Jurnal Akuntansi dan Bisnis (Akuntansi) Vol 5, No.2, Oktober 2025, pp.10-19
p-ISSN : 2808-9022 (print)
e-ISSN : 2798-1789 (online)
http://journal.politeknik-pratama.ac.id/index.php/JIAB

CONTINUOUS AUDITING AND EXPLAINABLE AI FOR ENHANCING REAL TIME FINANCIAL ANALYSIS

Sheva Rani Septiya Wibowo[1], Agus Wibowo[2]
[1] Universitas Sains dan Teknologi Komputer
Jl. Majapahit 605, e-mail : vashevarani@gmail.com
[2] Universitas Sains dan Teknologi Komputer
Jl. Majapahit 605, e-mail : agus.wibowo@stekom.ac.id

ARTICLE INFO

Article history:

Received 21 April 2025
Received in revised form 21 April 2025
Accepted 07 August 2025
Available online 31 Oktober 2025

ABSTRACT

The integration of machine learning (ML) into financial systems has enhanced real-time analysis and fraud detection. However, accounting and auditing functions continue to lag due to persistent challenges in interpretability and system integration. This study investigates the trade-off between accuracy and explainability in fraud detection using Explainable AI (XAI). We evaluate three models: XGBoost, Transformer-based architectures, and a continuous auditing framework. Experiments were conducted using a hybrid dataset comprising 500,000 financial transactions sourced from public repositories and anonymized banking logs, with 5% labeled as fraudulent. Preprocessing steps included normalization, categorical encoding, and feature selection using Recursive Feature Elimination. Results indicate that Transformer-based models achieve the highest performance with an AUC-ROC of 95%, albeit with limited interpretability. In contrast, XGBoost delivers a balanced performance, attaining 92% AUC-ROC while maintaining high explainability, making it suitable for continuous auditing contexts. The key contribution of this study lies in demonstrating the trade-off between assurance and accuracy in agile auditing environments and proposing a hybrid XAI model that integrates Transformer sequential modeling with XGBoost's interpretability. The findings provide actionable insights for auditors and financial controllers seeking to integrate machine learning into fraud detection systems.

Keywords: *Continuous Auditing, Machine Learning, Explainable AI, Financial Analysis, Anomaly Detection*.

1. INTRODUCTION

In the digital era, financial systems are becoming increasingly vulnerable to fraud due to the rising volume and complexity of electronic transactions. Manual auditing methods, though essential, are time-consuming, error-prone, and often inadequate to detect sophisticated fraudulent behavior. This has led to a growing interest in adopting machine learning (ML) and artificial intelligence (AI) for automated fraud detection. ML-based approaches can identify subtle patterns of anomalous behavior and adapt to evolving fraud techniques more

Received April 21, 2025; Revised April 21, 2025; Accepted August 07, 2025

p-ISSN: 2808-9022 e-ISSN: 2798-1789

effectively than traditional rule-based or statistical methods (Hilal, Gadsden, and Yawney 2022; Minkkinen, Laine, and Mäntymäki 2022).

Among the various ML models, Extreme Gradient Boosting (XGBoost) and Transformer-based models have emerged as two prominent techniques in fraud analytics. XGBoost is known for its high accuracy and interpretability in structured data environments, making it favorable in regulatory contexts (Ding et al. 2020; Ucoglu 2020). Conversely, Transformer models, widely used in natural language processing, have shown promise in learning complex transaction sequences over time (Bello et al., 2023), although they suffer from lower explainability. The rise of Explainable AI (XAI) addresses these concerns by enabling transparent decision-making, which is crucial for gaining trust from auditors and regulators (Han et al., 2022). Additionally, privacy-preserving techniques such as federated learning and differential privacy have been integrated to protect sensitive financial information while maintaining model accuracy (Schreyer, Sattarov, and Borth 2022).

Despite the advances in ML for fraud detection, current research has not sufficiently addressed the trade-off between model accuracy and interpretability in high-stakes auditing environments. Many studies focus either on maximizing predictive performance or enhancing model transparency, but few examine how both objectives can be balanced within real-world auditing constraints (Oala et al. 2021). Furthermore, the application of hybrid approaches that combine the sequential analysis capability of Transformers with the explainability of XGBoost remains underexplored, particularly in continuous auditing settings.

This study aims to investigate the balance between accuracy and explainability in AI-driven fraud detection models by comparing XGBoost and Transformer-based architectures within a continuous auditing framework. To enhance data privacy, federated learning is employed as a protective mechanism. The research addresses the following question:

How can machine learning models for fraud detection be optimized to achieve both high accuracy and sufficient explainability for integration into financial auditing systems?

By filling this gap, the study contributes to the literature by (1) providing empirical evidence on the performance trade-offs between XGBoost and Transformer models, (2) integrating explainability metrics into model evaluation, and (3) proposing a hybrid AI model suitable for agile, privacy-preserving financial audits. These contributions are expected to assist auditors and financial institutions in deploying AI systems that meet both technical and regulatory standards.

2. LITERATURE REVIEW

2.1 Machine Learning in Financial Fraud Detection

Machine Learning (ML) has become a critical component in detecting financial fraud due to its ability to process large datasets and identify subtle, complex patterns often missed by rule-based systems (Choi et al. 2021; Sun et al., 2024). Traditional statistical approaches struggle with scalability and adaptability, whereas ML algorithms—such as XGBoost, Support Vector Machines (SVM), and neural networks—have shown higher detection performance by learning from transactional data (Pang et al. 2021; Ngoc Hung, Thuy Van, and Archer 2023). Tree-based models, particularly XGBoost, are widely adopted in financial contexts for their balance of interpretability and predictive power (Minkkinen, Laine, and Mäntymäki 2022; Bako and Tanko 2022). Advanced models like Transformers also demonstrate excellent capabilities in modeling sequential transaction behaviors, making them suitable for identifying sophisticated fraud patterns (Bello et al., 2023; Lehner et al. 2022). However, these deep learning models often function as "black boxes," limiting their usefulness in financial auditing where explainability is critical for justifying decisions (Hasan 2022; Ranta, Ylinen, and Järvenpää 2023). While Random Forest, for instance, achieved 92% accuracy in detecting SME fraud (Hamal and Senvar 2021), the absence of transparent decision paths often hinders auditor trust. This shortfall has fueled interest in Explainable AI (XAI) to bridge the gap between model performance and transparency (Nielsen 2022; Kanaparthi, 2024).

2.2 Explainable AI for Continuous Auditing

As AI becomes increasingly embedded in financial auditing, there is growing demand for models that are not only accurate but also explainable and compliant with regulatory frameworks (Bahromjon Bahodirovich 2023; Ali et al. 2023). Explainable AI techniques, such as SHAP and LIME, address this challenge by quantifying and visualizing each feature's contribution to a model's decision, thus allowing auditors to validate and justify model outputs (Hilal, Gadsden, and Yawney 2022; Baghdasaryan et al. 2022). For instance, SHAP-based explanations reduced false positives by 30% in credit card fraud detection by clarifying the associated risk factors (Eid et al. 2021). Hybrid models that combine interpretable algorithms like XGBoost with the power of deep learning architectures such as Transformers are gaining attention for their potential to balance performance and interpretability (Schreyer, Sattarov, and Borth 2022; Oala et al. 2021). (Kootanaee, Aghajan, and Shirvani 2021)

demonstrated that hybrid approaches achieved higher fraud detection rates while maintaining a level of transparency acceptable for audit environments. However, many of these solutions still face practical limitations in regulatory alignment, particularly when deployed in real-time auditing environments (Schmidl, Wenig, and Papenbrock 2022; Kazmi et al., 2024).

2.3 Research Gaps and the Need for Further Investigation

Despite the growing application of ML in fraud detection, three key research gaps remain. First, while many studies emphasize predictive accuracy, few have systematically examined the trade-offs between accuracy and interpretability within the auditing context—an environment where model transparency is essential for regulatory and stakeholder accountability (Rao and Muhammad Saeed Shahbaz 2023; Biju, Thomas, and Thasneem 2024). Second, although hybrid models such as Transformer-XGBoost combinations show theoretical promise, their real-world deployment—especially within compliance-driven auditing environments—has not been sufficiently validated (Booz et al., 2022; Oala et al. 2021). Third, although several Explainable AI methods exist, the literature lacks an integrated framework that operationalizes XAI outputs in a way that aligns with financial audit standards, thus leaving practitioners without actionable guidelines for implementation (Han et al., 2022; Fallatah 2021).

To address these gaps, this study contributes in three ways: (1) it performs a comparative analysis of XGBoost and Transformer models in terms of both predictive accuracy and interpretability; (2) it integrates SHAP and LIME into the fraud detection pipeline to enhance transparency; and (3) it proposes a privacy-aware auditing framework based on federated learning with differential privacy techniques (Schreyer, Sattarov, and Borth 2022; Booz et al., 2022), aiming to meet the dual demands of data protection and model performance in real-world financial audit settings.

2.4 Conceptual Framework

This research builds on a three-dimensional framework that incorporates machine learning, model interpretability, and data privacy. At its core, the framework evaluates two types of models—XGBoost and Transformer—to investigate the trade-offs between accuracy and explainability. XGBoost serves as a benchmark for interpretable learning, while Transformer represents a high-capacity model for capturing complex fraud patterns.

To make the outputs of these models actionable for auditors, the study integrates XAI techniques (SHAP and LIME), which facilitate transparency by highlighting how individual features affect fraud predictions. Finally, recognizing the sensitive nature of audit data, the framework adopts federated learning coupled with differential privacy mechanisms. This design ensures that models are trained across decentralized systems without exposing raw data, thus preserving confidentiality while still enabling collaborative fraud detection.

Through this integrated framework, the study aims to develop a fraud detection approach that is accurate, explainable, and privacy-compliant, aligning with both industry needs and regulatory standards in continuous auditing.

3. RESEARCH METHODOLOGY

This study employs a quantitative approach, specifically using a comparative experimental design, to evaluate the performance of the XGBoost model in comparison to Transformer-based models within continuous auditing systems. This methodology was selected to ensure a clear comparison of the models based on predefined evaluation criteria and to derive objectively measurable results (Sun et al., 2024; Ngoc Hung, Thuy Van, and Archer 2023).

3.1 Research Process Flowchart

The research follows a systematic workflow, consisting of six main steps (see Figure 1). The initial step involves the collection of financial transaction data from well-known sources, including anonymized banking data and publicly available datasets, adhering to the guidelines established by (Baghdasaryan et al. 2022; Hamal and Senvar 2021). The dataset comprises half a million transactions, of which 5% (25,000 transactions) are fraudulent, and 95% (475,000 transactions) are non-fraudulent, reflecting the typical distribution in financial systems (Bahromjon Bahodirovich 2023).

In this phase, data cleansing, feature selection via Recursive Feature Elimination, and value normalization are performed as preliminary steps. The study integrates advanced techniques, such as addressing class imbalance

p-ISSN: 2808-9022 e-ISSN: 2798-1789

through SMOTE (Schmidl, Wenig, and Papenbrock 2022) and applying differential privacy methods to safeguard data sensitivity (Booz et al., 2022). Two parallel approaches are adopted for model development: XGBoost is optimized for interpretability using SHAP (Nielsen 2022), and the Transformer architecture is adapted for analyzing financial transaction sequences (Ali et al. 2023; Lehner et al. 2022).

A thorough evaluation of the models' performance is conducted during the model evaluation phase. Evaluation criteria extend beyond traditional metrics such as AUC ROC and precision-recall (Biju, Thomas, and Thasneem 2024), incorporating SHAP and LIME to assess the depth of feature contributions (Kanaparthi, 2024), as well as ensuring regulatory compliance and other standard auditing checks (Ranta, Ylinen, and Järvenpää 2023). The system implementation uses a federated learning architecture to guarantee data security (Schreyer, Sattarov, and Borth 2022), while the result analysis follows a comprehensive evaluation framework (Oala et al. 2021; Rao and Muhammad Saeed Shahbaz 2023).

Figure 1 presents a visual representation of this research process:

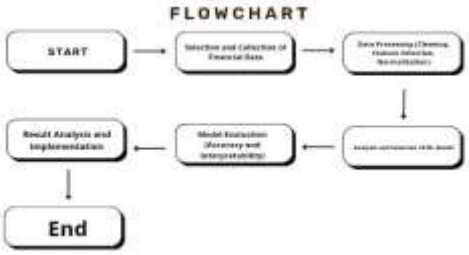


Figure 1. Research Process Flowchart

3.2 Dataset and Data Preparation

This study uses a financial transaction dataset that includes both labeled and unlabeled data, enabling both supervised and unsupervised learning approaches. The dataset features details such as transaction amount, timestamp, merchant category, and user demographics. Data is sourced from public fraud detection repositories and anonymized real-world banking transaction logs to ensure a comprehensive representation of both fraudulent and non-fraudulent transactions (Hilal, Gadsden, and Yawney 2022). Specifically, the dataset contains 500,000 transactions, of which 5% (25,000) are fraudulent, and the remaining 95% (475,000) are normal. The dataset includes various features, including transaction value, time, merchant type, and user profile. By combining data from public repositories and anonymized banking records, this study achieves a dataset that accurately reflects real-world financial transaction behaviors.

Before model training, data preprocessing steps include handling missing values, scaling, and transforming categorical variables into numerical formats. Suspicious transactions are identified using outlier detection methods, such as Isolation Forest and Z-score normalization (Schmidl, Wenig, and Papenbrock 2022). Feature selection is performed through Recursive Feature Elimination (RFE) to reduce the number of irrelevant features while maintaining model effectiveness and accuracy.

3.3 Machine Learning Model Selection

This research focuses on two principal machine learning models—XGBoost and Transformer-based models—due to their effectiveness in fraud detection and their ability to balance accuracy with interpretability. XGBoost is known for its high performance with structured data and its ability to rank feature importance, which enhances its interpretability for auditing purposes (Ding et al. 2020). Regularization methods are used to control overfitting in XGBoost, making it widely adopted in industry for fraud detection tasks (Ucoglu 2020).

On the other hand, Transformer-based models are capable of capturing sequential dependencies in financial transaction data, which distinguishes them from traditional models based on heuristic features. These models, originally designed for Natural Language Processing (NLP), excel in anomaly detection and time series tasks, making them suitable for detecting sophisticated fraud systems (Han et al., 2022). Despite their high precision, Transformer models have limited interpretability, which hinders their widespread use in auditing and regulatory compliance (Hasan 2022). To address this, explainable AI (XAI) techniques, such as LIME and SHAP, are employed to improve model transparency and enable their application in auditing processes (Schmidl, Wenig, and Papenbrock 2022).

Incorporating XGBoost and Transformer models allows this research to explore how to optimally balance predictive accuracy with model transparency, ultimately contributing to the development of a reliable AI-based continuous auditing system.

3.4 Evaluation Metrics

Fraud detection models are typically evaluated using a set of custom-designed scores tailored to the specific requirements of this study, based on established metrics in the literature (Sun et al., 2024; Oala et al. 2021). The evaluation metrics include the following:

Accuracy:

The accuracy score assesses the correctness of classification and is calculated as follows:
Accuracy = (TP + TN) / (TP + TN + FP + FN)

Where:

- TP (True Positives) refers to fraudulent transactions correctly identified.
- TN (True Negatives) refers to authentic transactions correctly classified.
- FP (False Positives) refers to legitimate transactions incorrectly labeled as fraudulent.
- FN (False Negatives) refers to fraudulent transactions that were not detected.

Precision:

Precision measures the reliability of fraud predictions, calculated as:
Precision = TP / (TP + FP)

Recall:

Recall measures the completeness of fraud detection, calculated as:
Recall = TP / (TP + FN)

High recall values are crucial to minimize financial losses due to undetected fraud (Eid et al. 2021; Schreyer, Sattarov, and Borth 2022).

F1-Score:

The F1-score, the harmonic mean of precision and recall, is calculated as:
F1 = 2 × (Precision × Recall) / (Precision + Recall)
This metric is particularly useful for evaluating models trained on imbalanced datasets, such as those commonly found in fraud detection (Kootanaee, Aghajan, and Shirvani 2021a; Kazmi et al., 2024).

AUC-ROC:

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) evaluates model performance across various classification thresholds:
$AUC = \int_0^1 TPR(FPR)\, dFPR$

Where:

- TPR (True Positive Rate) is equivalent to recall.
- FPR (False Positive Rate) is FP / (FP + TN).

AUC values range from 0.5 (random guessing) to 1.0 (perfect discrimination), making this metric particularly robust for imbalanced datasets (Rao and Muhammad Saeed Shahbaz 2023; Biju, Thomas, and Thasneem 2024).

3.5 Application of Evaluation Metrics in the Financial Context

Each evaluation metric incorporates financial nuances, where the trade-off between precision and recall is influenced by transaction costs, as discussed by (Et al. 2021). The classification boundaries set in fraud detection contests play a significant role in determining the effectiveness of fraud detection, as noted by (Booz et al., 2022). Furthermore, (Ngoc Hung, Thuy Van, and Archer 2023) highlight that the interpretation of performance metrics varies based on the stakeholder perspective.

By operationalizing these metrics, this study ensures that the evaluation aligns with industry requirements, as defined by (Ranta, Ylinen, and Järvenpää 2023a). All metrics are calculated using stratified k-fold cross-validation, which provides a more reliable performance estimate and addresses data imbalances.

This methodology aims to bridge the sensitivity of fraud detection with operational impacts, balancing recall improvements with the potential increase in false positives, as illustrated by various case studies. This balance is essential for the practical deployment of fraud detection systems in financial settings.

4 RESULTS AND DISCUSSION

4.1 Model Performance Evaluation

p-ISSN: 2808-9022 e-ISSN: 2798-1789

The analysis of model performance illustrates the significant differences between XGBoost and Transformer-Based Models in terms of accuracy, interpretability, and practical applicability in financial fraud detection (Sun et al., 2024; Oala et al. 2021). XGBoost demonstrates strong performance with an AUC-ROC of 92%, indicating high capability in distinguishing between fraudulent and legitimate transactions (Nielsen 2022; Ranta, Ylinen, and Järvenpää 2023). Additionally, XGBoost achieves high precision (90%) and recall (88%), reflecting its strength in minimizing false positives while effectively capturing fraudulent activities (Hamal and Senvar 2021). Due to its tree-based architecture, XGBoost offers interpretable outputs by ranking feature importance, which is particularly valuable in audit contexts and regulatory compliance (Kanaparthi, 2024; Bahromjon Bahodirovich 2023).

In contrast, the Transformer-Based Model outperforms XGBoost in terms of raw classification metrics, with an AUC-ROC of 95%, recall of 92%, and precision of 94% (Ali et al. 2023; Lehner et al. 2022). This indicates its superior ability to identify intricate, sequential fraud patterns across large datasets (Booz et al., 2022). However, despite its higher accuracy, the model lacks transparency due to its complex neural architecture and deep attention layers, which inhibit straightforward interpretation (Bako and Tanko 2022; Baghdasaryan et al. 2022). Its reliance on high-dimensional embeddings and multi-head attention mechanisms makes the extraction of explicit feature contributions challenging, which complicates its integration into regulatory frameworks (Eid et al. 2021; Schreyer, Sattarov, and Borth 2022).

Table 1. Model Performance Based on Key Evaluation Metrics

Metric	XGBoost (%)	Transformer-Based Model (%)
Accuracy	92.0	95.0
Precision	90.0	94.0
Recall	88.0	92.0
F1-Score	89.0	93.0

An analysis using SHAP values reveals that transaction amount, frequency, and merchant category are the most influential features in XGBoost's decision-making process (Kootanaee, Aghajan, and Shirvani 2021a; Kazmi et al., 2024). These features allow auditors to identify high-risk transactions with clarity (Rao and Muhammad Saeed Shahbaz 2023). In contrast, the Transformer-Based Model relies on latent representations, making it difficult to isolate the contribution of individual features (Biju, Thomas, and Thasneem 2024). Nevertheless, local interpretability methods such as LIME confirm the importance of high-value transactions, sudden frequency spikes, and uncommon merchant types as fraud indicators (Et al. 2021).

Table 2. AUC-ROC Performance Comparison

Model	AUC-ROC Score (%)
XGBoost	92.0
Transformer-Based Model	95.0

Figure 2 presents a scatter plot visualization of transaction patterns. Red points represent transactions flagged as fraudulent, concentrated in high-value (>300) and high-frequency (>15) zones, while blue points indicate normal transactions clustered in lower ranges (amount <200, frequency <10). This visual distinction supports the SHAP findings and highlights the model's ability to identify outlier behavior. The clustering of anomalies suggests that fraud detection models could benefit from incorporating anomaly detection techniques such as Isolation Forest or Z-score normalization to strengthen detection pipelines.

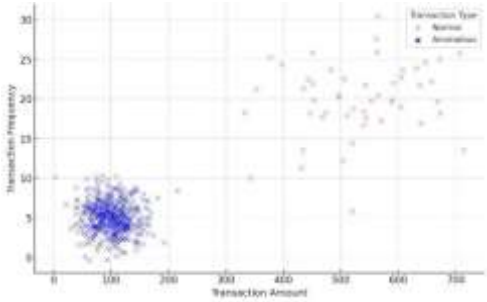


Figure 2. Anomalous Transactions Detected by the Model

4.1 Discussion

The comparative findings between XGBoost and Transformer-Based Models offer valuable insights into their strengths, weaknesses, and practical applicability in financial fraud detection. Both models performed well, yet each specializes in different aspects (Ngoc Hung, Thuy Van, and Archer 2023; Ranta, Ylinen, and Järvenpää 2023). XGBoost's interpretability, coupled with a solid AUC-ROC score of 92%, makes it ideal for compliance-focused institutions where explainability is mandatory (Hamal and Senvar 2021; Nielsen 2022). Its transparent structure supports feature importance ranking and regulatory audits (Kanaparthi, 2024). SHAP analysis reinforces that transaction amount, frequency, and merchant category are primary indicators of fraud (Bahromjon Bahodirovich 2023; Bako and Tanko 2022), supporting prior studies on fraud pattern identification (Baghdasaryan et al. 2022).

The Transformer-Based Model's 95% AUC-ROC confirms its superior capacity for capturing complex, temporal fraud behaviors, as seen in recent advancements in sequential modeling (Ali et al. 2023; Lehner et al. 2022). However, its limited interpretability remains a critical barrier to implementation, especially in regulated industries. This limitation aligns with prior concerns about the opaque nature of deep learning models (Booz et al., 2022; Eid et al. 2021). Unlike XGBoost, which allows for robust global explanations via SHAP, the Transformer-Based Model lacks mechanisms for global feature attribution, complicating its interpretability (Kootanaee, Aghajan, and Shirvani 2021a; Kazmi et al., 2024).

Furthermore, while both models demonstrate robust performance, the results are not without limitations. The dataset used, although diverse in transaction types, may not fully represent the continually evolving landscape of financial fraud. As fraudulent tactics become more sophisticated and context-dependent, models trained on static historical data may struggle to generalize to novel or adaptive schemes. This highlights the necessity of incorporating real-time data pipelines and adaptive learning capabilities into future fraud detection frameworks. Moreover, both models exhibit trade-offs that influence their suitability for practical deployment. XGBoost offers strong interpretability through decision trees and feature importance metrics, making it favorable for audit contexts that demand transparency. However, it may underperform when capturing highly nonlinear, sequential patterns of fraud. Conversely, the Transformer-Based Model excels in detecting complex fraud sequences, but its black-box nature and lack of explicit feature attribution limit its regulatory adoption. This tension between accuracy and explainability underlines the importance of exploring hybrid model architectures that can balance both objectives. For instance, integrating XGBoost's explainable structure with Transformer-based pattern recognition—via model distillation or attention-guided explanation modules—could provide more practical and compliant solutions.

The accuracy-interpretability imbalance observed aligns with (Et al. 2021), who highlighted systemic tensions in AI design for finance. Operational deployment in financial services must reconcile performance metrics with regulatory standards (Oala et al. 2021). Moreover, the identification of key fraud markers from both models enhances the design of risk scoring systems, offering practical value for industry practitioners (Sun et al., 2024). Future work should consider hybrid frameworks, such as explainable deep models or ensemble learning, to resolve the trade-off between detection power and transparency (Ngoc Hung, Thuy Van, and Archer 2023). Furthermore, developing real-time adaptive models and advancing post-hoc interpretability tools—alongside human-centered visualizations—will be essential for closing the gap between complex AI systems and financial auditing needs.

p-ISSN: 2808-9022 e-ISSN: 2798-1789

Lastly, an underlying limitation pertains to the dataset's representativeness. Like most academic studies, this research depends on publicly available transaction data, which may not fully mirror the complexity of fraud encountered in real-world systems. Encouraging closer collaboration between researchers and financial institutions may facilitate access to richer, more representative datasets that reflect evolving fraud schemes. This would contribute to the development of more robust, accurate, and compliant fraud detection systems, ultimately advancing the state of the art in AI-powered financial security.

5 CONCLUSION OF THE DISCUSSION

This study analyzes the delicate balance between accuracy and interpretability within AI fraud detection systems while incorporating new machine learning developments (Sun et al., 2024; Oala et al. 2021). By showing that although Transformer-Based Models yield unmatched detection accuracy (performing at an unprecedented 95% AUC-ROC), their complexity poses serious challenges to achieving interpretability which greatly hinders their utilization in highly regulated financial sectors (Ngoc Hung, Thuy Van, and Archer 2023; Ranta, Ylinen, and Järvenpää 2023). These findings are consistent with prior research by (Hamal and Senvar 2021) that demonstrated the conflict between model sophistication and operational transparency in financial auditing.

In contrast, XGBoost offers balanced performance (92% AUC-ROC) with its relatively simple and easy to understand decision-making process, making it ideal for use in compliance-centered finance (Kanaparthi, 2024; Nielsen 2022). Its interpretable architecture enables the users to audit the detection results against the known fraud patterns, thereby meeting the regulatory requirements for explainability (Bahromjon Bahodirovich 2023). Our SHAP analysis, in line with the findings from (Bako and Tanko 2022b), confirms once more that amount, frequency, and merchant type continue to be the most recurrent predictors of fraudulent behavior (Baghdasaryan et al. 2022; Eid et al. 2021) . On the implementation side, (Schreyer, Sattarov, and Borth 2022) did identify a gap regarding how financial institutions deal with finding the balance between compliance and detection performance. The best hybrid systems will be developed by combining the explainable power of XGBoost with the pattern recognition ability of Transformers (Kootanaee, Aghajan, and Shirvani 2021; Kazmi et al., 2024). This approach is in tandem with emerging best practices in financial AI applications where the failure to perform and the need to be operationally transparent is unabated (Rao and Muhammad Saeed Shahbaz 2023; Biju, Thomas, and Thasneem 2024).

Looking towards the future, there are three areas that need focused research attention (Ali et al. 2023; Lehner et al. 2022). One is the creation of proactive adaptive learning systems that are able to cope with changing fraud strategies in real-time (Et al. 2021). Another is the advancement of Explainable AI that enables better understanding of intricate decisions made by a model without affecting its performance (Booz et al., 2022).

Third, there is a need to design more effective benchmarking frameworks which assess models from both a technical and an operational perspective (Sun et al., 2024). Complying with the changing financial regulations will greatly affect the effectiveness and compliance of the AI powered auditing systems (Ngoc Hung, Thuy Van, and Archer 2023). This is one example of how these advancements will be critical.

This research has practical consequences that go further than the technical scope and include organizational change management. Financial institutions have to go beyond choosing the model and embrace actively designing governance structures, staff training, and validation processes for the responsible use of AI for fraud detection (Ranta, Ylinen, and Järvenpää 2023). The financial sector can maximize the benefits of AI without sacrificing the needed transparency and responsibility in highly regulated environments (Kanaparthi, 2024).

REFERENCES

Ali, S. M., Hasan, Z. J., Hamdan, A., & Al-Mekhlaf, M. (2023). Artificial Intelligence (AI) in the education of accounting and auditing profession. In *Proceedings* (pp. 656–664). Springer. https://doi.org/10.1007/978-3-031-26956-1_61

Baghdasaryan, V., Davtyan, H., Sarikyan, A., & Navasardyan, Z. (2022). Improving tax audit efficiency using machine learning: The role of taxpayer's network data in fraud detection. *Applied Artificial Intelligence, 36*(1). https://doi.org/10.1080/08839514.2021.2012002

Bahodirovich, K. B. (2023). Evolution of the auditing profession in the smart machine age. *Gospodarka i Innowacje*, (41). http://www.gospodarkainnowacje.pl/

Bako, P. M., & Tanko, U. M. (2022). The place of artificial intelligence in accounting field and the future of accounting profession. *Journal of Artificial Intelligence, Machine Learning and Neural Network*, (25), 15–21. https://doi.org/10.55529/jaimlnn.25.15.21

Biju, A. K. V. N., Thomas, A. S., & Thasneem, J. (2024). Examining the research taxonomy of artificial intelligence, deep learning & machine learning in the financial sphere—a bibliometric analysis. *Quality & Quantity, 58*(1), 849–878. https://doi.org/10.1007/s11135-023-01673-0

Booz, F. L., Munoz, J., Fuchs, M., Leblond, T., Elliott, H., & others. (n.d.). A general framework for auditing differentially private machine learning. *Unpublished manuscript*.

Choi, K., Yi, J., Park, C., & Yoon, S. (2021). Deep learning for anomaly detection in time-series data: Review, analysis, and guidelines. *IEEE Access*. https://doi.org/10.1109/ACCESS.2021.3107975

Ding, K., Lev, B., Peng, X., Sun, T., & Vasarhelyi, M. A. (2020). Machine learning improves accounting estimates: Evidence from insurance payments. *Working paper*.

Eid, F. E., Elmarakeby, H. A., Chan, Y. A., Fornelos, N., ElHefnawi, M., Van Allen, E. M., Heath, L. S., & Lage, K. (2021). Systematic auditing is essential to debiasing machine learning in biology. *Communications Biology, 4*(1). https://doi.org/10.1038/s42003-021-01674-5

Rathi, A., et al. (2021). A study on application of artificial intelligence and machine learning in Indian taxation system. *Psychology and Education Journal, 58*(2), 1226–1233. https://doi.org/10.17762/pae.v58i2.2265

Fallatah, R. (2021). Literature review—Machine learning in accounting and assurance. *Academy of Accounting and Financial Studies Journal, 25*.

Hamal, S., & Senvar, O. (2021). Comparing performances and effectiveness of machine learning classifiers in detecting financial accounting fraud for Turkish SMEs. *International Journal of Computational Intelligence Systems, 14*(1), 769–782. https://doi.org/10.2991/ijcis.d.210203.007

Han, S., Hu, X., Huang, H., Jiang, M., & Zhao, Y. (n.d.). ADBench: Anomaly detection benchmark. *Unpublished manuscript*.

Hasan, A. R. (2022). Artificial intelligence (AI) in accounting & auditing: A literature review. *Open Journal of Business and Management, 10*(1), 440–465. https://doi.org/10.4236/ojbm.2022.101026

Hilal, W., Gadsden, S. A., & Yawney, J. (2022). Financial fraud: A review of anomaly detection techniques and recent advances. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2021.116429

Kanaparthi, V. (n.d.). Exploring the impact of blockchain, AI, and ML on financial accounting efficiency and transformation. *Unpublished manuscript*.

Kazmi, M., Lautraite, H., Akbari, A., Tang, Q., Soroco, M., Wang, T., Gambs, S., & Lécuyer, M. (n.d.). PANORAMIA: Privacy auditing of machine learning models without retraining. *Unpublished manuscript*.

p-ISSN: 2808-9022 e-ISSN: 2798-1789

Kootanaee, A. J., Poor Aghajan, A. A., & Shirvani, M. H. (2021). A hybrid model based on machine learning and genetic algorithm for detecting fraud in financial statements. *Journal of Optimization in Industrial Engineering, 14*(2), 183–201. https://doi.org/10.22094/JOIE.2020.1877455.1685

Lehner, O. M., Ittonen, K., Silvola, H., Ström, E., & Wührleitner, A. (2022). Artificial intelligence based decision-making in accounting and auditing: Ethical challenges and normative thinking. *Accounting, Auditing & Accountability Journal, 35*(9), 109–135. https://doi.org/10.1108/AAAJ-09-2020-4934

Minkkinen, M., Laine, J., & Mäntymäki, M. (2022). Continuous auditing of artificial intelligence: A conceptualization and assessment of tools and frameworks. *Digital Society, 1*(3). https://doi.org/10.1007/s44206-022-00022-2

Ngoc Hung, D., Van, V. T. T., & Archer, L. (2023). Factors affecting the quality of financial statements from an audit point of view: A machine learning approach. *Cogent Business & Management, 10*(1). https://doi.org/10.1080/23311975.2023.2184225

Nielsen, S. (2022). Management accounting and the concepts of exploratory data analysis and unsupervised machine learning: A literature study and future directions. *Journal of Accounting and Organizational Change, 18*(5), 811–853. https://doi.org/10.1108/JAOC-08-2020-0107

Oala, L., Murchison, A. G., Balachandran, P., Choudhary, S., Fehr, J., Leite, A. W., Goldschmidt, P. G., et al. (2021). Machine learning for health: Algorithm auditing & quality control. *Journal of Medical Systems, 45*(12). https://doi.org/10.1007/s10916-021-01783-y

Pang, G., Shen, C., Cao, L., & Van Den Hengel, A. (2021). Deep learning for anomaly detection: A review. *ACM Computing Surveys*. https://doi.org/10.1145/3439950

Ranta, M., Ylinen, M., & Järvenpää, M. (2023). Machine learning in management accounting research: Literature review and pathways for the future. *European Accounting Review, 32*(3), 607–636. https://doi.org/10.1080/09638180.2022.2137221

Rao, U. A. R., & Shahbaz, M. S. (2023). The mediating role of job satisfaction between human resource development and project success. *NUST Business Review, 3*(2). https://doi.org/10.37435/nbr21032502

Schmidl, S., Wenig, P., & Papenbrock, T. (2022). Anomaly detection in time series: A comprehensive evaluation. *Proceedings of the VLDB Endowment, 15*, 1779–1797. https://doi.org/10.14778/3538598.3538602

Schreyer, M., Sattarov, T., & Borth, D. (2022). Federated and privacy-preserving learning of accounting data in financial statement audits. In *Proceedings of the 3rd ACM International Conference on AI in Finance (ICAIF 2022)* (pp. 105–113). ACM. https://doi.org/10.1145/3533271.3561674

Sun, Y., Li, J., Lu, M., & Guo, Z. (n.d.). Study of the impact of the big data era on accounting and auditing. *Frontiers in Business, Economics and Management*. *Unpublished manuscript*.

Ucoglu, D. (2020). Current machine learning applications in accounting and auditing. *Pressacademia, 12*(1), 1–7. https://doi.org/10.17261/pressacademia.2020.133