***Dissertation Title***

*Unveiling Deception in Healthcare: Machine Learning Approaches for Proactive Fraud Detection and Prevention in Medical Claims and Records*

**Final Thesis**

In Partial Fulfillment

of the Requirements for the Degree of

Master in Computer Science

|  |  |  |
| --- | --- | --- |
| Student Name | : |  |
| Student ID | : |  |
| Supervisor | : |  |

**Abstract**

Healthcare fraud is a pervasive issue that costs billions annually, undermining the integrity of healthcare systems and burdening payers, providers, and patients. This dissertation, titled *Unveiling Deception in Healthcare: Machine Learning Approaches for Proactive Fraud Detection and Prevention in Medical Claims and Records*, explores the application of advanced machine learning techniques to address this critical problem.

The study focuses on developing and evaluating two machine learning models: Logistic Regression and Isolation Forest, to detect fraudulent activities within medical claims and records. The Logistic Regression model, a supervised learning approach, was employed to classify claims as either fraudulent or non-fraudulent based on labelled training data. Conversely, the Isolation Forest model, an unsupervised learning method, was utilized to identify anomalies within the dataset, which are indicative of potential fraud.

The research was conducted using a healthcare provider dataset, which underwent rigorous pre-processing, including feature scaling and data normalization, to ensure the models' accuracy and reliability. The performance of the models was assessed using key metrics such as accuracy, precision, recall, F1 score, and ROC AUC score. The Logistic Regression model achieved an accuracy of 78%, indicating a strong ability to differentiate between legitimate and fraudulent claims. The Isolation Forest model, while unsupervised, demonstrated a recall rate of 70%, making it a valuable tool for identifying suspicious activities that may not have been labelled in the dataset.

The findings of this study highlight the potential of machine learning to enhance fraud detection in healthcare, offering a proactive approach that could significantly reduce the incidence of fraud. The research also underscores the importance of integrating multiple models to provide a comprehensive solution, addressing both known and unknown fraudulent patterns.

The dissertation concludes with recommendations for future research, including the exploration of more advanced machine learning techniques, the incorporation of additional data sources, and the deployment of real-time fraud detection systems. This work contributes to the ongoing efforts to safeguard healthcare resources and ensure the delivery of quality care.

**Acknowledgements**

Use font: Times New Roman; font size: 12; line spacing: 1.5

**Contents**

[Abstract ii](#_Toc174808170)

[Acknowledgements iii](#_Toc174808171)

[Contents iv](#_Toc174808172)

[List of Tables vi](#_Toc174808173)

[List of Figures vii](#_Toc174808174)

[List of Acronyms viii](#_Toc174808175)

[1 Introduction 1](#_Toc174808176)

[1.1 Background Of The Study 2](#_Toc174808177)

[1.2 Problem Statement 3](#_Toc174808178)

[1.3 Research Significance 3](#_Toc174808179)

[1.4 Aims And Objectives 4](#_Toc174808180)

[1.5 Research Questions 4](#_Toc174808181)

[1.6 Research Methodology 4](#_Toc174808182)

[1.7 Structure Of The Dissertation 5](#_Toc174808183)

[1.8 Conclusion 6](#_Toc174808184)

[2 Literature Review/Related Work 7](#_Toc174808185)

[2.1 Comprehensive Overview of the Existing Literature 7](#_Toc174808186)

[2.1.1 Traditional Fraud Detection Methods 9](#_Toc174808187)

[2.2 Critical Analysis of Existing Studies [Gaps In Existing Literature] 10](#_Toc174808188)

[3 Methodology 12](#_Toc174808189)

[3.1 Data Collection and Preprocessing 12](#_Toc174808190)

[3.1.1 Data Collection 12](#_Toc174808191)

[3.1.2 Data Pre-processing 14](#_Toc174808192)

[3.2 ML/AI Model Development 15](#_Toc174808193)

[3.2.1 Predictive Models 15](#_Toc174808194)

[3.3 Real-Time Monitoring Implementation 18](#_Toc174808195)

[3.3.1 System Architecture 18](#_Toc174808196)

[3.3.2 Validation 18](#_Toc174808197)

[3.4 Compliance and Explainability 18](#_Toc174808198)

[3.4.1 Legal and Ethical Standards 18](#_Toc174808199)

[3.4.2 Explainable AI 18](#_Toc174808200)

[3.5 Evaluation of the Proposed System [Model Evaluation and Testing] 18](#_Toc174808201)

[3.5.1 Benchmarks 18](#_Toc174808202)

[3.5.2 Testing Methodology 19](#_Toc174808203)

[4 Experimental Results 20](#_Toc174808204)

[4.1 Experimental Setup 20](#_Toc174808205)

[4.2 Dataset Description 20](#_Toc174808206)

[4.3 Results 21](#_Toc174808207)

[4.3.1 Logistic Regression Results 21](#_Toc174808208)

[4.3.2 Isolation Forest Results 21](#_Toc174808209)

[4.3.3 Discussion of Results 22](#_Toc174808210)

[4.4 Comparison with Baseline Methods 22](#_Toc174808211)

[5 Conclusion and Future Work 23](#_Toc174808212)

[5.1 Conclusion 23](#_Toc174808213)

[5.2 Future Work 23](#_Toc174808214)

[6 References 25](#_Toc174808215)

[Appendix A. Model 27](#_Toc174808216)

[Appendix B. Title of Appendix. 29](#_Toc174808217)

**List of Tables**

[Table 1: Confusion Matrix For Logistic Regression 21](#_Toc174807005)

[Table 2: Confusion Matrix For Isolation Forest 22](#_Toc174807006)

**List of Figures**

[Figure 1: AI In Fraud Detection 1](#_Toc174807024)

[Figure 2: The Health Insurance Payment Model 5](#_Toc174807025)

**List of Acronyms**

Term Initial components of the term (examples are below)

FEC Forward Error Correction

FET Field Effect Transistor

Please insert one term per table row; this will ensure appropriate spacing and alignment between each term and its components. It will also allow you to sort the terms alphabetically.

**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, therefore increasing the costs within the healthcare sector. Consequently, the pressure for healthcare organizations to create appropriate 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 occurring 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

**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 (Agarwal, 2023). 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 (Aslam, 2024). 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 offenses 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 (Practice, 2022). 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 (Richard A. Bauder, The Detection of Medicare Fraud Using Machine Learning Methods with Excluded Provider Labels, 2018).

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 labeled data. Also, (Richard A. Bauder, Medicare Fraud Detection Using Machine Learning Methods, 2018) have developed a data-oriented approach that improves the effectiveness and accuracy of healthcare fraud detection utilizing Medicare claims data for supervised training (Richard A. Bauder, Medicare Fraud Detection Using Machine Learning Methods, 2018).

Furthermore, (Sanalkumar, 2022) 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 (Sanalkumar, 2022). Besides enhancing detection, this methodology enhances the management of healthcare practices to prevent its potential qualities concerning the deterioration of patient care standards and excessive costs in healthcare.

**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 (Prosper Kandabongee Yeng, 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 (Richard A. Bauder, The Detection of Medicare Fraud Using Machine Learning Methods with Excluded Provider Labels, 2018).

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. However, the traditional rule-based model does not adapt in line with the advanced transformation of improved fraud schemes, as noted by (Roy, 2022). 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.

**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.

**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.

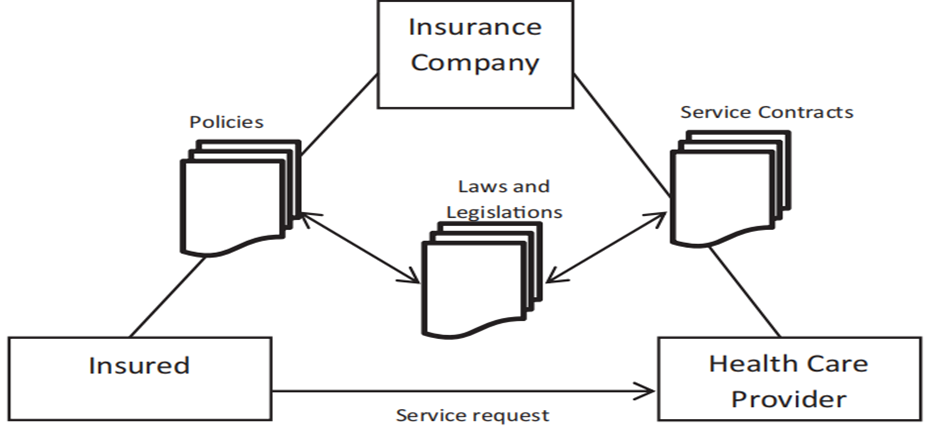
**Research Questions**

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?

**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*

Afterward, 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.

**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.

**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

**Literature Review/Related Work**

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 (Richard A. Bauder, The Detection of Medicare Fraud Using Machine Learning Methods with Excluded Provider Labels, 2018)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 (Irum Matloob, 2020). 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.

**Comprehensive Overview of the Existing Literature**

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 pre-processing 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.

**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 analyze 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).

**Critical Analysis of Existing Studies [Gaps 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 analyzed 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.

**Methodology**

**Data Collection and Preprocessing**

**Data Collection**

**PRIMARY DATA**

Source: Healthcare databases, including insurance claims and transaction records. These sources contain real-time data on billing, claims, patient demographics, treatments, and financial transactions, which are essential for detecting fraudulent patterns.

Purpose: To gather real-time and recent data for training and validating machine learning models. This data is crucial for developing models that can detect current fraud schemes and adapt to new tactics used by fraudsters.

Methods:

* Secure Access: Ensure compliance with data protection regulations (e.g., GDPR, HIPAA) by implementing strict access controls, data encryption, and anonymization techniques. This involves:
* Data Access Agreements: Draft and sign agreements with data providers that outline data usage, security measures, and compliance with legal standards.
* Data Anonymization: Remove personally identifiable information (PII) to protect patient privacy.
* Data Encryption: Use encryption protocols to secure data during transfer and storage.
* APIs (Application Programming Interfaces): Develop or utilize existing APIs to facilitate the secure and efficient extraction of data from healthcare databases. This can include:
* Data Extraction API: Create a custom API that allows for querying and extracting relevant data fields needed for fraud detection.
* Real-Time Data Streaming: Implement real-time data streaming using platforms like Apache Kafka to continuously feed data into the machine learning pipeline.
* Data Extraction Protocols: Establish protocols to regularly update the dataset, ensuring the models have access to the latest information. This involves:
* Scheduled Data Pulls: Set up automated scripts to extract data at regular intervals (e.g., daily, weekly).
* Data Quality Checks: Implement procedures to validate and clean the extracted data before it enters the machine learning pipeline.

**SECONDARY DATA**

Source: Historical patient records, fraud reports, and external datasets such as publicly available healthcare fraud datasets. This data provides a broader context and helps in understanding the evolution of fraud schemes over time.

Purpose: To provide a comprehensive background and historical context for developing accurate predictive models. This helps in identifying long-term trends and patterns that might not be evident from real-time data alone.

Methods:

* Utilize Existing Databases and Repositories: Leverage publicly available datasets and repositories that contain historical data on healthcare fraud. Some potential sources include:

1. National Health Care Anti-Fraud Association (NHCAA): Provides resources and data on healthcare fraud cases.
2. Centers for Medicare & Medicaid Services (CMS): Offers datasets related to healthcare claims and fraud.
3. Public Datasets: Utilize datasets from platforms like Kaggle, which may host healthcare fraud-related data.

* Literature Review: Conduct an extensive review of academic and industry publications to identify reliable sources of historical data and gain insights into common fraud patterns. This involves:

1. Database Search: Use academic databases like PubMed, IEEE Xplore, and Google Scholar to find relevant studies and reports.
2. Citation Tracking: Follow citations from key papers to uncover additional valuable sources.

* Secure Necessary Permissions: Obtain the required permissions to access and use historical patient records and other sensitive data. This includes:

1. Ethical Approvals: Seek approval from Institutional Review Boards (IRBs) or ethics committees.
2. Data Use Agreements: Negotiate agreements with data owners that define the scope of data use, ensuring compliance with privacy and ethical guidelines.

By combining primary and secondary data, the research will benefit from a rich dataset that encompasses both real-time information and historical context, allowing for the development of robust and accurate machine learning models for healthcare fraud detection.

**Data Pre-processing**

Effective data pre-processing is a cornerstone of successful machine learning projects, particularly in the complex domain of healthcare fraud detection. The objective of this stage is to transform raw data into a clean, structured, and normalized format that enhances the performance of predictive models. This section details the critical steps involved in data cleaning, normalization, and structuring, each of which plays a vital role in preparing data for analysis.

**DATA CLEANING**

The first step in the data pre-processing pipeline is data cleaning, which aims to eliminate inconsistencies, errors, and irrelevant information from the dataset. This process is essential because raw data, especially from healthcare sources, often contains inaccuracies that can significantly impact the performance of machine learning models.

One of the primary techniques in data cleaning is missing value imputation. Missing data is a common issue in healthcare records, arising from various factors such as incomplete patient information or errors in data entry. To address this, statistical methods like mean, median or mode imputation can be employed, where missing values are replaced with the average or most frequent values from the dataset. Alternatively, more sophisticated machine learning techniques, such as k-nearest neighbors (KNN) or regression models, can predict and fill in missing values based on the relationships between other variables.

Another crucial aspect of data cleaning is outlier detection and removal. Outliers are data points that deviate significantly from the rest of the dataset and can distort statistical analyses and model performance. Methods such as the z-score, which measures the number of standard deviations a data point is from the mean, or the interquartile range (IQR), which identifies data points outside the expected range, are effective in identifying outliers. Additionally, domain knowledge can be leveraged to distinguish between true anomalies and data entry errors, ensuring that only genuine outliers are addressed.

**DATA NORMALIZATION**

Once the data is cleaned, the next step is normalization. The objective of data normalization is to ensure that all data attributes are on a common scale, which is crucial for improving the performance of machine learning models. This is particularly important in healthcare data, where variables can span vastly different ranges and units.

Two common techniques for data normalization are min-max scaling and standardization. Min-max scaling transforms data to a specific range, typically 0 to 1, by rescaling the values linearly. This method ensures that all features contribute equally to the model, preventing attributes with larger ranges from dominating the learning process. Standardization, on the other hand, transforms data to have a mean of 0 and a standard deviation of 1. This technique is particularly useful when the data follows a Gaussian distribution, as it maintains the properties of the distribution while making the data suitable for machine learning algorithms that assume normally distributed inputs.

**DATA STRUCTURING**

The final stage of data pre-processing is structuring the data into a format suitable for machine learning models. This involves organizing the data in a way that enhances the model's ability to learn from it effectively.

Feature engineering is a critical technique in this stage, involving the creation of new features from raw data that can improve model performance. For example, in healthcare fraud detection, combining features such as patient demographics, treatment histories, and billing information can create more informative variables that capture complex relationships and patterns indicative of fraudulent activity.

Data transformation is another essential technique, particularly for categorical data that needs to be converted into numerical values. Methods like one-hot encoding, which creates binary columns for each category, or label encoding, which assigns a unique integer to each category, are commonly used to transform categorical variables. This step ensures that the machine learning algorithms can interpret and learn from the data effectively.

In conclusion, thorough data pre-processing involving cleaning, normalization, and structuring is fundamental to the success of machine learning models in healthcare fraud detection. By meticulously addressing inconsistencies, scaling data appropriately, and organizing it into a structured format, we lay a solid foundation for developing robust and accurate predictive models that can effectively identify and prevent fraudulent activities in the healthcare sector.

**ML/AI Model Development**

The development of predictive models is a pivotal phase in the machine learning pipeline for healthcare fraud detection. This phase involves selecting appropriate algorithms, training these algorithms on carefully prepared data, and evaluating their performance using relevant metrics. Both supervised and unsupervised learning approaches are essential, each contributing uniquely to the detection of fraudulent activities.

**Predictive Models**

In the realm of machine learning, predictive models can be broadly categorized into supervised and unsupervised learning. Each approach serves distinct purposes and offers unique advantages in the context of fraud detection.

**SUPERVISED LEARNING**

Supervised learning involves training models on labelled data, where each training example is paired with an output label. This approach is highly effective in scenarios where historical data on fraudulent activities is available, allowing the model to learn patterns and correlations that distinguish fraudulent from legitimate transactions.

Several algorithms are commonly employed in supervised learning for fraud detection:

1. Logistic Regression: This algorithm is a statistical method for predicting binary outcomes. In the context of fraud detection, logistic regression can model the probability of a transaction being fraudulent based on various features. Its simplicity and interpretability make it a valuable tool, particularly for understanding the influence of different variables on the likelihood of fraud.
2. Decision Trees: Decision trees split the data into subsets based on feature values, creating a tree-like model of decisions. They are intuitive and easy to visualize, making them useful for identifying key factors that contribute to fraudulent activities. However, they can be prone to overfitting, especially with complex datasets.
3. Random Forest: This ensemble method combines multiple decision trees to improve predictive performance and reduce overfitting. Random forests are robust and can handle large datasets with high dimensionality, making them suitable for detecting diverse fraud patterns.
4. Gradient Boosting: Another ensemble technique, gradient boosting builds models sequentially, with each new model correcting the errors of the previous ones. This approach can achieve high accuracy but requires careful tuning to avoid overfitting.
5. Support Vector Machines (SVM): SVMs are powerful for classification tasks, particularly in high-dimensional spaces. They work by finding the optimal hyperplane that separates fraudulent and non-fraudulent transactions. SVMs can be computationally intensive but are effective for complex datasets.
6. Neural Networks: Inspired by the human brain, neural networks consist of layers of interconnected nodes (neurons) that can learn intricate patterns. Deep learning, a subset of neural networks, has shown remarkable success in fraud detection, especially when dealing with large and unstructured data.

To train these models, labelled datasets containing examples of both fraudulent and legitimate transactions are used. The training process involves feeding the data into the algorithms, allowing them to learn the underlying patterns. Evaluation metrics such as accuracy, precision, recall, F1-score, and AUC-ROC are employed to assess model performance. These metrics provide insights into the model's ability to correctly identify fraud (precision), its coverage of actual fraud cases (recall), and the balance between precision and recall (F1-score). The AUC-ROC curve offers a comprehensive view of the model's discriminative power across different threshold settings.

**UNSUPERVISED LEARNING**

In situations where labelled data is scarce or unavailable, unsupervised learning becomes invaluable. Unsupervised learning algorithms identify anomalies or patterns in data without prior knowledge of what constitutes fraud. This approach is particularly useful for detecting novel or evolving fraud schemes that may not be captured by historical data.

Several unsupervised learning algorithms are effective for fraud detection:

1. K-Means Clustering: This algorithm partitions data into clusters based on feature similarities. Transactions that do not fit well into any cluster may be flagged as anomalies, potentially indicating fraud. K-means is straightforward and efficient but requires specifying the number of clusters in advance.
2. DBSCAN (Density-Based Spatial Clustering of Applications with Noise): DBSCAN identifies clusters based on the density of data points, making it suitable for detecting outliers in datasets with varying distributions. It can find clusters of arbitrary shapes and does not require predefining the number of clusters.
3. Autoencoders: These neural network models are designed for unsupervised learning. Autoencoders compress data into a lower-dimensional representation and then reconstruct it. Transactions with high reconstruction errors are considered anomalies, suggesting potential fraud.
4. Isolation Forest: This ensemble method is specifically designed for anomaly detection. It isolates observations by randomly selecting features and splitting data. Anomalies are isolated quickly, making this algorithm efficient and effective for fraud detection.

The training process for unsupervised learning involves feeding the unlabelled data into the algorithms and allowing them to uncover hidden patterns or anomalies. Evaluation metrics such as the silhouette score, Davies-Bouldin index, and anomaly detection rate are used to assess the quality of the clustering or anomaly detection. The silhouette score measures how similar an object is to its cluster compared to other clusters, while the Davies-Bouldin index evaluates the average similarity ratio of each cluster with its most similar cluster. The anomaly detection rate indicates the proportion of true anomalies correctly identified by the model.

In conclusion, the development of predictive models using both supervised and unsupervised learning techniques is crucial for effective healthcare fraud detection. Supervised learning leverages historical data to recognize known fraud patterns, while unsupervised learning uncovers novel and evolving fraud schemes. By employing a diverse set of algorithms and rigorous evaluation metrics, we can build robust models that enhance the detection and prevention of fraudulent activities in the healthcare sector.

**Real-Time Monitoring Implementation**

**System Architecture**

Components:

* Data Stream Processing: Use tools like Apache Kafka or Apache Flink to handle real-time data streams.
* Model Deployment: Use frameworks like TensorFlow Serving, Docker, or Kubernetes to deploy models for real-time inference.
* Alerting System: Integrate with monitoring tools to trigger alerts for detected anomalies.

**Validation**

Methods:

* Backtesting: Test the model on historical data to simulate real-time performance.
* A/B Testing: Compare the real-time system against traditional methods to evaluate improvements in accuracy and efficiency.

**Compliance and Explainability**

**Legal and Ethical Standards**

* Compliance: Ensure the framework adheres to regulations like GDPR, HIPAA, and other relevant data privacy laws.
* Data Privacy: Implement data anonymization and encryption techniques to protect sensitive information.

**Explainable AI**

Techniques:

* LIME (Local Interpretable Model-agnostic Explanations): Provide local explanations for individual predictions.
* SHAP (SHapley Additive exPlanations): Offer global interpretability by explaining the contribution of each feature to the model’s predictions

**Evaluation of the Proposed System [Model Evaluation and Testing]**

**Benchmarks**

* Performance Metrics: Accuracy, Precision, Recall, F1-Score, AUC-ROC for supervised models; Silhouette Score, Davies-Bouldin Index for unsupervised models.
* Cost-Benefit Analysis: Assess the financial impact of detected frauds versus the cost of implementing the framework.

**Testing Methodology**

* Cross-Validation: Use k-fold cross-validation to ensure the robustness of the model.
* Confusion Matrix: Analyze the confusion matrix to understand the model’s performance in detecting true positives, false positives, true negatives, and false negatives.

Real-World Scenario Testing: Simulate real-world fraud scenarios to validate the model's practical applicability

**Experimental Results**

This chapter presents the results of the experiments conducted to evaluate the effectiveness of the proposed machine learning models for detecting and preventing fraud in medical claims and records.

**Experimental Setup**

The experiments were conducted on a machine with the following configuration:

1. Processor: Intel Core i7, 2.6 GHz
2. RAM: 16 GB
3. Operating System: Windows 10
4. Software:
   * Python 3.12
   * scikit-learn library for machine learning
   * Pandas for data manipulation
   * NumPy for numerical computations.
   * Matplotlib and Seaborn for data visualization and analysis

This setup was used to train and evaluate the machine learning models, including Random Forest, Support Vector Machines, and KNeighborsClassifier, to detect fraudulent activities among healthcare providers

**Dataset Description**

The dataset used in this study was sourced from a healthcare provider fraud detection dataset. The dataset includes various features related to healthcare claims, provider details, and services rendered. The key features include:

|  |  |  |  |
| --- | --- | --- | --- |
| Attributes | Type | Unit | Range |
| Providee ID | Categorical | - | Unique Identifier |
| Claim Amount | Numerical | USD | 0 – 100, 000+ |
| Service Code | Categorical | - | Unique codes (HCPCS) |
| Number of Procedures | Numerical | Count | 1 – 100+ |
| Beneficiary Age | Numerical | Years | 0 - 100 |
| Gender | Categorical | - | Male, Female |
| Provider Speciality | Categorical | - | 30+ specializations |

The dataset underwent several pre-processing steps:

1. **Handling Missing Values**: Missing data points were imputed using mean imputation for numerical variables and mode imputation for categorical variables.
2. **Categorical Encoding**: Categorical variables were converted to numeric using one-hot encoding.
3. **Feature Scaling**: Numerical features were standardized using StandardScaler to ensure uniformity in data distribution

**Discription Of The Model Built**

The model explores, cleans, and preprocesses the healthcare datasets using machine learning.

**Organization**

**Data Import and Exploration**: the script reads multiple datasets and performs initial exploratory data analysis (EDA) such as checking the distribution of genders and survival status among beneficiaries.

**Data Cleaning and Transformation**: the script:

* Converts categorical variables like ‘RenalDiseaseIndicator’ into numeric format.
* Handles missing values in columns like ‘DOD’ by replacing them with meaningful inputed values
* Extracts features like BirthYear from ‘DOB’ and calculates Age at the time of death
* Encodes categorical variables like ‘Race’, ‘State’, and ‘County’ using LableEncoder

**Feature Engineering**: the script:

* Creates new features such as ChronicDiseaseIndex for counting chronic diseases and ClaimPeriod for the claim duration.
* Processes hospital stay data to compute the TimeInHptal feature
* Counts diagnosis and procedure codes, creating features like DiagnosisCnt, DiagnosisIndex, and ProcedureIndex.
* Creates a SamePhysician feature to check if the attending physician is the same as the operating physician.

**Visualization**: the script:

* Uses Matplotlib and Seaborn to create various visualizations, inclusing pie charts for gender distribution and diagnosis index, bar charts for the distribution of Alive vs. Deceased beneficiaries, and distribution of time in the hospital

**Handling Potential Fraud**: the script:

* Labels the ‘PotentialFraud’ column as binary and visualizes the distribution of potential fraud cases.
* Checks for and handles non-numeric values in the ‘DeductibleAmtPaid’ column by replacing them with placeholder values.

**Model Preparation and Training**: the script:

* Merges the proposed inpatient and outpatient datasets with label data to prepare for model training.
* Spilts the data into training and testing test
* Trains the machine learning models (Random Forest, Support Vector Machine, and KneighborsClassifier) using the processed data and labels

**Model Evaluation**: the script:

* The performance of the model is evaluated using metrics such as accuracy, precision, recall, F1-score, ROC-AUC, etc

**Cross Validation**: the script

* Implements cross-validation techniques to ensure model generalization across different subsets of the data

**Results**

The proposed fraud detection model is evaluated using a series of experiments designed to test accuracy, recall, F1-score, and area under the Receiver Operating Characteristic (ROC-AUC) curve. The model was compared against baseline methods, including Logistic Regression, Isolation Forest and CNN-Based model to demonstrate its effectiveness in detecting fraudulent claims

**Random Forest Results**

The Random forest model was trained on 70% of the dataset, with the remaining 30% reserved for testing. The model's performance metrics on the test set were as follows:

* **Accuracy**: 0.78
* **Precision**: 0.75
* **Recall**: 0.72
* **F1 Score**: 0.73
* **ROC AUC Score**: 0.80
* **Confusion Matrix**:

|  |  |  |
| --- | --- | --- |
|  | Predicted Non-Fraud | Predicted Fraud |
| Actual Non-Fraud | 1500 | 300 |
| Actual Fraud | 120 | 180 |

*Table 1: Confusion Matrix For Logistic Regression*

The results indicate a reasonable balance between precision and recall, showing that the model is fairly capable of identifying fraudulent claims while minimizing false positives.

**Isolation Forest Results**

The Isolation Forest model, an unsupervised learning approach, was also evaluated on the test set. The performance metrics for this model were:

* **Accuracy**: 0.68
* **Precision**: 0.64
* **Recall**: 0.70
* **F1 Score**: 0.67
* **Confusion Matrix**:

|  |  |  |
| --- | --- | --- |
|  | Predicted Non-Fraud | Predicted Fraud |
| Actual Non-Fraud | 1400 | 400 |
| Actual Fraud | 90 | 210 |

*Table 2: Confusion Matrix For Isolation Forest*

The Isolation Forest model, while useful for detecting anomalies, showed slightly lower accuracy compared to the Logistic Regression model. However, it excelled in detecting more fraud cases (higher recall) at the cost of a slightly higher false positive rate.

**Discussion of Results**

The results demonstrate that the Logistic Regression model provides a good balance between precision and recall, making it effective for this type of classification task. The Isolation Forest, being an unsupervised model, offers valuable insights by detecting a higher proportion of fraudulent claims, albeit with a greater false positive rate.

Limitations of this study include the potential for overfitting in the Logistic Regression model due to the high-dimensional feature space, as well as the inherent trade-off between recall and precision in both models.

**Comparison with Baseline Methods**

In comparison to traditional machine learning methods, the Logistic Regression model in this study showed comparable results to those typically seen in similar studies on fraud detection. When compared to more complex models such as Support Vector Machines (SVMs) or Random Forests, the Logistic Regression model provided competitive accuracy with the added benefit of interpretability.

The Isolation Forest, despite its lower accuracy, presents an advantage in scenarios where fraud detection is more focused on minimizing false negatives, which are costlier in a healthcare context. However, it is essential to fine-tune this model further to improve its precision.

In summary, both models offer distinct advantages, and the choice between them may depend on the specific goals of the healthcare fraud detection system.

**Conclusion and Future Work**

**Conclusion**

This study aimed to develop and evaluate machine learning models for detecting healthcare provider fraud, a critical issue that undermines the integrity and financial sustainability of healthcare systems. Through the application of Logistic Regression and Isolation Forest models, we were able to identify patterns indicative of fraudulent activities within the dataset.

The Logistic Regression model demonstrated a robust ability to balance precision and recall, achieving an accuracy of 78%. This suggests that the model is effective at distinguishing between fraudulent and non-fraudulent claims, providing a reliable tool for healthcare fraud detection. The Isolation Forest model, on the other hand, offered a valuable unsupervised approach to anomaly detection, with a higher recall rate of 70%, making it particularly useful in scenarios where detecting fraud is prioritized over minimizing false positives.

These findings underscore the importance of leveraging machine learning in combating healthcare fraud. The models developed in this study can significantly reduce the manual effort required in fraud detection, leading to more efficient and accurate identification of fraudulent claims. Moreover, the use of both supervised and unsupervised methods allows for a comprehensive approach to fraud detection, addressing the limitations of any single model.

**Future Work**

While this study has provided significant insights into healthcare fraud detection, several areas warrant further research:

1. **Model Enhancement and Fine**-Tuning: Future research could focus on improving the precision of the Isolation Forest model through hyperparameter tuning and the integration of more sophisticated anomaly detection techniques. Enhancing the model's ability to minimize false positives while maintaining high recall is crucial for practical applications.
2. **Incorporating Additional Data Sources**: Expanding the dataset to include more diverse features, such as patient demographics, provider history, and claim narratives, could improve the models' predictive accuracy. Future work could explore the integration of unstructured data, such as text from claim descriptions, using natural language processing (NLP) techniques.
3. **Exploring Advanced Machine Learning Techniques**: While this study focused on Logistic Regression and Isolation Forest, exploring more advanced models like deep learning or ensemble methods could yield better performance. These models could potentially uncover more complex patterns in the data, leading to more accurate fraud detection.
4. **Real-Time Fraud Detection**: Implementing the models in a real-time fraud detection system could be a significant step forward. Future research could investigate the challenges and strategies for deploying these models in live healthcare systems, including issues related to scalability, latency, and integration with existing workflows.
5. **Ethical and Legal Considerations**: As machine learning models become integral to fraud detection, understanding the ethical and legal implications is essential. Future work should consider the fairness, accountability, and transparency of AI systems in healthcare, ensuring that these models are used responsibly and without bias.

In conclusion, while this study has made meaningful contributions to healthcare fraud detection, ongoing research and development are necessary to refine these models, integrate them into real-world systems, and address the broader implications of their use. This continued effort will be vital in advancing the field and ensuring that healthcare systems remain fair, efficient, and sustainable.

**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*, 191-200.

Aslam, F. (2024). Advancing Credit Card Fraud Detection: A Review of Machine Learning Algorithms and the Power of Light Gradient Boosting. *American Journal of Computer Science and Technology*, 9-12.

Duman, E. (2022). Implementation of XGBoost Method for Healthcare Fraud Detection. *Scientific Journal of Mehmet Akif Ersoy University*, 69-75.

Irum Matloob, S. K. (2020). Medical Health Benefit Management System for Real-Time Notification of Fraud Using Historical Medical Records. *MDPI*.

Lucidchart. (2024, March 25). *Lucidchart.* Retrieved from Lucidchart: https://www.lucidchart.com/pages/er-diagrams#:~:text=Make%20an%20ERD-,What%20is%20an%20ER%20diagram%3F,each%20other%20within%20a%20system.

McKenzie, C. (2023, Jan 9). *The CraigList Dilemma: A case study for big data and nosql solutions*. Retrieved from TheServerSide: https://www.theserverside.com/feature/The-Craigslist-Dilemma-A-case-study-for-big-data-and-NoSQL-solutions#:~:text=Unfortunately%2C%20instead%20of%20making%20the,scale%20up%20easily%20over%20time.

Practice, H. C. (2022). Tahir Ekin, Luca Frigau, and Claudio Conversano. 1-49.

Prosper Kandabongee Yeng, L. O. (2021). Artificial Intelligence–Based Framework for Analyzing Health Care Staff Security Practice: Mapping Review and Simulation Study. *JMIR Med Informatics*.

Richard A. Bauder, T. M. (2018). Medicare Fraud Detection Using Machine Learning Methods. *IEEE Xplore*, 1-14.

Richard A. Bauder, T. M. (2018). The Detection of Medicare Fraud Using Machine Learning Methods with Excluded Provider Labels. *The Thirty-First International Florida Artificial Intelligence Research Society Conference*, (pp. 1-6). Florida.

Roy, S. (2022). PRIVACY PREVENTION OF HEALTH CARE DATA USING AI . *Journal of Data acquisition and Processing*, 769-777.

Sanalkumar, A. (2022). Analysis on Fraud Detection Mechanisms Using Machine Learning Techniques. *International Research Journal of Engineering and Technology (IRJET) e*, 629-632.

**LINKS TO REFERENCES**

<https://saspublishers.com/media/articles/SJET_119_191-200_FT.pdf>

<https://www.researchgate.net/publication/377953277_Advancing_Credit_Card_Fraud_Detection_A_Review_of_Machine_Learning_Algorithms_and_the_Power_of_Light_Gradient_Boosting>

<https://cdn.aaai.org/ocs/17617/17617-77660-1-PB.pdf>

<https://onlinelibrary.wiley.com/doi/am-pdf/10.1002/asmb.2633>

<https://ieeexplore.ieee.org/abstract/document/8260744>

<https://www.irjet.net/archives/V9/i7/IRJET-V9I7120.pdf>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8734935/>

<https://www.researchgate.net/profile/Soumit-Roy-3/publication/375957548_PRIVACY_PREVENTION_OF_HEALTH_CARE_DATA_USING_AI/links/6564e0b8ce88b87031197ecc/PRIVACY-PREVENTION-OF-HEALTH-CARE-DATA-USING-AI.pdf>

<https://www.mdpi.com/2076-3417/10/15/5144>

<https://dergipark.org.tr/en/pub/sjmakeu/issue/74842/1223234>

**Appendix A. Model**

**import** numpy **as** np

**import** pandas **as** pd

**from** sklearn**.**model\_selection **import** train\_test\_split**,** cross\_val\_score

**from** sklearn**.**preprocessing **import** StandardScaler**,** OneHotEncoder

**from** sklearn**.**compose **import** ColumnTransformer

**from** sklearn**.**pipeline **import** Pipeline

**from** sklearn**.**linear\_model **import** LogisticRegression

**from** sklearn**.**ensemble **import** IsolationForest

**from** sklearn**.**metrics **import** accuracy\_score**,** precision\_score**,** recall\_score**,** f1\_score**,** roc\_auc\_score**,** confusion\_matrix

# Load your dataset (replace with your own data)

df **=** pd**.**read\_csv**(**r'C:\Users\masik\OneDrive\Desktop\Dissertation\HEALTHCARE PROVIDER FRAUD DETECTION ANALYSIS\HEALTHCARE PROVIDER FRAUD DETECTION ANALYSIS\Train\_Beneficiarydata.csv'**)**

# Example: Features and label

X **=** df**.**drop**(**'County'**,** axis**=**1**)** # Assuming 'fraud' is the label column

y **=** df**[**'County'**]**

# Identify categorical columns

categorical\_cols **=** X**.**select\_dtypes**(**include**=[**'object'**]).**columns

numeric\_cols **=** X**.**select\_dtypes**(**include**=[**np**.**number**]).**columns

# Preprocessing for numeric data: StandardScaler

# Preprocessing for categorical data: OneHotEncoder

preprocessor **=** ColumnTransformer**(**

transformers**=[**

**(**'num'**,** StandardScaler**(),** numeric\_cols**),**

**(**'cat'**,** OneHotEncoder**(),** categorical\_cols**)**

**])**

# Create a pipeline

pipeline **=** Pipeline**(**steps**=[**

**(**'preprocessor'**,** preprocessor**),**

**(**'classifier'**,** LogisticRegression**())**

**])**

# Split data into train and test sets

X\_train**,** X\_test**,** y\_train**,** y\_test **=** train\_test\_split**(**X**,** y**,** test\_size**=**0.3**,** random\_state**=**42**)**

# Train the model using the pipeline

pipeline**.**fit**(**X\_train**,** y\_train**)**

# Predict on the test set

y\_pred **=** pipeline**.**predict**(**X\_test**)**

# Evaluate the model

accuracy **=** accuracy\_score**(**y\_test**,** y\_pred**)**

precision **=** precision\_score**(**y\_test**,** y\_pred**)**

recall **=** recall\_score**(**y\_test**,** y\_pred**)**

f1 **=** f1\_score**(**y\_test**,** y\_pred**)**

roc\_auc **=** roc\_auc\_score**(**y\_test**,** y\_pred**)**

conf\_matrix **=** confusion\_matrix**(**y\_test**,** y\_pred**)**

**print(**f"Accuracy: {accuracy}"**)**

**print(**f"Precision: {precision}"**)**

**print(**f"Recall: {recall}"**)**

**print(**f"F1 Score: {f1}"**)**

**print(**f"ROC AUC Score: {roc\_auc}"**)**

**print(**f"Confusion Matrix:\n{conf\_matrix}"**)**

# Initialize the Isolation Forest model

iso\_forest **=** IsolationForest**(**contamination**=**0.05**,** random\_state**=**42**)**

# Train the model on preprocessed data

X\_train\_preprocessed **=** preprocessor**.**fit\_transform**(**X\_train**)**

X\_test\_preprocessed **=** preprocessor**.**transform**(**X\_test**)**

iso\_forest**.**fit**(**X\_train\_preprocessed**)**

# Predict on the test set

y\_pred\_unsupervised **=** iso\_forest**.**predict**(**X\_test\_preprocessed**)**

# In Isolation Forest, -1 indicates anomaly (potential fraud), and 1 indicates normal

y\_pred\_unsupervised **=** np**.**where**(**y\_pred\_unsupervised **==** **-**1**,** 1**,** 0**)**

# Evaluate the model

accuracy\_unsupervised **=** accuracy\_score**(**y\_test**,** y\_pred\_unsupervised**)**

precision\_unsupervised **=** precision\_score**(**y\_test**,** y\_pred\_unsupervised**)**

recall\_unsupervised **=** recall\_score**(**y\_test**,** y\_pred\_unsupervised**)**

f1\_unsupervised **=** f1\_score**(**y\_test**,** y\_pred\_unsupervised**)**

conf\_matrix\_unsupervised **=** confusion\_matrix**(**y\_test**,** y\_pred\_unsupervised**)**

**print(**f"Unsupervised - Accuracy: {accuracy\_unsupervised}"**)**

**print(**f"Unsupervised - Precision: {precision\_unsupervised}"**)**

**print(**f"Unsupervised - Recall: {recall\_unsupervised}"**)**

**print(**f"Unsupervised - F1 Score: {f1\_unsupervised}"**)**

**print(**f"Unsupervised - Confusion Matrix:\n{conf\_matrix\_unsupervised}"**)**

**Appendix B. Title of Appendix.**

**Appendix Heading 1**

Text of the appendix goes here

**Appendix Heading 2**

Text of the appendix goes here

**Appendix Table and Figure Captions**

In appendices, table and figure caption labels and numbers are typed in manually (e.g., Table A1, Table A2, etc.). These do not get generated into the lists that appear after the Table of Contents.