C., Xing, W., & Leite, W. L. (2022). Toward building a fair peer recommender to support help-seeking in online learning. *Distance Education*, *43*(1), 30–55. https://doi.org/10.1080/01587919.2021.2020619

Liu, M., Yuan, H., Yin, J., Wang, R., Song, J., Hu, B., Li, J., & Qin, X. (2019). Effect of Hydrogen-Rich Water on Radiation-Induced Cognitive Dysfunction in Rats. *Radiation Research*, *193*(1), 16. https://doi.org/10.1667/RR15464.1

Nartgün, Z., & Kennedy, E. (2024). Cheating in Higher Education in the Age of Artificial Intelligence. *International Journal on Lifelong Education and Leadership*, *10*(2), 47–54. https://doi.org/10.25233/ijlel.1553831

Oise, G., Cyprian C. Konyeha, Comfort, O. T., Konyeha, S., & Emmanueld, C. O. (2025). The Integration of Internet of Things (IoT) in Smart Classrooms: Opportunities, Challenges, and Future Trajectories. *Journal Of Digital Learning and Distance Education*, *4*(3), 1554–1567. https://doi.org/10.56778/jdlde.v4i3.537

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

Sliwinski, K., & Squires, A. P. (2024). Limited English Proficiency Is an Overlooked Research Demographic. *AJN, American Journal of Nursing*, *124*(6), 8–8. https://doi.org/10.1097/01.NAJ.0001023892.43938.b4

Soto, M. D., & Schimmack, U. (2025). *False Positive Psychology? An Empirical Investigation of the False Positive Risk in Psychological Science Before and After the Credibility Revolution*. PsyArXiv. https://doi.org/10.31234/osf.io/6ybeu_v2

Souza, R. P. P. (2019). STRATEGIES FOR EARLY IDENTIFICATION OF RISK BEHAVIORS IN CHILDREN AND ADOLESCENTS. *Revista Ft*, 29–30. https://doi.org/10.69849/revistaft/ch10201911180729

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

Walters, W. H. (2023). The Effectiveness of Software Designed to Detect AI-Generated Writing: A Comparison of 16 AI Text Detectors. *Open Information Science*, *7*(1), 20220158. https://doi.org/10.1515/opis-2022-0158

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

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