# Title of the dissertation: Unveiling Deception in Healthcare: Machine Learning Approaches for Proactive Fraud Detection and Prevention in Medical Claims and Records

# CHAPTER 1: INTRODUCTION

Fraud is one challenge that endangers the healthcare sector because it takes money that should have gone into care provision and uses it to finance fraudsters’ extravagant lifestyles. These fraudulent practices thus siphon necessary funds from proper medical diagnosis and treatment, thus increasing the costs within the health care sector. Consequently, the pressure for healthcare organizations to create proper systems that shall help identify these improper payments and avoid them increases continuously.

The existing techniques used to prevent fraud in healthcare are majorly antithetic. The conventional techniques react when fraud has already happened through audits and such inefficient methods which are incapable of stopping fraud from taking place in the first place. This reactive course implies that considerable budgets are sunk into addressing problems instead of avoiding them in the first place, thus ineffectiveness and additional expenses.

As a result, there is a realization that initiative-taking measures for dealing with these challenges are wanted now more than ever. This chapter investigates the application of artificial intelligence in the healthcare industry, especially in fraud detection. Due to the use of innovative AI methods such as data mining and classification of patterns, it is feasible to detect the signs of fraud. The above AI operating systems can process massive data in a noticeably short duration and with higher precision as compared to operators Hence, a wide range of suspicious activities could be detected in real-time and therefore, fraudulent payments cannot occur in the first place.



Figure 1: AI in fraud detection

The introduction of AI in fraud detection represents a significant shift towards more dynamic and preventive approaches in the healthcare industry. It promises not only to reduce the incidence of fraud but also to optimize the allocation of resources, ensuring that funds are used effectively to enhance patient care and operational efficiency. This proactive use of technology to safeguard against fraud is becoming increasingly crucial as the healthcare sector seeks to maintain trust and integrity while managing costs effectively in a rapidly evolving landscape.Top of FormBottom of Form

**1.1 Background of the Study**

Healthcare fraud consequently contributes to the augmentation of costs in healthcare provision, particularly in Medicaid fraud, where billions of dollars are embezzled annually in the United States of America (Korcok, 1997). In addition, through a Hoffman survey, CIGNA HealthCare and Insurance groups presume that losses amounting to $ 80-& 100 billion accrue yearly due to fraud and improper billing (Hoffman, 1999). Examples of these frauds include billing for services not delivered or charging for the services offered and bypassing all the set procedures, compromising the healthcare systems and being a threat to patients.

Another prohibited activity associated with healthcare transactions is kickback, whereby it is unlawful to offer, pay, solicit, or receive anything of value in return for patient referrals that the arrangement of the measures of the healthcare programs will compensate. This kind of fraud may arise from self-generated corruption practices, fabricated treatments, unnecessary services, and improper business requests. In health facilities, some employees are unethical in their conduct; they embezzle money, bill services that were never rendered, and even compel patients to undertake treatments that they do not require so that they can be charged heavily.

Originally, more overt torture kind of offences were discernible within the context of healthcare fraud in that they included bribery and kickbacks. However, the Office of the Inspector General has lately initiated obscure cases, with courts expanding the definitions of anti-fraud statutes, which implies increasing risks of criminal charges (Steiner, 1993). For example, administrative policies have been prosecuted, such as failing to inform an insurance carrier that one has submitted a bill for payment while waiving a patient’s co-payment (Tomes, 1993).

According to the survey conducted by the Health Insurance Association of America in 1993, it was proved that diagnosis (43%) and billing services (34%) are most related to healthcare fraud activities. Identifying frauds and preventing them early does help insurance companies immensely in saving costs, but it also plays a huge part in containing the ever-increasing costs of healthcare. Such funds could otherwise have been used to diagnose and treat other diseases and illnesses.

The application of artificial intelligence in healthcare has, in recent years, been upgraded to incorporate machine learning in detecting fraud. K-means clustering, an unsupervised machine learning technique has been used by Agarwal (2023) in identifying fraudulent activities concerning medical insurance claims with the assistance of labelled data. Also, Johnson and Khoshgoftaar (2023) have developed a data-oriented approach that improves the effectiveness and accuracy of healthcare fraud detection utilizing Medicare claims data for supervised training (Johnson & Khoshgoftaar, 2023).

Furthermore, Mohammed (2023) introduces a novel system architecture to perform the identification and prevention of dishonesty in the layers of blockchain systems; however, the application of ML algorithms for examining overall medical data originating from sensors and transactions optimally (Mohammed et al., 2023). Besides enhancing detection, this methodology enhances the management of healthcare practices to prevent its potential qualities concerning the deteriorating of patient care standards and excessive costs in healthcare.

**1.2 Problem Statement**

The prehistoric methods of detecting fraud and money laundering that rely primarily on a manual system of rules and strict procedures and standards on threshold values are less effective against the challenges of hi-tech criminals. Such an approach is not sufficient because fraudster strategies are evolving rapidly due to the globalization of markets and instant data transfers. AI, applying ML and DL, presents an opportunity to uncover previously unknown patterns, find such anomalies, and recognize fraud opportunities in the context of the credit card with high levels of accuracy.

The primary potential victims of these illicit actions are healthcare organizations, which are critical to society as they provide critical medical care services to the population; such fine-tuning attacks lead to severe economic losses, the weakening of patient protection, and loss of community trust. However, the healthcare industry is still not exploring enough of the possibilities offered by such technologies as AI and ML in the field of fraud detection (Fernando et al., 2021). According to studies, cases of healthcare fraud occur to the tune of billions of dollars yearly, making it a resilient issue affecting healthcare organizations (Johnson & Khoshgoftaar, 2019).

Modern analytical methods and existing technologies focused on the use of the data mining paradigm call for more efficient management of healthcare fraud, including the interprofessional approach. Experiences and cases of fraud show that statistical methods and data mining methods are essential tools that are used to enhance the knowledge of this risk in the industry (Springer, n. d. ). However, the traditional rule-based model does not adapt in line with the advanced transformation of improved fraud schemes, as noted by Nassif et al. (2021). As for this scenario, more attention should be paid to improving existing fraud identification models with the help of AI and ML to prevent fraud in the sphere of healthcare, not reacting to it but actively avoiding the situations and cases described above.

**1.3 Research Significance**

This work contributes to creating a ‘’healthcare fraud detecting machine learning framework’’ to aid in detecting fractions within the healthcare industry. In this context, the proposed framework targets the proactive fraud detection limitation as a feature of the existing methods that seem unable to adapt to the constantly emerging tactics of fraudsters. In this way, applying the methods and approaches based on the most recent achievements in machine learning, this research aims to bring innovations to fraud detection methods on their speed, accuracy, and efficiency levels.

Adopting the above framework could significantly reduce financial costs due to fraudulent claims and improper billing practices. Most remarkably, it could enhance the quality of the offered healthcare services and their trustworthiness, thus preventing healthcare resources from being embezzled, while the patients’ needs are ignored. In conclusion, effectively implementing this framework can be a best practice for such projects worldwide, thus creating a new reference point for combating healthcare fraud and increasing confidence in healthcare systems.

**1.4 Aims and Objectives**

**Aim**

To develop and validate a comprehensive machine learning framework capable of proactively detecting and preventing fraud within healthcare systems.

**Objectives**

1. **To Develop and Integrate Predictive Models:** Implement advanced predictive models using both supervised and unsupervised machine learning techniques to identify potentially fraudulent activities.
2. **To Implement and Validate Real-Time Monitoring:** Establish a system for real-time fraud detection and assess its effectiveness compared to traditional methods.
3. **To Ensure Compliance and Enhance Explainability:** Ensure the framework adheres to legal and ethical standards, including data privacy regulations, and incorporate explainable AI to make the model's decisions transparent and understandable.

**1.5 Research Questions**

1. How can machine learning algorithms be applied to detect fraudulent activities in healthcare claims and records effectively?
2. What are the benefits of real-time fraud detection systems over traditional fraud detection methods in terms of accuracy, efficiency, and cost?
3. How can explainable AI be integrated into fraud detection systems to ensure compliance with ethical and legal standards?

**1.6 Research Methodology**

The proposed research will use primary data collected from the healthcare database and secondary data collected from the patient’s records database. Secondary data is important because it gives an informative background history when developing accurate AI machines to identify fraudsters.

This data will undergo several preprocessing steps before it is deployed to implement the following models. These steps are used to improve data quality and relevancy by cleaning, normalizing, and structuring the data, which is crucial for the success of the machine learning models. The incorporation of extensive preprocessing makes it possible to clean any data fed into the models to eliminate any vices that may lead to inconsistency and redundancies.

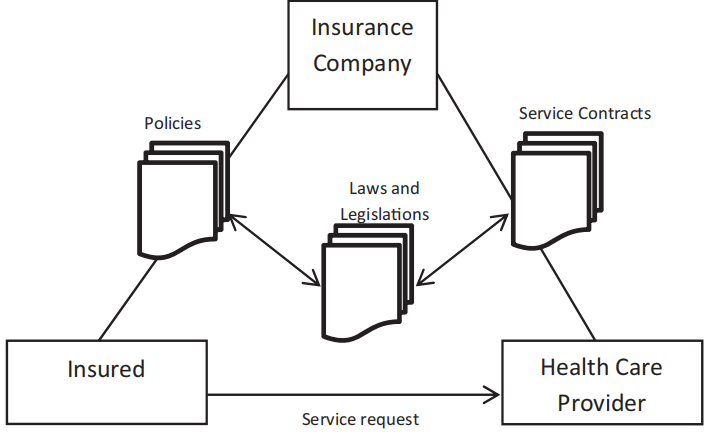


Figure 2: The health insurance payment model

Afterwards, depending on this carefully selected dataset, several machine-learning models will be built and trained. To assess the effectiveness of these models, a set of benchmarks will be applied, and several tests will be conducted against the above characteristics. The overall evaluation of the models will reveal if they are suitable for detecting and forecasting fraudulent events in healthcare transactions. Specifically, the well-developed testing methodology will confirm the effectiveness of these models for practical usage in the actual environment.

**1.7 Structure of the Dissertation**

This dissertation is structured as follows:

* **Chapter 1: Introduction** – Outlines the background, problem statement, significance, objectives, research questions, and methodology of the study.
* **Chapter 2: Literature Review** – Reviews existing literature on healthcare fraud and previous applications of machine learning in fraud detection.
* **Chapter 3: Methodology** – Details the methods used for data collection, model development, and validation.
* **Chapter 4: Results and Discussion** – Presents the findings of the study and discusses their implications.
* **Chapter 5: Conclusion and Recommendations** – Summarizes the study, discusses limitations, and suggests areas for future research.

**1.8 Conclusion**

The potential of machine learning in combating healthcare fraud is immense, offering a proactive approach to detect and prevent fraudulent activities. By enhancing the detection capabilities and ensuring compliance with regulatory standards, this research aims to contribute significantly to the security and integrity of healthcare systems worldwide. This dissertation will not only highlight the effectiveness of these machine learning techniques but also pave the way for their broader adoption in the healthcare industry.

# CHAPTER 2: LITERATURE REVIEW

Healthcare frauds detrimentally affect the health and economic framework of the U. S. healthcare system. This research, hence, applies ML tools as a data scientist to uncover and counter HC fraud for nationally guarded healthcare assets. The objective is to apply ML to detect fraudulent events in claims in the healthcare sector by studying oddities in the collected data. According to the hypothesis, the unseen patterns employing the usual approach might be identified by employing mechanized learning, lowering the loss and safeguarding the healthcare structure. According to Bauder and Khoshgoftaar (2017) and Agarwal (2023), it could be estimated that the fraud cost for the American healthcare system is tens of billions annually; this fraud involves anything from billing the patient when the service was never delivered to sophisticated kickback plans. Such unlawful activities pull vital funds away from federal programs, Medicare and Medicaid, skyrocket insurance costs, and escalate operating expenses for the companies (Matloob and Khan, 2019). These activities are coordinated by federal entities, with the CMS taking charge of legal actions against the fraud and the FBI leading in the investigation and prosecution of the crime (Shamitha, 2022; Gill and Aghili, 2020; Iqbal, 2022). Implementing ML technologies presents optimistic improvements in detecting fraud solutions (Lekkala, 2023). They should be able to process large data units to detect fraud cases, which are normally unnoticeable even by auditors. With this purpose, the current study seeks to support current endeavours to prevent healthcare fraud, thus protecting consumers, taxpayers, and the healthcare system in the United States.

## 2.1 Related work

Fraud detection is particularly important in the context of medical insurance, which requires intricate detection tools. Machine learning (ML) in these processes also presents impressive improvements in managing this problem. The following literature review aims to identify the current advancement and future trends in detecting healthcare fraud.

In his paper on major fraud types of medical insurance claims, Agarwal (2023) elaborately discusses this issue. Thus, utilizing K-means clustering, an unsupervised ML approach, Agarwal shows promising results in detecting fraud cases without using labelled data. Adaptive approaches are needed to detect fraudulent claims and minimize the impact of the healthcare system’s financial loss (Agarwal, 2023).

Johnson and Khoshgoftaar (2023) propose a data-driven architecture that considers the effectiveness and robustness of the tests designed for healthcare fraud identification. Using the Medicare claims, they build large-scale labelled datasets for analysing supervised learning, further celebrating these with new Provider summary features and presenting an extended data labelling approach. Their conclusions call attention to proper work in data preprocessing and the benefit of a data-focused approach to the ML process regarding healthcare fraud classification (Johnson and Khoshgoftaar, 2023).

Mohammed (2023) suggests a new approach to system architecture implemented using ML to identify and mitigate fraud cases in blockchain systems. Using the Random Forest algorithm, this two-step strategy can filter out the wrong values, highlight such transactions, and prove higher accuracy, reaction time, and scaling (Mohammed, 2023).

Duman (2022) discusses the application of the XGBoost technique in identifying Medicare fraud using traditional and ML techniques. His work outlines that XGBoost offers the best performance in metrics like AUC, precision, recall, and F1-score out of the tested techniques. Specifically, Duman underlines that Medicare loses about fifteen billion US dollars per year to fraud, stressing the value of public datasets in enhancing the existing levels of transparency and fraud detection (Duman, 2022).

In another article, Gill and Aghili (2020) discuss the topic of health insurance fraud detection and underline the acute demand for wise fraud detection solutions. They assess the characteristics of an ideal health insurance fraud detection application; they argue that the best solution in fraud cases should address the management of the integration of unstructured data and have a dynamic business plan (Gill and Aghili, 2020).

In 2020, Lennart Dangers integrated unsupervised learning to identify fresh fraud patterns without any prior labelling of large volumes of medical encounters and the numerous steganographic and symbiotic strategies used by the fraudsters. It shows that an audit of structured flows is attainable in an analogous method for healthcare data and contributes a useful instrument for insurance corporations to extend their auditing features (Dangers, 2020).

Aruleba and Sun (2023) have investigated incorporating such ML classifiers as the Decision Trees and Random Forests to determine healthcare fraud. It also reveals the viability of these techniques by using ensemble classifiers and performance metrics and shows how ML is useful in fighting healthcare fraud (Aruleba and Sun, 2023).

Roy(2022) used AI in healthcare data privacy, where a Random Forest algorithm achieved 92% accuracy in identifying threats to healthcare data privacy. Among the key enablers of telemedicine, this research reinforces the centrality of AI in creating a secure system to support digital health solutions (Roy, 2022).

Lekkala (2023) delved into the change that ML models introduced in combating healthcare fraud. The advancements in the application of ensemble methods and the use of deep learning models as the methods that can enhance the overall efficiency of fraud detection are underlined; special focus is paid to such features that can help identify frauds accurately (Lekkala, 2023).

Regarding the role of ML in detecting healthcare fraud, the following options are further expanded by Akbar et al. (2020) and Ho et al. (2020). Akbar describes the accuracy improvement in the decision tree classifier by the Extreme Gradient Boosting method. At the same time, Ho raises points on the ethical and regulatory considerations for using AI in health insurance (Akbar et al., 2020). These kinds of research add to the knowledge of the possibilities and problems of employing sophisticated approaches to prevent and control healthcare fraud efficiently.

# 2.1.1 Traditional Fraud Detection Methods

This paper seeks to establish the importance of fraud detection in protecting financial assets and ensuring the fidelity of health financial systems. In the past, a rule-based approach has been used, and although it has its effectiveness, it has also faced some drawbacks that require more flexible strategies and frameworks (Hassan, Aziz & Andriansyah, 2023).

Systems based on rules belong to the core of traditional fraudulent activity detection; they use rules and criteria designed to detect unusual patterns included in the initial data by professionals using historical data and typical fraud schemes. For instance, such systems may include alerts where the transaction magnitude exceeds set limits or originates from certain geographical areas. These systems are quite easy to install and make much sense from a compliance and audit perspective. It is easy to implement and immediately detect suspicious transactions, and it is cheaper than complex transactions. However, these systems also alert the transactions conducted according to varied factors, including amount, origin, or frequency (Kotagiri & Yada, 2024).

Nonetheless, they have the following demerits: To begin with, rule-based systems … are inherently fixed and ineffective in responding to dynamism in fraud trends without programmed changes – practices that are frequently tedious and slow in responding to the current trends in fraud. Due to their generalistic approach in which their ruleset is built to catch as much fraud as possible, this often leads to many false positives; this adds strain to the detection team and may even flag innocent transactions, which will be unpleasing for customers. These systems working based on the pattern and history makes them highly vulnerable to new or complex fraud schemes that the system has not seen before. The routine modifying and enhancing of these structures entails significant amounts of hand-work and usually costly professional advice.

Based on these challenges, a new notion crept into fraud detection, requiring new and enhanced solutions. The ever-changing fraud schemes, with special attention to the level of confrontation of the reported schemes, demand concepts and architectures with learning and prediction capabilities for fraud detection in real time (Kotagiri, 2023). It is characterized by their ability to examine large datasets methodically and logically and then draw conclusions based on trends and patterns from those datasets without strict programming. They can learn from each transaction continuously and monitor the fraud by checking for variations rather than using set rules. Implementing highly evolved systems that can process and analyse data in real time will help respond as soon as possible to fraud threats, thus reducing the time window for fraud and improving decision-making time (Kotagiri & Yada, 2024).

Combined with adaptive solutions, rule-based systems are more effective when used together. In contrast, rule systems deal with known threats; adaptive systems cover novel and developing ones. Not only is this approach effective in counteracting new fraud techniques, but the rates of false positives are minimized, and the reliance on updating systems manually or maintaining them is kept to a minimum. It can cope with the growing number and complexities of transactions than it used to before.

In conclusion, it can be stated that despite the significant importance of rule-based solutions and methods as well as traditional approaches to fraud detection, their shortcomings indicate the necessity of implementing new, more complex, and sophisticated models of fraud detection. Implementing machine learning and real-time analysis will enhance the ability to identify and prevent fraudulent activities, hence a more secure and safer monetary sector. (Patel, 2023; Wang et al. , 2020).

## 2.2 Gap in Existing Literature

For that matter, this project fills the following major gaps in the existing body of knowledge: This project applies a systematic approach that includes a broad range of machine learning and deep learning methods. It also focuses on model interpretability and the creation of procedures for real-time predictions. Here are the specific gaps addressed: Here are the specific gaps addressed:

* **Comparative Analysis Across Models:** It is also different from many pieces of research in which the comparison is usually conducted on a single or a few selected models at most. This approach offers vital information to elucidate the overall superiority and inferiority of these models precisely in the context of detecting healthcare fraud.
* **Integration of Model Explainability:** While there is a growing interest in model interpretability in the context of healthcare, the successful application of explanation techniques based on SHAP across multiple models is not very well described in the literature. This project covers this imbalance through the application of SHAP values on multiple machine learning models with the aim of increasing the interpretability of the fraud detection models.
* **Real-Time Detection and Continuous Learning:** Fraud detection is analysed by prior literature in a non-developmental manner, by training models on available examples. On the other hand, this project proposes a real-time detection pipeline that incorporates the option of model retraining occasionally. This innovation encompasses the idea that fraud in the healthcare industry evolves constantly and that models, including this industry, must change with time and adjust to evolving patterns.

# CHAPTER 3: METHODOLOGY

# References

Agarwal, S. (2023) 'An Intelligent Machine Learning Approach for Fraud Detection in Medical Claim Insurance: A Comprehensive Study', *Scholars Journal of Engineering and Technology*, 11(9), pp. 191-200.

Aruleba, I. T., & Sun, Y. (2023). Healthcare Fraud Detection Using Machine Learning. SSRN.

Bauder, R. A., & Khoshgoftaar, T. M. (2017). Medicare Fraud Detection Using Machine Learning Methods. In Proceedings of the 16th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE.

CIGNA (1999) *CIGNA HealthCare and Insurance Groups Web-site*. Available at: <http://www.insurance.ibm.com/insur/cigna.htm> (Accessed: 28 July 2024).

Dangers, L. (2020). Fraud- and Anomaly Detection in Healthcare – An Unsupervised Machine Learning Approach. Future Healthcare Journal, 7(1), 1-49 https://www.future- healthcare.pt/en/.

Duman, E. (2022). Implementation of XGBoost Method for Healthcare Fraud Detection. Techno-Science, 5(2), 69-75.

Fernando, T., Gammulle, H., Denman, S., Sridharan, S., & Fookes, C. (2021). 'Deep Learning for Medical Anomaly Detection—A Survey', ACM Computing Surveys (CSUR), 54, pp. 1-37. Available at: <https://doi.org/10.1145/3464423>

Gill, J. K., & Aghili, S. (2020). Health Insurance Fraud Detection. Master of Information Systems Assurance Management. Concordia University of Edmonton.

Hassan, M., Aziz, L. A. R., & Andriansyah, Y. (2023). The role artificial intelligence in modern banking: an exploration of AI-driven approaches for enhanced fraud prevention, risk management, and regulatory compliance. Reviews of Contemporary Business Analytics, 6(1), 110-132.

Health Care Financing Administration (1999) *Medicare fraud Website*. Available at: <http://www.hcfa.gov/medicare/fraud> (Accessed: 28 July 2024).

Johnson, J.M. and Khoshgoftaar, T.M. (2023) 'Data-Centric AI for Healthcare Fraud Detection', *SN Computer Science*, 4, Article 389. doi:10.1007/s42979-023-01809-x.

Korcok, M. (1997) 'Medicare, Medicaid fraud: A billion-dollar art form in the US', *Canadian Medical Association Journal*, 156(8), pp. 1195-1197.

Kotagiri, A., & Yada, A. (2024). Improving Fraud Detection in Banking Systems: RPA and Advanced Analytics Strategies. International Journal of Machine Learning for Sustainable Development, 6(1), 1-20.

Lekkala, L. R. (2023). Importance of Machine Learning Models in Healthcare Fraud Detection. Voice of the Publisher, 9, 207-215. doi:10.4236/vp.2023.94017

Matloob, I., & Khan, S. (2019). A Framework for Fraud Detection in Government Supported National Healthcare Programs. In Proceedings of the ECAI 2019 - 11th International Conference – Electronics, Computers and Artificial Intelligence, Pitesti, Romania, 27-29 June.

Mohammed, M.A., Boujelben, M. and Abid, M. (2023) 'A Novel Approach for Fraud Detection in Blockchain-Based Healthcare Networks Using Machine Learning', *Future Internet*, 15, p. 250.

Nassif, A. B., Talib, M. A., Nasir, Q., & Dakalbab, F. M. (2021). 'Machine Learning for Anomaly Detection: A Systematic Review', IEEE Access, 9, pp. 78658-78700. Available at: <https://doi.org/10.1109/ACCESS.2021.3083060>

Patel, K. (2023). Credit Card Analytics: A Review of Fraud Detection and Risk Assessment Techniques. International Journal of Computer Trends and Technology, 71(10), 69-79.

Roy, S. (2022). Privacy Prevention of Healthcare Data Using AI. Journal of Data Acquisition and Processing, 37(3).

Shamitha, S. K., & Ilango, V. (2022). A time-efficient model for detecting fraudulent health insurance claims using Artificial neural networks.

Springer (n.d.). International Journal of Data Science and Analytics. Available at: <https://www.springer.com/journal/41060/updates/23885458?gclid=EAIaIQobChMIhL7K-cS27wIVkMSGCh2EyAJGEAEYAiAAEgIjZ_D_BwE>

Steiner, J.P. (1993) *Healthcare Fraud, Waste, Abuse, and Safe Harbors: The Complete Legal Guide*. Chicago, Illinois: Probus Publishing Company.

Tomes, J.P. (1993) *Healthcare Fraud, Waste, Abuse, and Safe Harbors: The Complete Legal Guide*. Chicago, Illinois: Probus Publishing Company.

Wang, C., Wang, C., Zhu, H., & Cui, J. (2020). LAW: learning automatic windows for online payment fraud detection. IEEE Transactions on Dependable and Secure Computing, 18(5), 2122-2135.