

Exploring Data-Driven Approach for Financial Fraud Detection: A Comprehensive Literature Review

Qayoom Abdul¹, Wu Yadong², Umair Saeed³

¹PhD Scholar, School of Computer Science and Technology, Southwest University of Science and Technology (SWUST), Mianyang, P.R-China.

²Professor, School of Computer Science and Technology, Southwest University of Science and Technology (SWUST), Mianyang, P.R-China.

³PhD Scholar, Department of Computer Science, Bahria University Islamabad Campus, Islamabad-Pakistan.
aqbuzdar@mails.swust.edu.cn, wyd028@163.com, umairsaeedmixit@gmail.com

DOI: 10.5281/zenodo.14338985

ABSTRACT

Financial fraud detection has emerged as a critical area of research with the growing complexity and scale of fraudulent activities in the financial sector. Traditional methods of fraud detection, which are based on rule-based systems and manual oversight, fail to capture the dynamic and sophisticated nature of modern fraud schemes. This comprehensive literature review examines data-driven approaches that take into account the advancement of machine learning, artificial intelligence, and big data analytics to improve fraud detection. Some of the key methodologies covered are supervised, unsupervised, and hybrid models. The survey reflects growing usage in neural networks, ensemble methods, and anomaly detection techniques, emphasizing their performance in identifying complex fraud patterns in different financial datasets. Discussions include the difficulties with unbalanced datasets, evolving tactics for frauds, and requirements for explainability that remain future areas of interest. Drawing upon recent relevant research work, this review synthesis aims at informing readers concerning the landscape evolution in fraud detection against finances and presenting possible innovations in order for these to remain robust yet adaptive, clear, and transparent in nature.

Keywords: Fraud Detection, Machine Learning, Deep Learning, Fraud Detection Systems, Federated Learning, Financial Fraud Detection.

Cite as: Qayoom Abdul, Wu Yadong, & Umair Saeed. (2024). Exploring Data-Driven Approach for Financial Fraud Detection: A Comprehensive Literature Review. *LC International Journal of STEM*, 5(3), 64–72. <https://doi.org/10.5281/zenodo.14338985>

INTRODUCTION

Financial fraud detection has emerged as a highly critical area of research and practice, especially in contexts where complexities in financial transactions are increasing and digital payment systems are becoming more prominent. The financial sector, in particular, is one of the most vulnerable institutions to fraudulent activities, since such activities can lead to huge economic losses and thereby undermine consumer trust. The detection approaches become inefficient since fraudsters keep developing newer sophisticated schemes as more reliance is on traditional manual processing and rule-based systems. Inevitably, inadequacies call for an evolution to data-driven solutions with algorithms and machine learning that could strengthen detection approaches (Zhao & Bai, 2022; Kumar, 2024; Baesens et al., 2021).

The significance of financial fraud detection extends beyond the immediate loss in financial terms. Instead, it reaches a wider scale in regulatory compliance, risk management, and integrity of financial systems. In fact, incorporating data-driven methodologies into fraud detection systems enhances not only the accuracy but also the speed of detection, thereby allowing institutions to take proactive steps in the face of possible threats (Albashrawi, 2022; Hanae, 2023). As such, it becomes pertinent to examine these methodologies in further detail to shape more stringent frameworks for mitigating the developing aspects of financial fraud.

Financial fraud detection has always been dependent on statistical methods and audits, but the emergence of big data and learning machines transformed this field to allow analysis based on massive datasets to uncover hidden patterns that could indicate fraudulent behavior. Technics used in this type include decision trees, support vector machines, and logistic regression which can be used in determining if a transaction is legitimate or not based on its previous historical data (Chen, 2024; Liang, 2023; Xie et al., 2021).

These developments have been linked to a shift from traditional statistical-based methods to more complex algorithms in machine learning. The recent literature reports excellent results based on ensemble learning methods and neural networks in the detection of complex fraud patterns (Ngai et al., 2011; Liao et al., 2022; Chen & Wu, 2022). This shift underscores the necessity for continuous innovation in the methods of detection to keep pace with the sophistication of the tactic's fraudsters employ.

This encompasses multiple techniques using data mining and machine learning to detect fraud through data-driven means. In fraud detection, the approach will be based on analysis of transaction data that can unveil patterns of anomalies unusual to those known patterns of behavior. The improvement of machine learning models has also been evidenced when class imbalance in fraud detection datasets was corrected by applying the Synthetic Minority Over-sampling Technique (SMOTE) (Zhao & Bai, 2022; Pozzolo et al., 2018).

Besides, unsupervised learning methods like clustering and anomaly detection can detect unknown fraud patterns without the requirement of labeled data. This is especially useful in situations where there are few examples of fraud that are labeled (Wang, 2023; Hamza et al., 2023; Li et al., 2021). Data-driven approaches implemented into these systems improve detection rates and decrease false positives, hence boosting the overall effectiveness of fraud detection systems.

LITERATURE REVIEW

The literature on data-driven approaches for financial fraud detection has been quite evolved, showing different methodologies and challenges that researchers and practitioners face in this domain. Starting with the foundational work by (Chimonaki et al., 1970), the authors present a comprehensive classification framework that employs data mining techniques to detect financial statement fraud. Their investigation into feature selection and comparative studies using multicriteria analysis is in the context of laying the stage for understanding complexity and necessity in intelligent fraud detection practices. In 2015, (Lari Dashtbayaz & Mohammad, 2015) expanded on this base by emphasizing the critical role of fraud indicators within financial statements. They emphasize the decreasing success rates of detection mechanisms and suggest adaptive management processes that evolve with the fraud patterns across industries. This call for a more responsive approach to fraud detection highlights the need for continuous learning and adaptation in methodologies.

(West et al., 2015) further expands the discourse in terms of developing customized computational intelligence methods in several fraud categories. They observed that algorithmic tuning plays an important role in improving its performance. According to their view, previous research work has the disadvantage of representing several fraud types inadequately, as evidenced by the scarce literature of money laundering and securities fraud types. Therefore, it can be found out that more studies are necessary, including multiple fraud types beyond financial statement fraud.

Moving in lock step, as the postpandemic landscape of financial fraud detection advanced; to the growing complexity of stealth, Zhu et al (2021) addressed challenges towards developing a model that enables smooth multiple-source information integration of ever-innovative fraud patterns developed during the pandemic, pushing into the future. Insight developed during a discussion on hidden challenges created toward fraud detection calls upon one to seek advanced analytics as in knowledge graphs which increase the capability of tracking changed fraud schemes in any system.

In 2022, (Isangediok & Gajamannage, 2022) examined various machine learning techniques for fraud detection with an emphasis on the anomaly detection aspect of fraud cases. They highlight the problems caused by imbalanced datasets and the need for hybrid techniques that combine statistical methods with machine learning. Their results emphasize the need for robust models that can learn effectively from limited fraud observations, a recurring theme in the literature.

(Xu et al., 2023) further delve into deep learning applications in fraud detection, discussing the advantages and limitations of conventional methods versus deep learning techniques. They emphasize that training datasets should address the rare occurrence of fraud cases as it can significantly impact classification algorithms. This exploration of deep learning capabilities reflects a growing trend toward leveraging advanced technologies to improve fraud detection processes.

Recent research, such as by (Vivek et al., 2023), focused more on the use of real-time fraud detection machine learning models. Their work, on ATM fraud detection by using streaming data analytics, describes the need for scalable and efficient systems that cope with imbalances in data and can respond in due time to fraudulent activities. If-supervised and unsupervised models is a new trend towards innovative solutions that can function well even in situations of limited data availability.

Awosika et al., (2023) discuss transparency and privacy at the crossroads of financial fraud detection, arguing for explainable AI and federated learning approaches. They emphasize the need for comprehensive models that consider multiple factors influencing fraud for a proper understanding of fraudulent behavior, thereby enriching the dialogue on effective risk mitigation strategies.

Together, these studies highlight the multi-faceted nature of financial fraud detection and reveal an ongoing evolution in methodologies and the urgent need for adaptive, intelligent systems to address both current and emerging challenges in the field.

It should be noted that Luo et al., (2023), in their discussion, do give meaning to this by elaborating on the analysis of fraud detection using AI in decentralized finance. They consider data-scarcity and skewed-data-related factors. They present the need for sound approaches that can adapt to a DeFi environment with an ability to handle issues far superior to traditional data-gathering methods.

AI-driven methods, as presented by Narayan et al., leverage machine learning models to detect fraudulent activities with a high degree of accuracy, even in sparse data environments (Narayan et al., 2024). Mothukuri et al. have proposed a multi-model system that aggregates different factors for trust

scoring DeFi projects. This approach reduces the effect of skewed datasets by incorporating diverse data sources, including social media sentiment (Mothukuri et al., 2024).

The review of some recent studies is presented in Table 1 describing the approaches used in the study, type of the fraud, dataset and the key findings of these studies.

Table 1: A review of some available studies on fraud detection.

Study	Approach	Fraud Type	Dataset	Key Findings
Dal Pozzolo et al. (2015)	Class Imbalance Learning (Under-sampling)	Credit card fraud	Public credit card dataset	Addressed class imbalance challenges, showing improvement in fraud detection using under-sampling techniques.
Jurgovsky et al. (2018)	Recurrent Neural Networks (RNNs)	Credit card fraud	European card dataset	Demonstrated that RNNs effectively capture sequential transaction patterns to improve detection accuracy.
Wang et al. (2019)	Semi-supervised graph attentive network (SemiGNN)	Bank fraud	Real-world dataset (dataset from AliPay)	Connected the labeled and unlabeled data through their social relations. The node-level attention can better correlate neighbors and the view-level attention can better integrate different views.
Fiore et al. (2019)	Generative Adversarial Networks (GAN)	Credit card fraud	Credit card dataset	Combined machine learning with rule-based systems to reduce false negatives in fraud identification.
Chu & Yong (2021)	Machine Learning	Accounting fraud	Financial dataset	Explored real-time data analysis to detect fraudulent patterns in e-commerce, emphasizing scalability.
Kaur, Rani, & Kalra (2022)	Blockchain with AI	Transaction validation	Healthcare adapted to finance	Proposed a blockchain-based predictive model, emphasizing transparency and fraud prevention.
Zheng et al. (2024)	k-means clustering mining algorithm.	Machine learning for fraud prevention	Stock exchange dataset	Compared with traditional accounting fraud identification methods, the overall misjudgment rate of data mining algorithms based on smart cities has decreased by 3%. Data mining algorithms can improve the accuracy of accounting fraud and contribute to audit objectivity and effectiveness.

CHALLENGES IN FRAUD DETECTION

Despite the advancements made in data-driven approaches, a number of challenges remain with financial fraud detection. One of these is the imbalanced dataset problem, which has fraudulent transactions far outnumbered by legitimate ones. These biased models cannot identify fraud accurately because of this imbalance (Meng, 2022; Nobel, 2024). Researchers have suggested many strategies to address this issue, such as cost-sensitive learning and ensemble methods that combine multiple algorithms to improve the detection performance (Kumar & Nalini, 2021; Ileberi et al., 2022).

Another challenge is that fraud schemes are dynamic and change with any detection efforts. This resulted in the demand to construct adaptive models that can learn from new data inputs for adapting to new changes in fraud patterns (Shoetan, 2024; Yang et al., 2020). Finally, the interpretability of the machine learning models has also emerged as a major challenge; for stakeholders to have the decision-making process of such algorithms, they should gain the trust and to maintain adherence to regulatory standards (Baesens et al., 2021; Meng, 2022). Othman (2021) mentions that historical financial statement data may not be the best way to use it because fraud detection models need to adapt to the specific conditions of firms. The study supports the incorporation of governance factors in fraud detection models since the deficiencies in corporate governance have been related to financial scandals. Additionally, Zhou et al. (2021) address the limitations of conventional rule-based expert systems for the identification of financial fraud, particularly as the scale of financial data keeps expanding. They propose a distributed big data approach that utilizes graph embedding algorithms to enhance the classification and prediction of fraudulent activities. This innovative approach underscores the necessity for adaptive analytical models that can keep pace with the evolving nature of financial fraud.

FUTURE DIRECTIONS IN FRAUD DETECTION RESEARCH

Looking ahead, financial fraud detection is going to see further innovation with advancements in artificial intelligence and machine learning. Future research is likely to focus on the integration of deep learning techniques, which have shown promise in capturing complex relationships within data (Shoetan, 2024; Aslam, 2024). Moreover, exploration into hybrid models, which use more than one algorithm, may provide improved detection capabilities, especially in real-time fraud detection. (Hanae, 2023; Carcillo et al., 2018).

In addition, the implementation of XAI in the fraud detection system will be a must to ensure transparency and accountability. Since most financial institutions are switching towards the automated systems, the rationale behind the fraud detection decisions should be explained (Meng, 2022; Ahmed et al., 2021). With the concentration on interpretability and more efforts towards rectifying the data imbalance and model adaptability, it shall determine the landscape of the future of financial fraud detection.

The future research in the financial fraud detection domain lies in the continuous evolution of data-driven techniques and the inclusion of emerging technologies. Gupta & Gill (2012) provide a framework for data mining to prevent and identify financial statement fraud, stating that new methodologies are urgently needed to be able to adapt to changing fraud patterns. This framework serves as a initial point for additional research that tries to improve the precision and effectiveness of fraud detection systems.

Big data analytics with the involvement of machine learning is most probably to be game-changing fraud detection practices. According to Abrol (2023), big data analytics is being seen as an enabler in the process of detection of financial fraud. Big data analytics allows organizations to scan high volumes of information, including insights and patterns that could have remained unnoticed otherwise. Such is of utmost importance to those looking to reduce risks associated with financial fraud.

CONCLUSION

In conclusion, data-driven approaches to financial fraud detection represent a dynamic and highly evolving field that is imperative to safeguard the integrity of financial systems. With ever-changing tactics by fraudsters, the need for more sophisticated methodologies of detection will also be paramount. By using the power of machine learning and data mining, financial institutions will be able to enhance their detection and prevention capabilities of frauds, thereby building a much more trusted and secure financial ecosystem. The literature in data-driven approaches toward financial fraud detection reflects a wide range of methodologies, challenges, and future directions. The data mining and machine learning application techniques has substantially improved the ability to identify and prevent financial fraud, while forensic accounting continues to play a significant role in this domain. As financial fraud evolves, ongoing research and innovation will be essential in developing adaptive and effective fraud detection systems that safeguard the integrity of financial markets.

CONFLICT OF INTEREST

There is no conflict of interest of all authors.

REFERENCES

- Abrol, S. (2023). Role of big data analytics in financial fraud detection-a bibliometric analysis. *Corporate Governance Insight*, 5(1), 82-111. <https://doi.org/10.58426/cgi.v5.i1.2023.82-111>
- Ahmed, M., Ansar, K., Muckley, C., Khan, A., Anjum, A., & Talha, M. (2021). A semantic rule based digital fraud detection. *Peerj Computer Science*, 7, e649. <https://doi.org/10.7717/peerj-cs.649>
- Albashrawi, M. (2022). Detecting financial fraud using data mining techniques: a decade review from 2004 to 2015. *Journal of Data Science*, 14(3), 553-570. [https://doi.org/10.6339/jds.201607_14\(3\).0010](https://doi.org/10.6339/jds.201607_14(3).0010)
- Aslam, F. (2024). Advancing credit card fraud detection: a review of machine learning algorithms and the power of light gradient boosting. *AJCST*. <https://doi.org/10.11648/ajcst.20240701.12>
- Awosika, T., Mani Shukla, R., & Pranggono, B. (2023). Transparency and Privacy: The Role of Explainable AI and Federated Learning in Financial Fraud Detection
- Baesens, B., Höppner, S., & Verdonck, T. (2021). Data engineering for fraud detection. *Decision Support Systems*, 150, 113492. <https://doi.org/10.1016/j.dss.2021.113492>

Carcillo, F., Pozzolo, A., Borgne, Y., Caelen, O., Mazzer, Y., & Bontempi, G. (2018). Scarff : a scalable framework for streaming credit card fraud detection with spark. *Information Fusion*, 41, 182-194. <https://doi.org/10.1016/j.inffus.2017.09.005>

Chen, J. (2024). Lightgbm model for detecting fraud in online financial transactions. *Highlights in Science Engineering and Technology*, 93, 363-371. <https://doi.org/10.54097/xw0bng93>

Chen, Y. and Wu, Z. (2022). Financial fraud detection of listed companies in china: a machine learning approach. *Sustainability*, 15(1), 105. <https://doi.org/10.3390/su15010105>

Chimonaki, C., Papadakis, S., Vergos, K., & Shahgholian, A. (1970). Identification of financial statement fraud in Greece by using computational intelligence techniques

Chu, M. K., & Yong, K. O. (2021). Big data analytics for business intelligence in accounting and audit. *Open Journal of Social Sciences*, 9(9), 42-52.

Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015, December). Calibrating probability with undersampling for unbalanced classification. In 2015 IEEE symposium series on computational intelligence (pp. 159-166). IEEE.

Fiore, U., De Santis, A., Perla, F., Zanetti, P., & Palmieri, F. (2019). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences*, 479, 448-455.

Gupta, R. and Gill, N. (2012). Financial statement fraud detection using text mining. *International Journal of Advanced Computer Science and Applications*, 3(12). <https://doi.org/10.14569/ijacsa.2012.031230>

Hamza, C., Abrouk, L., Cullot, N., & Nicolas, C. (2023). Semi-supervised method to detect fraudulent transactions and identify fraud types while minimizing mounting costs. *International Journal of Advanced Computer Science and Applications*, 14(2). <https://doi.org/10.14569/ijacsa.2023.0140298>

Hanae, A. (2023). End-to-end real-time architecture for fraud detection in online digital transactions. *International Journal of Advanced Computer Science and Applications*, 14(6). <https://doi.org/10.14569/ijacsa.2023.0140680>

Ileberi, E., Sun, Y., & Wang, Z. (2022). A machine learning based credit card fraud detection using the ga algorithm for feature selection. *Journal of Big Data*, 9(1). <https://doi.org/10.1186/s40537-022-00573-8>

Isangediok, M. & Gajamannage, K. (2022). Fraud Detection Using Optimized Machine Learning Tools Under Imbalance Classes

Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P. E., He-Guelton, L., & Caelen, O. (2018). Sequence classification for credit-card fraud detection. *Expert systems with applications*, 100, 234-245.

Kaur, J., Rani, R., & Kalra, N. (2022, November). A Blockchain Enabled Predictive, Analytical Model for Fraud Detection in Healthcare Data. In *2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC)* (pp. 319-324). IEEE.

Kumar, D. (2024). Analyzing the impact of machine learning algorithms on risk management and fraud detection in financial institution. *International Journal of Research Publication and Reviews*, 5(5), 1797-1804. <https://doi.org/10.55248/gengpi.5.0524.1135>

Kumar, G. and Nalini, D. (2021). Accuracy analysis for logistic regression algorithm and random forest algorithm to detect frauds in mobile money transaction. *Revista Gestão Inovação E Tecnologias*, 11(4), 1228-1240. <https://doi.org/10.47059/revistageintec.v11i4.2182>

Lari Dashtbayaz, M. & Mohammad, S. (2015). Data search and discovery process for financial statement fraud.

Li, C., Ding, N., Dong, H., & Zhai, Y. (2021). Application of credit card fraud detection based on cs-svm. *International Journal of Machine Learning and Computing*, 11(1), 34-39. <https://doi.org/10.18178/ijmlc.2021.11.1.1011>

Liang, Z. (2023). A study of identification of corporate financial fraud using neural network algorithms in an information-based environment. *Informatica*, 47(9). <https://doi.org/10.31449/inf.v47i9.5220>

Liao, B., Huang, Z., Cao, X., & Li, J. (2022). Adopting nonlinear activated beetle antennae search algorithm for fraud detection of public trading companies: a computational finance approach. *Mathematics*, 10(13), 2160. <https://doi.org/10.3390/math10132160>

Luo, B., Zhang, Z., Wang, Q., Ke, A., Lu, S., & He, B. (2023). AI-powered Fraud Detection in Decentralized Finance: A Project Life Cycle Perspective.

Madhusudan, Narayan., P.R., Shukla., Rajeev, Kanth. (2024). 1. AI-Driven Fraud Detection and Prevention in Decentralized Finance. *Advances in finance, accounting, and economics book series*, doi: 10.4018/979-8-3693-6321-8.ch004

Meng, Q. (2022). Credit card fraud detection using feature fusion-based machine learning model. *Highlights in Science Engineering and Technology*, 23, 111-116. <https://doi.org/10.54097/hset.v23i.3208>

Ngai, E., Hu, Y., Wong, Y., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature. *Decision Support Systems*, 50(3), 559-569. <https://doi.org/10.1016/j.dss.2010.08.006>

Nobel, S. (2024). Unmasking banking fraud: unleashing the power of machine learning and explainable ai (xai) on imbalanced data. *Information*, 15(6), 298. <https://doi.org/10.3390/info15060298>

Othman, I. (2021). Financial statement fraud: challenges and technology deployment in fraud detection. *International Journal of Accounting and Financial Reporting*, 11(4), 1. <https://doi.org/10.5296/ijaf.v11i4.19067>

Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit card fraud detection: a realistic modeling and a novel learning strategy. *Ieee Transactions on Neural Networks and Learning Systems*, 29(8), 3784-3797. <https://doi.org/10.1109/tnnls.2017.2736643>

Shoetan, P. (2024). Transforming fintech fraud detection with advanced artificial intelligence algorithms. *Finance & Accounting Research Journal*, 6(4), 602-625. <https://doi.org/10.51594/farj.v6i4.1036>

Viraaji, Mothukuri., Reza, M., Parizi., James, L., Massa., Abbas, Yazdinejad. (2024). 2. An AI Multi-Model Approach to DeFi Project Trust Scoring and Security. doi: 10.1109/blockchain62396.2024.00013

Vivek, Y., Ravi, V., Anand Mane, A., & Ramesh Naidu, L. (2023). ATM Fraud Detection using Streaming Data Analytics.

Wang, D., Lin, J., Cui, P., Jia, Q., Wang, Z., Fang, Y., ... & Qi, Y. (2019, November). A semi-supervised graph attentive network for financial fraud detection. In *2019 IEEE international conference on data mining (ICDM)* (pp. 598-607). IEEE.

Wang, Y. (2023). Fraud detection based on fs-smote model for credit card. *Highlights in Science Engineering and Technology*, 70, 316-323. <https://doi.org/10.54097/hset.v70i.12479>

West, J., Bhattacharya, M., & Islam, R. (2015). Intelligent Financial Fraud Detection Practices: An Investigation.

Xie, Y., Li, A., Gao, L., & Liu, Z. (2021). A heterogeneous ensemble learning model based on data distribution for credit card fraud detection. *Wireless Communications and Mobile Computing*, 2021(1). <https://doi.org/10.1155/2021/2531210>

Xu, B., Wang, Y., Liao, X., & Wang, K. (2023). Efficient Fraud Detection Using Deep Boosting Decision Trees.

Yang, Y., Chen, R., Bai, X., & Chen, D. (2020). Finance fraud detection with neural network. *E3s Web of Conferences*, 214, 03005. <https://doi.org/10.1051/e3sconf/202021403005>

Zhao, Z. and Bai, T. (2022). Financial fraud detection and prediction in listed companies using smote and machine learning algorithms. *Entropy*, 24(8), 1157. <https://doi.org/10.3390/e24081157>

Zheng, X., Hamid, M. A. A., & Hou, Y. (2024). Data mining algorithm in the identification of accounting fraud by smart city information technology. *Heliyon*, 10(9).

Zhou, H., Sun, G., Sha, F., Wang, L., Hu, J., & Gao, Y. (2021). Internet financial fraud detection based on a distributed big data approach with node2vec. *Ieee Access*, 9, 43378-43386. <https://doi.org/10.1109/access.2021.3062467>

Zhu, X., Ao, X., Qin, Z., Chang, Y., Liu, Y., He, Q., & Li, J. (2021). Intelligent financial fraud detection practices in post-pandemic era.