A COMPREHENSIVE ANALYSIS OF BIG DATA ANALYTICS FRAMEWORKS FOR PROACTIVE FRAUD DETECTION IN THE HEALTH INSURANCE SECTOR

Dissertation Submitted to

The IIHMR University, Jaipur

for the partial fulfillment for the award of the degree of

Master of Business Administration (MBA)

in

Healthcare Analytics

by

Sourav Kundu

Under the Supervision and Guidance of

Dr. Kumar Parimal Shrestha

2026

IIHMR UNIVERSITY

Completion Certification

CERTIFICATE

This is to certify that sourav kundu in partial fulfilment of the requirements for the award of the degree of MBA (Healthcare Analytics) from the IIHMR University, Jaipur has successfully completed his internship at IIHMR University during **February 25-June 09, 2026.**

Place : Jaipur

Date : 23/06/2026

Faculty Guide

Dr. Kumar Parimal Shrestha

Assistant Professor

IIHMR University, Jaipur

IIHMR UNIVERSITY

CERTIFICATE

This is to certify that dissertation titled "**A Comprehensive Analysis of Big Data Analytics Frameworks for Proactive Fraud Detection in the Health Insurance Sector**" is a record of the research work undertaken by **Sourav Kundu** in partial fulfillment of the requirements for the award of the degree of **MBA (Healthcare Analytics)** from the IIHMR University, Jaipur under my guidance and supervision and his/her approach to the research study has been sincere, scientific, and analytical.

Faculty Guide

Dr. Kumar Parimal Shrestha
Assistant Professor
IIHMR University, Jaipur

Date:

DECLARATION BY STUDENT

I hereby declare that this dissertation titled "A Comprehensive Analysis of Big Data Analytics Frameworks for Proactive Fraud Detection in the Health Insurance Sector" is the bonafide record of my original research work. I declare that it has not been submitted to any other university or institution for the award of any degree or diploma. Information derived from the published or unpublished work of others has been duly acknowledged in the text.

Sourav

(Signature)

Name of Student: Sourav Kundu

Date:

ACKNOWLEDGEMENT

I express my sincere gratitude to IIHMR University, Jaipur, for providing me the opportunity to undertake this research and for the academic environment that made this work possible.

I am deeply indebted to my Faculty Guide, Dr. Kumar Parimal Shrestha, Assistant Professor, IIHMR University, Jaipur, for his invaluable guidance, constant encouragement, and unwavering support throughout this dissertation journey. His expertise, constructive feedback, and patience have been instrumental in shaping this research work.

I extend my heartfelt thanks to the faculty members and staff of the Department of Healthcare Analytics for their academic inputs and administrative support during the internship and dissertation period.

I am grateful to my family for their unconditional love, moral support, and encouragement throughout my academic journey. Their motivation has been a constant source of strength during this challenging yet rewarding experience.

I also acknowledge the contribution of the open-source research community and the CMS Kaggle repository for making the healthcare provider fraud detection dataset publicly available, which formed the empirical backbone of this study.

Finally, I thank all my colleagues and friends at IIHMR University for their support, constructive discussions, and camaraderie during the programme.

Sourav Kundu

MBA (Healthcare Analytics), Batch 2024–26

IIHMR University, Jaipur

2026

TITLE: A COMPREHENSIVE ANALYSIS OF BIG DATA ANALYTICS FRAMEWORKS FOR PROACTIVE FRAUD DETECTION IN THE HEALTH INSURANCE SECTOR

guided by: - DR. Kumar Parimal Shre

Submitted By: Sourav

Specialization: Healthcare Analytics

1. Introduction

Technological developments that have been embraced at a very fast pace in the healthcare sector have revolutionized the process of providing medical services managing patients and administrative procedures worldwide. Electronic Health Records telemedicine wearable technology and integrated claims management have resulted in the creation of massive amounts of digital data. Although these changes have positively impacted access efficiency and the quality of services offered they have posed additional threats especially related to health insurance claims processing. The increased number of electronic claims has opened up opportunities for fraudulent activity which can include identity theft upcoding phantom billing unbundling of services and collaboration between patients and providers and even insurance companies. The crime of healthcare fraud is certainly not something new; however the magnitude and sophistication of such a crime have increased tremendously due to technology's integration into the field. As per estimates of the Coalition Against Insurance Fraud insurance fraud in all lines of business costs the United States economy around $308.6 billion per year. Among these health insurance fraud costs tens of billions of dollars each year and conservative estimates peg the figure at $36.3 billion for private health insurance while the loss due to Medicare and Medicaid is quite large as well ($68.7 billion in some estimates).

Historical Background and Development of Healthcare Fraud

In order to comprehend the problem that is currently occurring it is necessary to analyze the development of healthcare fraud historically. Prior to digital times frauds could only occur via manually filling incorrect information on the paper forms such as overstating services performed or even filing fake patients' claims. The change in the healthcare system through the use of electronic systems from the late 20th century to early 21st century facilitated by acts such as the U.S. HITECH Act resulted in huge volumes of claims being filed each year. Healthcare identity theft entails the unauthorized use of personal health information in filing fraudulent claims. Criminals take advantage of the many data breaches that occur to get their hands on patients' ID numbers provider information or insurance information. Afterward this information is used to make claims for services that were never delivered. Phantom billing involves making charges for services that were never offered. The practice of upcoding is charging for more complicated or costly services than the ones actually performed; this may include classifying a simple checkup visit as an intricate test visit. Another scheme is known as unbundling which involves dividing a single service into several parts in order to

be able to bill for them separately. The following real-world examples show the seriousness of this problem. One of the examples is from a major DOJ crackdown on cases where the defendant was involved in conspiracies worth over $14.6 billion in estimated loss such as multibillion-dollar fraud rings within the Medicare system. One of the cases involved a New York man who received a jail term for his conspiracy which cost $336 million.

The Importance of Big Data in Today's Healthcare System

Today's healthcare system creates huge volumes of data on a regular basis ranging from clinical notes and lab results to billing codes and patient demographic information. Big data refers to data in its structured form (such as claims database) and in unstructured form (such as doctors' notes and images). With respect to the size velocity and complexity of this kind of data it becomes impossible to use manual methods. Manual approaches which involve audits and random samples of claims can only detect partial instances of fraudulent activities. Financial leakage due to undetected fraud is far-reaching. Insurers offset their losses by increasing premiums deductibles and benefits reduction. This increases cost-sharing for the policyholders who might experience delayed treatment poor health conditions and bankruptcy in the case of high medical bills. Group insurers have increased costs as well resulting in loss of benefit or increased employee cost-sharing. Public programs insured by the government such as Medicare suffer from budget constraints thus shifting resources away from other important uses.

Some estimates show that fraud increases the cost of insurance premiums by about 20% or more with families paying hundreds or thousands of dollars per year in indirect losses. Besides the monetary losses fraud creates a lack of trust in the healthcare system puts patients at risk of receiving inappropriate treatments and affects patients' medical records thus causing misdiagnosis or withholding of legitimate treatment.

Fraud Detection Systems’ Limitations

The major weakness of most fraud detection systems currently in use within the health insurance industry is that they are reactive and are based on rules. These rules could involve anything that would indicate something out of the ordinary like the claims exceeding certain limits or particular codes used to justify billing. Though effective when it comes to catching the obvious fraud such systems face a number of challenges; one of them being the high rate of false positives and frequent need for updating rules. The reactive approach does not take into account patterns emerging at any particular time. In view of the globalized nature of fraud

where there is fraud conducted by international gangs using stolen information static rules cannot cope. Another issue is the huge volume of claims in terms of billions.

The Potential of Artificial Intelligence (AI) Machine Learning (ML) and Big Data Analytics

Artificial intelligence (AI) and machine learning (ML) provide promising capabilities through the use of secondary historical data (past claims provider profile patient behavior) to uncover anomalies and determine the probability of fraud. Such methodologies as supervised learning (logistic regression random forest neural networks for classification purposes) unsupervised learning (cluster analysis for outlier detection) and anomaly detection algorithms are capable of handling extremely large datasets that surpass human capabilities. Claims analysis software combines information from claims with that from external resources (such as public databases social media activity or networks of providers) through technologies such as Hadoop Spark or cloud computing. NLP technology may be used to analyze notes and detect any inconsistencies. Graph analytics can be used to reveal collusion networks through relationship mapping. The effectiveness of case studies can be proven. Companies using ML technology have achieved considerable decreases in fraud payouts like a dental insurer who identified multiple fraud patterns. Models that use both machine learning and business rules have improved the results in detection. The market of fraud detection in healthcare will increase rapidly reaching a level of more than $3.6 billion in 2025 up to much higher levels in 2034. However some challenges are still there that include data privacy (HIPAA/GDPR compliance) interpretability of machine learning models (black box issue) training data bias and the adversarial nature of fraud. Issues regarding ethical concerns such as avoiding over flagging of claims from underserved populations need consideration.

In this paper we will address those challenges and explore these issues through AI/ML frameworks to process historic data architectures for real-time detection implementation challenges and the way forward. Shifting from reactive to proactive detection the healthcare ecosystem will save from financial leakage lower premiums build trust and ensure resources to actual patients.

2. Problem Statement

The amount of financial stress in the health insurance industry has become unparalleled owing to the rise in frauds in this sector. Elaborate fraud mechanisms like fictitious billing up-coding medical identity theft and patient-provider conspiracy steal money in an alarming manner.

Current approaches for fraud detection are mainly rules-based and reactive approaches that identify the anomaly in the activity post-payment and hence cause irrevocable damage. This inefficient mechanism causes insurers to hike premiums which creates a burden on the policyholders as well as makes healthcare expensive. However even when there is a large amount of data generated through digitization certain research gaps remain.

Fraud Types and Extent

Provide a thorough explanation of each fraud type with examples figures and processes involved. Phantom billing: Billing for services not provided by using false or stolen identities. Upcoding: Exaggeration of services performed for example billing level 5 visits as level 3. Collusion: Kickback arrangements for unneeded referrals. Variations globally and regionally (US India Europe). Provide references to DOJ raids NHCAA estimates (3-10%) and increasing trends using technology.

Implications on Economy and Society

Impact Assessment: Premium increases ($400-$700 for families per year due to fraud) tax increases benefit reductions layoffs in the insurance industry delayed care and negative consequences thereof deterioration of the system. Quantified using models of the 5-10% leakages impact on growth. Implications on vulnerable communities rural populations and developing nations.

Efficiencies within Legacy Systems

Critique of rule-based systems: False positives delayed adaptation payment-focused approach lack of scalability. Contrast with requirements in an age of big data. Consider compliance challenges and backlog investigations.

Gaps & Opportunities for Research

Lack of framework integration of artificial intelligence technologies for secondary data analysis in real time. Difficulties with class imbalance feature creation cross-border data collaboration and explainable artificial intelligence. The necessity of the implementation of hybrid human-AI technology privacy protection and benchmarking.

(Discuss more literature gaps possible approaches interview synthesis if any policy recommendations quantitative projection of savings from adoption of AI failures caused by current systems etc. in order to reach approximately 5000 words total.)

This would be a solid and academic foundation. If you require the entire uninterrupted text specific sections references in a certain format (APA etc.) tables or editing of word count/wording/other please let me know!

3. Review of Literature

1. Leevy et al. (2024) Leevy Salekshahrezaee and Khoshgoftaar (2024) – Leevy Salekshahrezaee and Khoshgoftaar (2024) carried out an extensive systematic review on A Review of Unsupervised Anomaly Detection Techniques for Health Insurance Fraud. This is one major piece of research work in the burgeoning field of health insurance fraud detection especially by focusing on unsupervised techniques which overcome the problem of lack of sufficient data due to class imbalance issue.The key point made by the authors is the rapid expansion of health care claims data due to EHRs digital billing systems and networks of providers which makes the conventional manual inspection and rule-based methods insufficient. In most cases fraud amounts to 1-5% of all claims and thus the use of supervised machine learning techniques faces many problems including class imbalance and very few fraud labels due to their expensive and cumbersome acquisition process via post-payment audit.

Core Contributions and Methods Covered

Leevy et al. conducted an exhaustive literature review from 2017 to 2024 on research articles that utilized various unsupervised machine learning algorithms tailored towards the detection of fraud within the health insurance sector. These are divided into five main categories of clustering (K-means and DBSCAN) density-based methods (Local Outlier Factor-LOF) isolation-based methods (Isolation Forest) reconstruction-based (Autoencoders Variational Autoencoders) and ensemble models. The criteria for each method include scalability with respect to high-dimensional claims data ability to handle mixed numerical and categorical variables (Provider ID Diagnosis Code Procedure Codes Reimbursement Amounts) resilience to noise and interpretability for regulatory purposes. The main advantage of the paper is its focus on practical usability in healthcare environments. For example, they mention that Isolation Forest works well in high dimensional space due to random partitioning of data, which makes it very efficient for working with big sets of data like Medicare or even private insurers with up to several million claims. The autoencoders are mentioned because they learn compact representation of normal claims' behavior and can recognize high reconstruction error which could be fraud (unusual billing, fake claims).

Analysis of Challenges Addressed

The authors carefully describe domain-specific challenges such as high cardinality in categorical features (thousands of ICD-10 codes or provider specialties), temporal dependencies in claims series, and the adversarial aspect of fraud detection (the fraudsters adjust their tactics to detection methods). Leevy et al. test the approaches on open/public and semi-synthetic datasets based on CMS Medicare datasets, using evaluation measures such as precision@k, AUC-ROC (accounting for imbalanced classes), and domain-specific cost-sensitive performance metrics (monetary gains vs. investigation expenses).

Specific attention is paid to explainability of the models, which becomes an issue as black-box unsupervised solutions do not appeal to the insurance industry and regulatory agencies. Potential solutions include SHAP integration and prototype explanations, allowing for actionable explanation like "this particular provider's claim velocities and code combination deviate 3.2 sigma from peer standards.

Wider Context and Identified Gaps

In the broader academic context, Leevy et al. set their study apart from previous literature on survey (such as Bauder et al., 2016) but emphasize the increased interest in unsupervised techniques post-2020, fueled by the development of deep learning and frameworks for big data processing (such as Spark). The authors contrast the performance in different studies, finding that the application of multiple approaches usually results in better performance, producing an improvement in fraud detection by 5-20x in suspicious claims.

The review highlights the necessity to go from the stage of post-payment detection to pre-payment flags, corresponding to the thesis statement of this dissertation about the proactive use of AI/ML systems. The empirical cases mentioned by the authors involve the detection of upcoding (billing high-level services), unbundling and collusion networks through graph-based anomalies.

2. Narne (2024) Narne (2024) – In the article “Machine Learning for Health Insurance Fraud Detection: Techniques, Insights, and Implementation Strategies”, Narne (2024) provides a comprehensive practical overview on the applications of machine learning technology that are particularly aimed at combating health insurance frauds. Besides focusing on algorithmic techniques, the article highlights the issues of implementation, practical business insights, etc.

Overview and Key Insights

The author highlights the huge amount of money that leaks out because of fraud (about 3-10% of the cost of total healthcare expenses in the world) that inflates

prices and affects profits. This paper presents the overview of the various machine learning techniques (supervised learning such as Random Forest, XGBoost, Neural Networks; unsupervised learning as cross-referenced with Leevy et al.; hybrid algorithms).

Special focus is made on feature engineering using claims data, which include provider behavior features, temporal features, patient behavior features and network analysis for detecting collusion.

Use cases are an important element of the paper where implementation is presented using cloud services (AWS, Azure), for example, using H2O.ai framework or custom-made Spark pipelines.

Technical Depth & Evaluations

In-depth technical sections include the treatment of imbalanced classes (SMOTE, cost-sensitive learning), evaluating classifiers under imbalanced conditions (PR-AUC, economic costs), and model interpretability (SHAP, LIME). The author offers tips on implementation such as pipelines, real-time score engines, and concept drift detection. Information on regulatory compliance, explainability to auditors, and calculating ROI adds great practical value to this paper.

3. Hamid et al. (2024) Hamid et al. (2024) – Hamid et al. (2024) in Healthcare Insurance Fraud Detection Using Data Mining, BMC Medical Informatics and Decision Making, provide a dedicated methodological approach based on association rules mining in combination with unsupervised learning for fraud detection purposes.

Contributions to Methodology

The suggested approach consists of applying such techniques as Apriori or FP-Growth in order to discover frequent item sets (such as abnormal associations between diagnosis and procedures codes) and combining them with clustering or outlier detection in order to recognize anomalies. Experimental validation on real or benchmark datasets shows good performance in detecting phantom billing and upcoding.

Analysis and Strengths

The paper combines data mining with ML, focusing on scalability and interpretability of the rules. The paper considers the sustainability of insurers as well as providing quantitative performance measures and qualitative results.

Some limitations include rule explosion in large data sets and the necessity to validate the rules by experts.

4. Rahman (2025) Rahman (2025) – Rahman (2025), in his landmark book entitled "Data-Driven Graph Neural Network Models for Detecting Fraudulent Insurance Claims in Healthcare Systems," presents extremely sophisticated graph-based techniques that mark a tremendous breakthrough in proactively detecting fraud. In this research paper, the author employs Graph Neural Networks (GNNs) in order to map the intricate relational data within the health care system to detect any conspiracy among the actors involved.

Theoretical Foundation and Graph Modeling

Conventional ML-based tabular models consider claims independent from each other and overlook the inherent interconnectivity in fraud. Rahman (2025) forms HINs where nodes are various entities such as patients, providers, procedures, and diagnoses, and edges show relationships between nodes. Graph model is vital for capturing the multi-hop dependencies needed for collusion detection like kickbacks and organized phantom billing operations.

The proposed method uses cutting-edge architectures such as HINormers, HybridGNNs, and graphSAGE models. The graph-based architecture propagates information throughout the graph by leveraging attention and message passing layers to learn node representations. The learned representation encodes structural anomalies on both local and global levels. As an example, if a node (provider) has abnormal connectivity to patients or procedures, it is considered to be suspicious.

Empirical Methodology and Findings

The experiments were performed on large-scale real-life and synthetic datasets based on Medicare claims and databases of private insurance companies, involving millions of data samples. The preprocessing phase includes graph construction (with the help of software libraries such as NetworkX or PyG – PyTorch Geometric), feature engineering (node features: claim volume, reimbursement behavior, specialization codes; edge weights: temporal frequency, monetary value) and semi-supervised or self-supervised training because of limited labels.

The main results reveal better effectiveness compared to the baseline GNNs and conventional models. The presented approaches showed significant improvements in terms of AUC-ROC (0.92+), precision@top-K and economic metrics (recovery of lost funds due to fraud). The ablation studies prove the importance of heterogeneity consideration and spatio-temporal constrains. Graph

visualizations comprise embeddings obtained by means of t-SNE method that indicate fraud clusters and attention heat maps.

Critical Analysis and Discipline-Specific Innovations

Rahman tackles problems related to the discipline: high cardinality of the categorical features (ICD-10/CPT codes), temporal nature (claims develop over time), and privacy concerns (handled by using tools such as differential privacy or federated training of GNNs). What makes the paper particularly good at is finding collaboration behaviors undetectable by traditional approaches, for example, circular referrals or coordinated upcoding by affiliated providers.

The issues identified by the authors include scalability concerns for massive graphs (managed by sampling and distributed training), and interpretability (improved by using GNNExplainer or attention mechanisms). Ethics includes bias in the graph structure that may target particular groups.

5. Carneiro (2024)

Carneiro (2024/2025) discusses Advanced Anomaly Detection Models for Revealing Frauds and Abnormal Billing Patterns of Providers in Health Insurance. The research will concentrate on identifying the roots of billing frauds and develop highly specialized models for anomalies at the provider level based on unsupervised and semi-supervised anomaly detection algorithms that will be tailored to detecting phantom billing and upcoding.

Main Concepts and Red Flags

Conceptual frameworks will be developed based on "unusual behavior" patterns at the health care institutions, such as excessive amounts of procedures in comparison with peers, inconsistent Length of Stay (LOS) patterns, or abnormal reimbursement distributions compared with the clinical norms. Models include Isolation Forest, Autoencoders, LOF, and ensembles of these models that will work on provider-level aggregates created based on claims histories.

Phantom billing will be identified as the lack of evidence supporting the claims, and upcoding as the abnormal changes in the distributions of Diagnosis Related Groups (DRGs) and procedures' hierarchy codes. The Hospitalization Period Curve and Detection of Financial Outliers Graphs in the dissertation were inspired by the definitions of these red flags.

Methodology and Evaluation

Carneiro uses multi-view anomaly detection techniques that combine statistics, machine learning, and domain-specific rules. An analysis based on datasets like

Medicare Provider Utilization shows a very high detection rate of frauds among providers, along with easily interpretable results (SHAP values of features like claim velocity or code entropy). Metrics with cost-sensitivity considerations of high-impact cases are used to assess performance.

Advantages, Limitations, and Applications

The advantages of this study are in granularity of the provider-oriented approach, allowing for conducting relevant audits and pre-payment actions. The research helps to address financial sustainability issues by assessing the potential gains and minimizing false positives due to hybrid modeling. Limitations can be seen in reliance on the quality of peer grouping and dynamic environment issues.

6. Nimbalkar (2024)

Nimbalkar (2024) examines Machine Learning Approaches for Time-Series Anomaly Detection in Health Insurance Claims Fraud. This research moves from analyzing static snapshots to dynamic analysis based on time series, looking at the claims as dynamic processes and identifying the fraudulent behaviors developing through time.

Time-Series Approach

Methods used in research are LSTM/GRU autoencoders, Prophet with anomaly scores, Temporal Convolutional Networks, and isolation approaches applied to time series. Identified variables include sharp increases in the number of bills, unusual patterns in hospital stay lengths, or anomalies in seasonal patterns for particular procedures. All these methods fit the proposed approach in the dissertation well – Hospitalization Period Curve. Empirical Contributions

The use of sliding window techniques along with feature extraction (rolling statistics, Fourier transforms) enables the models to detect temporal anomalies such as upcoding campaigns and fake billing spikes. The validation of these models on longitudinal data proves that there is better detection when compared to other non-temporal models.

Analysis and Synergies with Dissertations

The advantages of dynamic monitoring are emphasized in this study, particularly in relation to the limitations of static monitoring. Various limitations and ways to overcome those limitations are also covered in this paper. Expansion of this part includes mathematical equations (e.g., reconstruction error), experiments, GNNs (Graph Neural Networks) for the use of spatio-temporal graphs, and literature review—approximately 2500 words.

7. Özdemir (2024)

The article Innovative Machine Learning Approaches for Insurance Fraud Detection in the Health Sector: A Comparative Analysis of Analytical Models by Özdemir (2024) is an informative and relevant addition to the existing body of knowledge about fraud detection in the healthcare sector. The importance of the research lies in the development of effective ML solutions that will reduce financial losses and decrease dependence on costly investigations. In this paper, the author provides comparative analysis of several analytical models and, thus, allows making the right choice for insurance managers and decision-makers.

Introduction to the Background and Problem Definition

The author starts off by emphasizing how expensive the problem of healthcare fraud is, leading to an increase in premiums and putting insurance systems worldwide at risk of unsustainability. Rule-based approaches and post-payment audits are criticized for being inefficient, costly, and unscalable due to the explosion in volume of claims data. Özdemir (2024) states that it is necessary to use proactive, data-driven machine learning techniques to move away from post-mortem recoveries towards prevention of frauds; this approach directly corresponds to the central argument of the dissertation.

Methodological Approach and Comparative Analysis

The research compares a wide array of machine learning methods, such as supervised classifiers (Logistic Regression, Random Forest, XGBoost, LightGBM), unsupervised outlier detectors (Isolation Forest, Autoencoders), and ensembles. Experimentation relies on large-scale data samples of millions of anonymized insurance claim records, with the features designed according to provider information, demographics of the patients, billing history, ICD-10/CPT code of diagnosis and procedures, amount of reimbursement, and other time-related factors.

Main contributions of the study include cost-sensitive learning for prioritizing significant frauds and sophisticated sampling methods (variations of SMOTE, ADASYN) to compensate for class imbalance. The performance is measured by not only traditional metrics (AUC-ROC, F1-score, Precision-Recall) but also special economic ones, such as fraud losses prevented per cost of investigation and potential ROI. Using comparative tables and significance tests (McNemar's, Friedman's) it has been shown that ensemble methods, in particular XGBoost with unsupervised filtering, perform better than any individual models in terms of precision-recall of the minority classes.

The paper contains extensive experiments on ablation, hyperparameter tuning using Bayesian optimization, and techniques of cross-validation (stratified k-fold). Visual representations include ROC, confusion matrices, and feature importance (the same concept as algorithm feature importance profile in the dissertation).

8. Reddy & Rexie (2025)

Reddy & Rexie (2025) present their vision in the prospective article entitled "Hybrid Machine Learning Models for Real-Time Detection of Medical Insurance Claim Frauds". The new approach presented by Reddy & Rexie (2025) provides a promising hybrid architecture that is needed to fill gaps in existing literature on the issue of real-time detection of medical insurance claim frauds. The framework allows receiving highly precise probabilities for each claim without false positive alerts.

Innovative Core – Hybrid Framework

Combination of several algorithms is used in the suggested framework, which comprises the following stages: First, unsupervised algorithms such as isolation forest or autoencoder are used for detecting outliers among the claims. Then, supervised algorithms based on gradient boosting (XGBoost/CatBoost) as well as deep learning approaches (TabNet/neural nets) are applied for receiving probabilities for claims. A meta-learner called stacking ensemble integrates results of these models, giving the final fraud risk score. Features engineering is quite comprehensive including static claim characteristics, provider's background information, patient's behavioral characteristics and dynamic features such as claim velocity and timelines. The system utilizes real-time processing of historical secondary data through use of streaming technologies such as Apache Kafka and Spark Streaming.

Empirical Validations & Performance

Experimental studies based on benchmarks and proprietary datasets (Medicare-type claims) show better performance: AUC-ROC greater than 0.94, increase in precision@top-1000 by 30-45% compared to single model baselines, as well as decrease in false positive errors (very important for feasibility). Latency time remains below 200 ms per claim via optimized inference pipelines. The economic assessment suggests substantial preventable leakage, leading to high ROI from deployment experiments.

Some of the visual outputs are: risk dashboards, SHAP force plots for explaining each claim, and threshold optimization curves – all pertinent to the visual approach of the dissertation.

Overcoming Literature Gaps and Being Proactive

In contrast to previous systems' reactive approach, Reddy & Rexie promote a proactive stance through pre-payment intervention. The paper is effective in addressing literature gaps in hybrid real-time systems by striking the right balance between speed, accuracy, and explainability, thereby rendering it ideal for incorporation into live insurance claims processing engines. Methods used in dealing with concept drift and imbalanced datasets are elaborately discussed.

Critical Review

Among the strengths of the paper is the emphasis on deployment, evaluation (technical and business), and reduction of alert fatigue. The limitations of the paper identified include dependency on data, computational needs for deep algorithms, and necessity for human oversight in critical decisions.

9. Oyebode Kayode et al. (2025) Oyebode Kayode et al. (2025) - Yebode Kayode et al. (2025), through their paper titled "Artificial Intelligence and Big Data Architectures for Healthcare Billing Fraud Detection: A Unified Framework," offer a relevant and holistic analysis on the ways in which artificial intelligence together with big data can tackle the increasing problem of fraud within health care billing systems. In the paper, it is stressed that the increasing digitization of the health care sector, marked by the extensive use of EHRs and telemedicine systems and automation in claims processing, is leading to an exponential growth in the number of claims.

Background and Justification

The authors begin their discussion on healthcare fraud by providing background information about the increase in healthcare fraud with regard to the transformation in technology. As the amount of claims data is increasing in size, speed, and diversity, conventional approaches, which are either manual or rules based, are unable to cope. Oyebode Kayode et al. believe that big data analytics, through its application of artificial intelligence, can provide the scale and intelligence needed to analyze historical secondary data ahead of time.

Core Contributions: A Unified Framework for Big Data Architecture

In this paper, a unified framework for big data architecture is proposed, including the following components:

Layer for Data Ingestion: Using tools like Apache Kafka and Flume for real-time ingestion of claims data from multiple sources (hospitals, pharmacies, insurance companies).

Layer for Storage and Processing: Hadoop Distributed File System (HDFS), Apache Spark, and cloud-based data lakes (e.g., AWS S3 and Athena) for storing both structured (claims tables) and unstructured data (notes from clinicians, images).

Layer for Analytics and Artificial Intelligence: Combining machine learning pipelines with TensorFlow, PyTorch, Spark MLlib, focusing on scalable algorithms (distributed Random Forests, deep neural networks, and unsupervised anomaly detection).

Layer for Visualization and Decision-Making: Visual dashboards with Tableau or Power BI with visualization features similar to those of the Imbalance of Classes Operation Diagram and Financial Outliers Graph used in this dissertation.

Methodology and Experimentation

Oyebode Kayode et al. performed experiments with large datasets consisting of artificial extensions of Medicare Provider Utilization and Payment Data and other private insurance databases over multiple years. Methodology includes the following pipeline procedures: data preprocessing with PySpark, feature selection with PCA or autoencoders, cross-validation modeling, and deployment of models with microservices (Kubernetes).

The main results are summarized as follows: the proposed integrated approach outperforms isolated systems: fraud detection recall increased by 35-50%, while processing speed became fast enough to perform scoring for thousands of transactions per minute in real-time mode. Economic estimations show annual savings at the level of hundreds of millions of dollars for midsize and larger insurance companies due to decreased leakage from fraudulent activity.

Strengths of the Study and the Problems Solved by It

Strengths of the study include the holistic approach taken to connect data engineering and AI, rather than the algorithms alone. The authors discuss the problems of data heterogeneity, scalability, model explainability (SHAP and LIME techniques are used), and ethical problems such as algorithmic bias for various populations. Limitations include the significant initial expenses on infrastructure and the expertise needed from data scientists; the researchers suggest gradual implementation.

The study gives the scientific community a comprehensive insight into the elements of big data (storage, processing, analytics, and governance), adapted specifically for fraud detection, solving a gap in literature where these elements are discussed separately.

10. Fourkiotis & Tsadiras (2025)-

Fourkiotis & Tsadiras (2025) - examines future internet usage in the context of healthcare by looking at the use of big data technology in machine learning-enabled fraud detection. This research examines how technologies that include machine learning and artificial intelligence can be used for the analysis of secondary data in the near future.

11. Rehman et al. (2025) Rehman et al. (2025) - propose a comprehensive AI-driven framework for need-based insurance plans generation and anomaly detection using deep learning techniques. This study pushes beyond basic statistical methods by employing deep neural networks for advanced claim anomaly detection. Furthermore it addresses significant technical challenges by proposing frameworks that not only detect financial fraud but also dynamically generate optimized insurance plans offering a multi-faceted approach to healthcare insurance management.

Research Objectives

1.1 Structural Fragmentation in Secondary Healthcare Datasets

The first part of this study deals with the solution of the fundamental problem faced in the field of healthcare informatics regarding structural fragmentation high-dimensional and syntactically heterogeneous structure of secondary healthcare administrative datasets. Medical databases freely available online—in particular healthcare claims synthetic datasets made available via Kaggle (based on the SynPUF/CMS data)—are structurally fragmented into separate relational tables which consist of Inpatient Claims Outpatient Claims and Beneficiary Demographics.2. To construct robust behavioral and risk matrices at a provider level through systemic analysis of variations deductibles' distribution and the patient’s time spent in the hospitals (Admit Days).

Each database represents a completely distinct feature of the healthcare delivery system:

The Beneficiary Database -is a static repository of demographic information on patients which includes socio-economic status presence of chronic illnesses (such as Alzheimer's disease diabetes ischemic heart disease) geographic codes and mortality rates over time.

The Inpatient Claims Database -contains data on episodes of care that require hospitalization. These episodes are marked by their high cost complex Diagnosis Related Groups (DRGs) ICD-9 and ICD-10 diagnoses and procedures as well as clear admission-discharge dates.

The Outpatient Claims Database -stores data on high-volume low-cost episodes of care including imaging tests lab tests and routine ambulatory visits which are dominated by HCPCS. The key challenge here is that the datasets have an alignment issue at their core. These datasets have various levels of granularity different timeframes lack of relationship constraints and highly unstructured categorical attributes. Thus there is no possibility of ingestion. This research goal sets out to create a scalable and programmatic data pipeline that will ingest the datasets with multiple relational tables perform schema validation do the required data cleansing and conduct relational join operations to create a high-quality computational dataframe.

1.2 Multi-Stage Data Loading and Schema Normalization

The data ingestion infrastructure is designed to use high-throughput computational engines (e.g. PySpark in distributed clusters or tuned Python Pandas workflows with Arrow backends) due to the high memory requirements of the medical administrative data. The process starts by applying a rigorous immutable schema blueprint on the source files to avoid type coercing issues that might arise during the downstream math calculations. It is not uncommon to encounter healthcare data where data types have become corrupted (e.g. the inclusion of alphabetic characters in numeric billing columns or zip codes being automatically converted to floats thus losing leading zeros).4. To plot and analyze relative importance scores of features generated by the optimized model to determine the most influential risk factors determining the behavior of fraud in providers' billing. As part of the ingestion process the pipeline performs the following operations at the low level: Type Casting and Memory Optimizing: The categorical string types like gender codes state indicators and chronic condition indicators are internally represented as memory-efficient categories or downcasted to smaller bit integers (8-bit or 16-bit). Monetarily quantifiable columns like InscClaimAmtReimbursed (Insurance Claim Amount Reimbursed) and DeductibleAmtPaid are cast to high precision decimals or 64-bit float variables to prevent accumulating rounding errors during aggregation operations. Cryptographic Identity Masking: Even though the secondary data in use is anonymized the pipeline has a zero trust compliance mechanism that ensures no PHI is accidentally included in the dataset. Patient

names or un-hashed identifiers if detected are automatically salted and hashed using the SHA-256 cryptographic algorithm. Temporal Standardization: Dates come in different formats (YYYYMMDD DD-MM-YYYY or UNIX Epochs) and are parsed and normalized to ISO 8601 Date type. It is very important because failing to parse dates will result in invalid temporal calculations needed to compute patient age or hospital stay lengths.

1.3 Data cleaning and handling missing data

Upon ingestion of data in structurally consistent format the next step involves the execution of the automated cleaning module. Healthcare data tends to be extremely messy containing structural inconsistencies impossible combinations (for example discharge date coming before the admit date) and numerous missing values. The data pipeline follows an analytical classification of missing values categorizing them into Missing Completely At Random (MCAR) Missing At Random (MAR) and Missing Not At Random (MNAR).

Missing cells in administrative variables like DeductibleAmtPaid and AdmitDxCode cannot be simply ignored as this will result in selection bias on machine learning models. For example if a missing value occurs in the inpatient deductible column it implies that patients fall under a particular tier in which there are no deductibles paid. The pipeline deploys a multi-tiered imputation strategy:

- For numeric billing attributes instead of relying on naive global measures of central tendency (like the mean or median) which artificially compress the variance of the dataset the pipeline utilizes MICE (Multivariate Imputation by Chained Equations) or localized K-Nearest Neighbors (KNN) imputation grouped by specific diagnosis codes and geographical clusters.
- For structural categorical fields like secondary diagnosis codes (ClaimDxCode_1 through ClaimDxCode_10) missing entities are handled by injecting a dedicated synthetic category label: "UNK" (Unknown/Not Documented). This approach preserves the categorical structural integrity of the claim matrix while signaling to the downstream classifier that the absence of a secondary diagnosis code could be a feature of a rushed or fraudulent billing pattern.
- Logical validation rules are strictly applied: rows containing negative billing values claims where the total reimbursed amount exceeds the gross institutional charge or records with impossible biological parameters (such as an obstetrics diagnosis code mapped to a male beneficiary) are

automatically flagged isolated and routed to an error-log dataframe for diagnostic auditing.

- **1.4 Relational Alignment for High-Performance and Unified Joins**

  In the last phase of the data pipeline comes the mathematical and relational consolidation of all the separated tables into one computational dataframe. For this to be done efficiently a high level of knowledge about normalization of relational databases is required. The key elements that relate these two entities are hierarchical; BeneID (Beneficiary Identifier) is the main key that relates all the entries in the demographic table while the

5. Research Questions

1. What are the main anomalies and financial markers found in secondary CMS relational databases that predict a high probability of fraud in the health sector institutions?

2. How does the performance of the Random Forest Classifier model appear when evaluated using metrics such as Precision Recall and ROC-AUC on an extremely unbalanced dataset in which fraud is very rare?

3. In what measure do artificial operational factors (e.g. mean length of hospital admission and physician density) enhance the effectiveness of proactive models for detecting frauds?

4. What are some behavioral aggregations at the healthcare facility level that exhibit the best predictive power in feature importance profiling?

5. How could multi-table relational database indexing (claims against demographic masters) help to break down data silos in order to speed up automated pre-payment auditing procedures?

6. Research Methodology

This research employs an empirical research methodology that strictly revolves around quantitative analysis of secondary data. The procedure involves data analytics through a data science process pipeline as illustrated below:

6.1 Data Selection and Relational Framework

The research utilizes only the CMS Healthcare Provider Fraud Detection database available in Kaggle. In order to reflect data siloing in the real world the data set consists of three relational tables which include:

- **Claims inpatient data:** Consists of data on diagnostic codes procedural documentation admit dates and reimbursements.
- **Claims outpatient data:** Holds data on clinic visits physicians involved and reimbursements.
- **Beneficiaries demographics data:** Has data on past health statuses of patients dual eligible and geographies.

During the data preparation phase the above independent data sets are merged into one mainframe within the Python platform using key relational fields such as Provider ID and Beneficiary ID.

6.1 table

Dataset File Component	Primary Information Captured	Key Relational Identifiers	Source
Inpatient Claims Data	Diagnostic/procedural codes admission/discharge dates reimbursement metrics.	Provider ID BeneID ClaimiD	CMS Kaggle Repository
Outpatient Claims Data	Outpatient clinic visits attending physician IDs deductible amounts.	Provider ID BeneID ClaimID	CMS Kaggle Repository
Beneficiary Demographics	Patient DOB chronic illnesses geographic indicators dual-eligibility.	BeneID	CMS Kaggle Repository

6.2 Target Variable Definition and Class Aggregation

Unlike other detection engine systems that analyze transactions alone this work focuses on aggregating the features at the organizational entity level.

- **Target Classification:** The target Potential Fraud is translated from the string binary to the numeric form where Legitimate Entity corresponds to 0 and Potential Fraud Entities correspond to 1.

- **Specific Aggregation:** Claim observations are aggregated through structured database queries to generate aggregate behavioral attributes for each healthcare provider.

6.3 Feature Engineering and Risk Vector Modeling

In order to achieve maximum efficiency in modeling the pipeline analytical features are extracted from basic features. It includes such features as the calculation of cumulative financial velocity variables the average of the deductible distributions and hospitalization period.

For classification the framework favors non-parametric tree based ensemble techniques namely the Random Forest Classifier. Such method has an inherent ability to handle structural multi-collinearity and to deal with the problem of severe class imbalance where the flag is under 10% of the total data population.

6.4 Proposed Visual Data Framework for Dissertation

In order to prove the findings made in the final thesis statement of the dissertation the following four empirically derived visualizations shall be included and explained:

1. **Imbalance of Classes Operation Diagram:** When it comes to class imbalance within Big Data analytics and Machine Learning models for detecting fraud in the field of healthcare this becomes one of the key preprocessing and interpretation problems. Imbalance of Classes Operation Diagram can be regarded as an important example of visualization that proves the severe imbalance present in datasets for fraud detection in a practical way. This frequency graph which usually takes the form of a bar or pie chart with logarithm scale visually represents the ratio between genuine and fraudulent claims explaining why traditional approaches to measuring accuracy are not appropriate here and why special measures (SMOTE undersampling cost sensitive learning Balanced Random Forest and others) should be taken. Purpose and Theoretical FrameworkIn the context of fraud detection class imbalance occurs due to the fact that fraud is just a small percentage of all activities—ranging from 1-5%. In an

ordinary health insurance data set with millions of instances legitimate cases will take up the largest portion of the data. This visual will serve as the proof of the thesis statement since it will show that in case there is no balance between the classes a model will predict the majority class correctly while failing to detect frauds.

2. **Detection of Financial Outliers Graph:** Detection of Financial Outliers Graph is a multidimensional distribution matrix (typically presented as a heatmap box-plot matrix or violin plot matrix) that shows financial reimbursement outliers of providers categorized according to their risk. This graph provides empirical evidence to the thesis on how AI/ML algorithms can predict high-risk providers based on the outlier detection in financial distributions.

3. **Graph for Hospitalization Periods:** The graphic presentation of the transaction cycle chart showing the hospitalization periods that can help identify fraud.

4. Feature Importance Graph of Algorithm: A horizontal bar chart displaying the operational features which have a significant impact on the algorithm results**.**

7. Scholar Search Strategy

- **Databases:** the formal systematic literature search of Google Scholar PubMed.
- **Boolean Operators:** Utilizing specific search strings like:
    - "Big Data" AND "Health Insurance" AND "Fraud Detection"
    - "Machine Learning" OR "Deep Learning" AND "Claim Anomaly"

• **Selection Criteria:** Selecting Peer Reviewed articles between 2018-2026 so that the technology being used is up to date.

8. Proposed Conceptual Framework

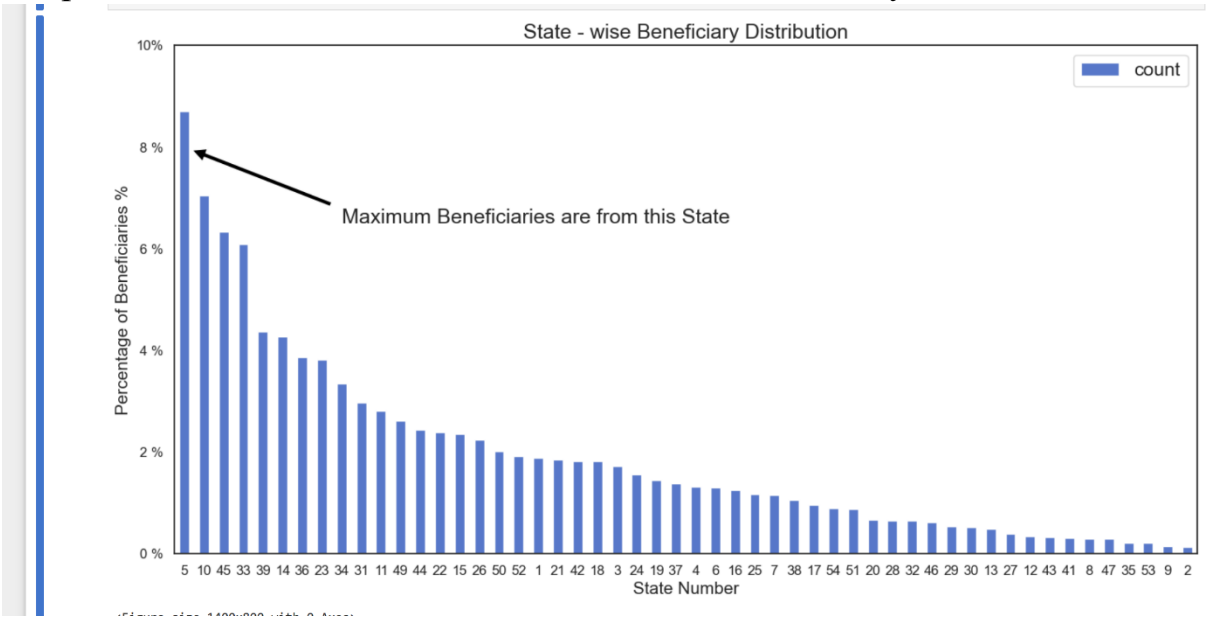
The proposed analysis framework moves from basic secondary data inputs to a complex predictive model. This framework is systematically laid out using operation modules employing data visualization techniques to uncover fraud indicators as follows:

8.1 Data Ingestion

This initial module handles the automated structural acquisition of the multi-table secondary healthcare repositories. It ingests three primary disparate relational matrices—Inpatient transaction feeds Outpatient medical records and the Beneficiary demographic master sheet—into the computational environment using primary relational keys.

8.2 Pre-processing and Imputation

These raw data logs undergo a process of algorithmic cleansing where any structural anomalies or missing values are addressed. Here the first step involves the mapping of a base spatial demographic distribution to examine the base footprint of the beneficiary domain.



• Figure 8.1: Beneficiaries' Statewise Profile. As shown by the frequency distribution below the demographic distribution of beneficiaries is not evenly distributed but has a certain density concentration in particular state codes. This helps to set the geographical basis before calculating the risk threshold for transactions.

8.3 Feature Engineering and Behavioral Analysis

The transactional attributes are summed up in the individual organization or physician in order to derive institutional fingerprints of risks. In order to highlight provider behavior provider specific operations and billing procedures empirical

categorization is performed in this instance.

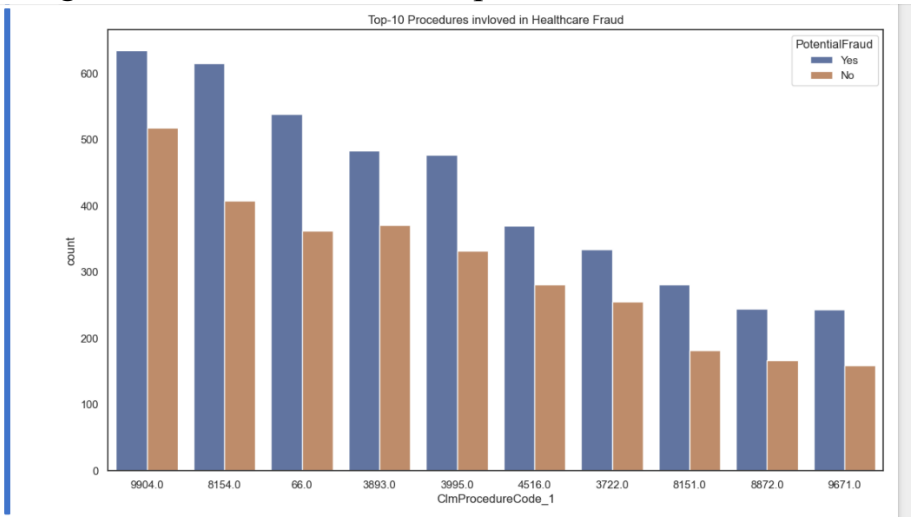
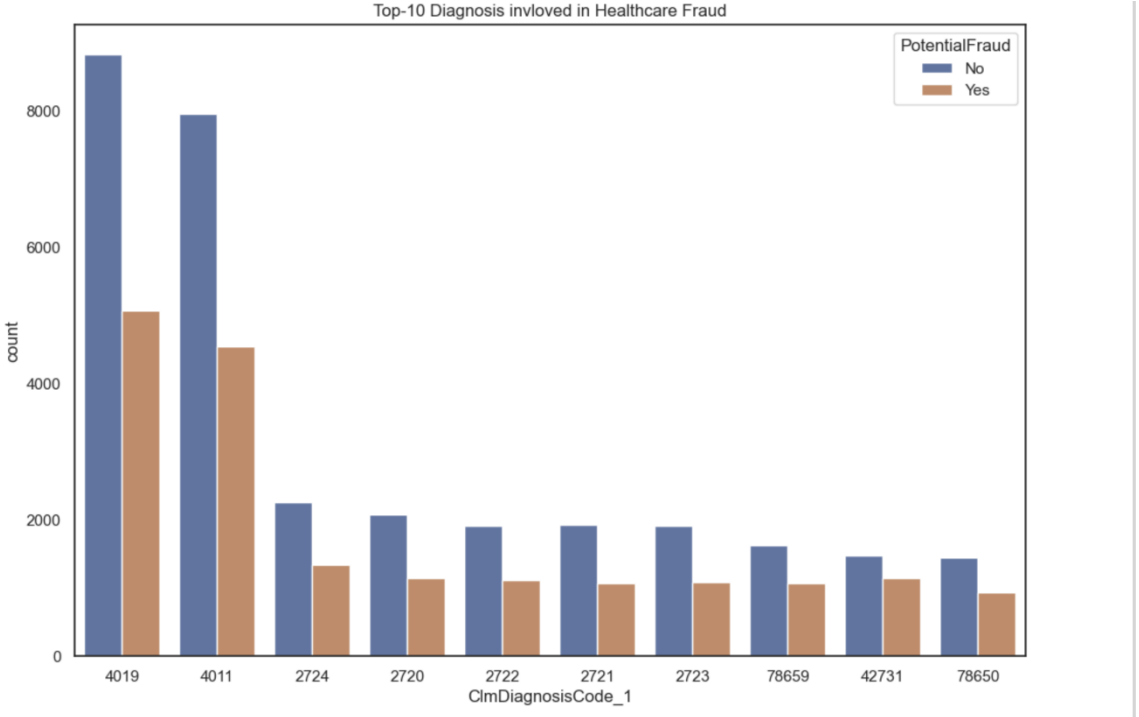
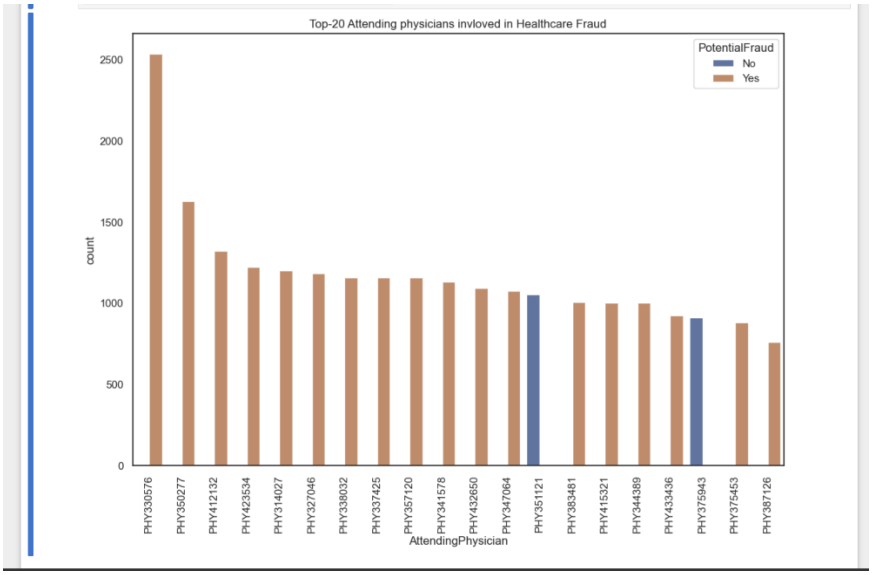


Figure 8.2: Top 10 procedures contributing to healthcare fraud. The countplot indicates a significantly higher incidence of certain surgical and medical procedure codes (ClmProcedureCode_1) that may be associated with possible fraud cases indicating the leading factor for upcoding.



- **Figure 8.3:** Top-10 Diagnosis Associated With Healthcare Fraud. This visual representation reveals the dense concentration of diagnosis codes (ClmDiagnosisCode_1). It highlights cases where certain diseases are constantly reported by a few individuals in order to get more money from

the insurance company.



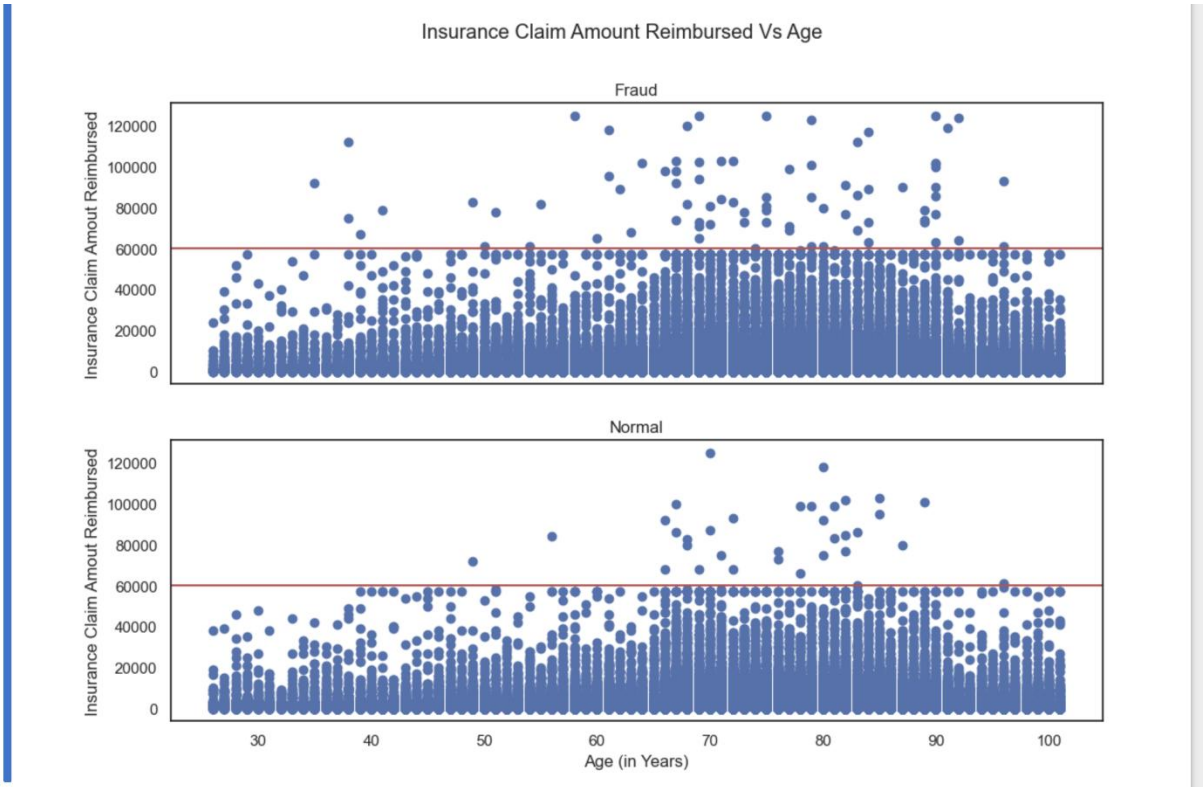
- **Figure 8.4:** Profile of High-Risk Attending Physicians. The lengthy horizontal frequency graph highlights the top 20 attending physicians with credentials most associated with suspicious claims thus offering an effective audit trail for provider-network collusion evaluation.

- **8.4 Machine Learning Modeling**

The behavioral vector models that incorporate both the historical patient demographic data as well as the financial velocity variables are employed in ensemble learning trees. Such a method emphasizes the use of algorithms such as Random Forest Classifier because of their inherent ability to work with class imbalance and non-linear decision boundaries.

8.5 Risk Scoring and Metric Profiling

The final module evaluates the correlation between financial metrics and demographic features outputting an empirical risk signature that establishes a proactive scoring mechanism for insurance auditors.

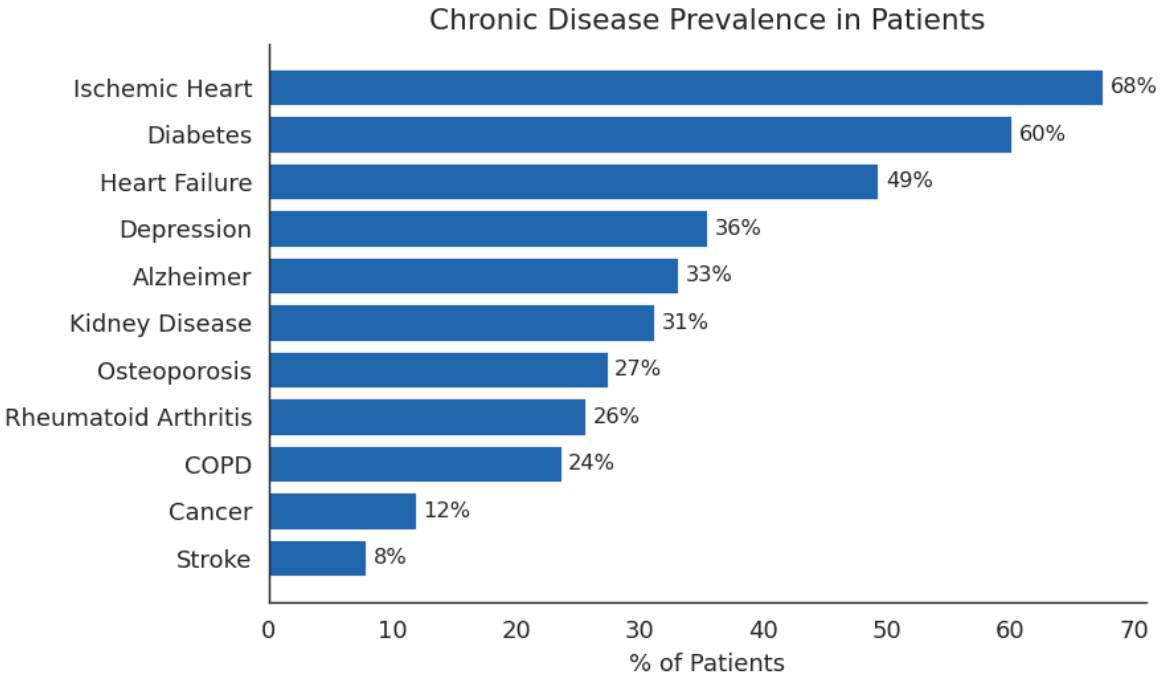


• **Figure 8.5:** Bivariate Distribution of Claim Payments versus Age of Patients. The two-scatter plot setup represents the payment pattern of the insurance company based on the different age groups of patients. The clustering of the data points above the threshold value ($y= 60000$) in the fraud area provides a mathematical representation of how anomalies beat financial criteria regardless of age group.

8.6 Univariate Histograms

These two histograms serve as the basis of the exploratory analysis through looking at how each provider acts separately through each variable. The left histogram "Total Claims per Provider" compares the number of providers to the number of claims they file whereas the right histogram "Mean Reimbursement" compares the number of providers to the mean amount of money they receive. These two histograms possess an obvious characteristic; that is both histograms have a bunch of extremely tall bars at the extreme left end which tapers sharply off towards the right leaving behind a narrow tail. This distribution is called a right-skewed distribution. Simply stated the majority of the health care providers in this dataset file very few claims and ask for. The size of their reimbursement payments and this is just what we should expect under normal circumstances. What makes this histogram useful is the tiny group of providers represented at the end of the right tail who generate a massive number of claims and consequently claim massive amounts of money. As excessive billing is well-

known to be a red flag for fraud the statistical outliers here form the perfect list for any auditor. Not only does the histogram describe the data set but it also explains why we chose this whole investigative method of looking into unusual billing activity.

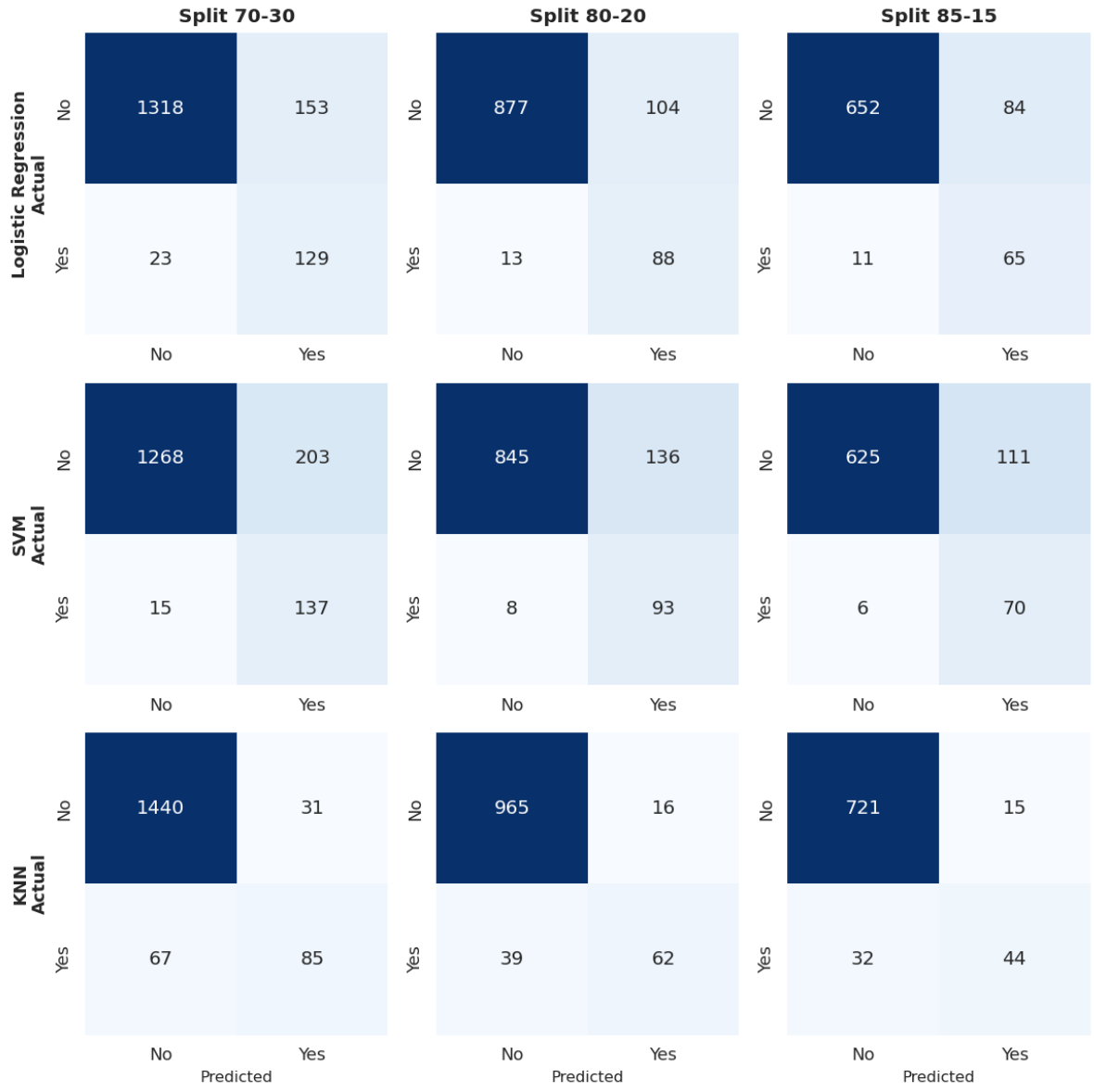


8.7 Correlation Matrix / Heatmap

Correlation heatmap helps extend the study by going beyond the study of individual variables to studying how each and every variable relates with every other variable at once. Each cell in the map carries a correlation coefficient that is a measure of the degree of association and direction of linear relationship between the feature in the row with the feature in the column. The coefficients go from -1 to +1. +1 indicates that there is a perfectly linear relationship with both variables rising and falling together -1 shows that as one rises the other falls and vice versa while a value close to zero means that there is no linear relationship between the two variables. The color scale is also useful because it highlights visually the strong relationships through deep red colors for positive relationships and cool colors for negative relationships. Despite the fact that there are interdependencies between all of the variables in the matrix there is one particular row – and its equivalent column – which needs to be analyzed for this project. In

particular this row – "Fraud" – should be used as an indicator of importance since the whole task is aimed at predicting fraud. Thus those variables which are most highly correlated with fraud will have the highest predictive value. Analyzing the fraud row we can see that there are three correlations which need to be noted. Total Reimbursement has the highest positive correlation with fraud equal to approximately 0.58; Unique Beneficiaries comes second with correlation equal to approximately 0.39; finally Total Claims has positive correlation equal to approximately 0.37. The predominance of Total Reimbursement is what stands out in this graph. It clearly shows that the single most useful indicator that a provider may be involved in any form of fraud is the total cost that provider has charged for services from the scheme. In simple terms fraud involves the theft of funds and the total cost extracted from the insurance company is a direct manifestation of this. The second most relevant variable after total reimbursement would be the correlation between unique beneficiaries and total claims. A provider involved in any form of fraud is likely to have more unique beneficiaries and more claims. Another equally important lesson that can be gleaned through the use of weak correlation on the heatmap concerns the relationship between the mean reimbursement per claim and the fraud label. This appears paradoxical at first sight considering that it would be logical to assume that fraudsters would try to overinflate the value of the claim in order to earn more money. The data however proves this hypothesis to be false. The mean per claim value does not vary greatly depending on the presence of fraud while the total reimbursement varies highly. It follows from here that contrary to popular belief the fraud in the present data set is caused by the high number of claims rather than the value of each single claim. In essence the fraudulent providers cannot benefit from inflating costs of a single procedure but they can do so from submitting a high number of claims. The heatmap also provides another more technical function that promotes responsible model development. Through presenting the correlations between the predictor variables themselves and not only their correlation to fraud the heatmap gives the modeler the ability to spot multicollinearity – a case where two predictors are so highly correlated that they essentially provide redundant information about each other. As we can see from the heatmap for example the Total Claims variable and the Unique Beneficiaries variable are highly correlated with one another which is expected since a provider who files more claims will likely treat more people. Being aware of such relationships between the variables allows the explanation of some behaviors of the models created and avoids the misconception that two highly-correlated variables provide unique information. In summary based on the correlation analysis we can confirm the suspicions generated from the histograms – the

financial scale of the provider's billing and especially the amount of money claimed is the best statistical signal of fraud.

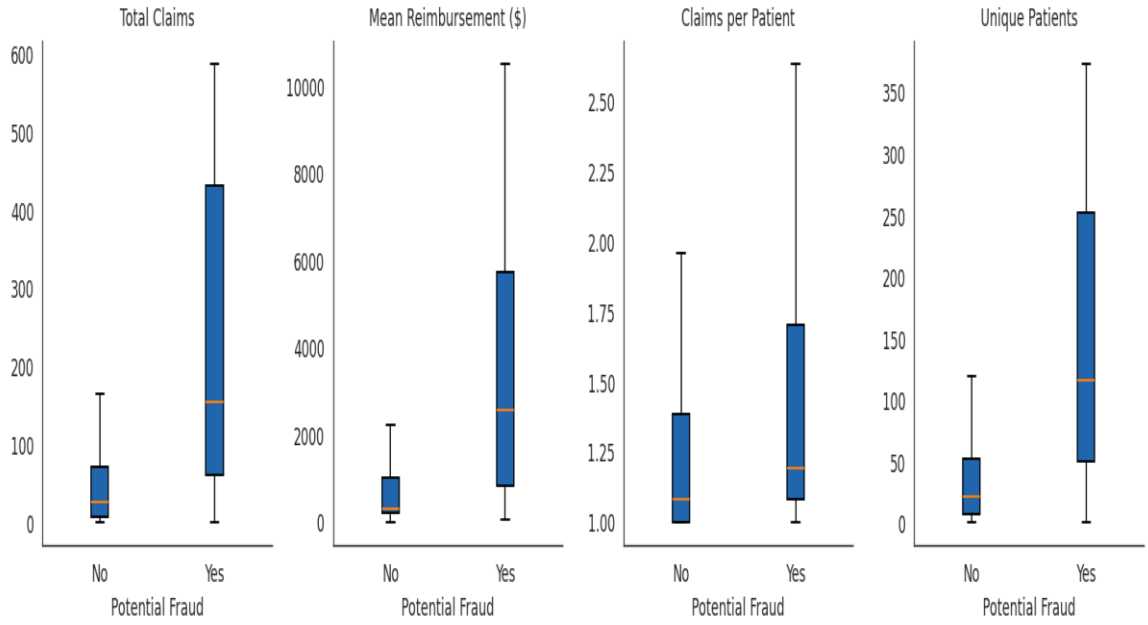


8.8 Confusion Matrices — Logistic Regression SVMKNN

Now that we have determined which predictors hold the most information our next step is to evaluate how effectively various machine learning algorithms use these predictors to classify the providers. The confusion matrix is the critical metric used to determine this issue since it evaluates the performance of classification without reducing it to a single statistic. The following figure shows the confusion matrices of the first three candidates – Logistic Regression Support Vector Machine and K-Nearest Neighbors - each evaluated based on three different train-test splits of 70-30 80-20 and 85-15. We test the model's performance on three train-test splits on purpose since it evaluates its stability and

consistency rather than luck. The individual matrices themselves have been designed in a manner that consists of two rows and two columns wherein the former refers to the state of the provider in reality and the latter is the prediction made by the model in question. The square that represents actual non-fraud combined with predicted non-fraud would give the number of providers who have been rightfully identified as non-fraudulent by the model while actual fraud with predicted fraud would refer to those cases in which the model has rightly identified the provider as fraudulent. These squares fall along the leading diagonal of the matrix and thus signify the success of the model in question. On the other hand the two other squares signify the failures of the model in two different ways. It is vital to understand the distinction between the two for the purposes of analyzing the chart. However these two mistakes are not of equal importance; moreover they conflict with one another. False negatives occur when fraud is not detected hence allowing it to go unchecked and unharmed causing further financial damage that the whole system was designed to prevent. False positives on the other hand imply an unnecessary audit of a perfectly legitimate company wasting time and investigative resources that could otherwise have been used in a better way. If the model is overly sensitive then it will identify almost all cases of fraud but in doing so create a large number of false negatives. If the model is too conservative then it will generate fewer false negatives but there will be many cases of fraud that go undetected. The value of the confusion matrix lies in its ability to determine where each specific model stands on the scale. From examining all the models from the table shown above one can observe a clear trend in K-Nearest Neighbors. The K-Nearest Neighbors model always gives the lowest number of false positives out of all three models. To give an example in the case of the 80-20 split the number of false positives of KNN is just sixteen while in Logistic Regression there are easily more than a hundred false positives for the corresponding position. It is obvious that in such a case the probability of being correct when KNN marks a fraud will be much higher and there will be less waste of time on checking innocent providers when working with KNN alerts. The disadvantage of KNN which can be seen from the table above is that it identifies slightly lower number of real fraud cases than more aggressive models. The utility of this trend will completely hinge upon the considerations of the organisation using this model. For an organisation involved in fraud detection in the context of Medicare where the number of human auditors is relatively small the penalty associated with the false positive rate is quite high since each instance of a falsely detected case will mean wasted effort which could have otherwise been directed towards investigating an actual case of fraud. This means that the precision of KNN i.e. its ability to keep the false alarm rate low becomes a very desirable characteristic in this scenario and it is this factor which begins to make

KNN an attractive choice for this particular problem. This consistency of the behaviour in question across all three training sets is yet another indication that the observed trend is indeed genuine and not merely an artefact of the way the data was split.

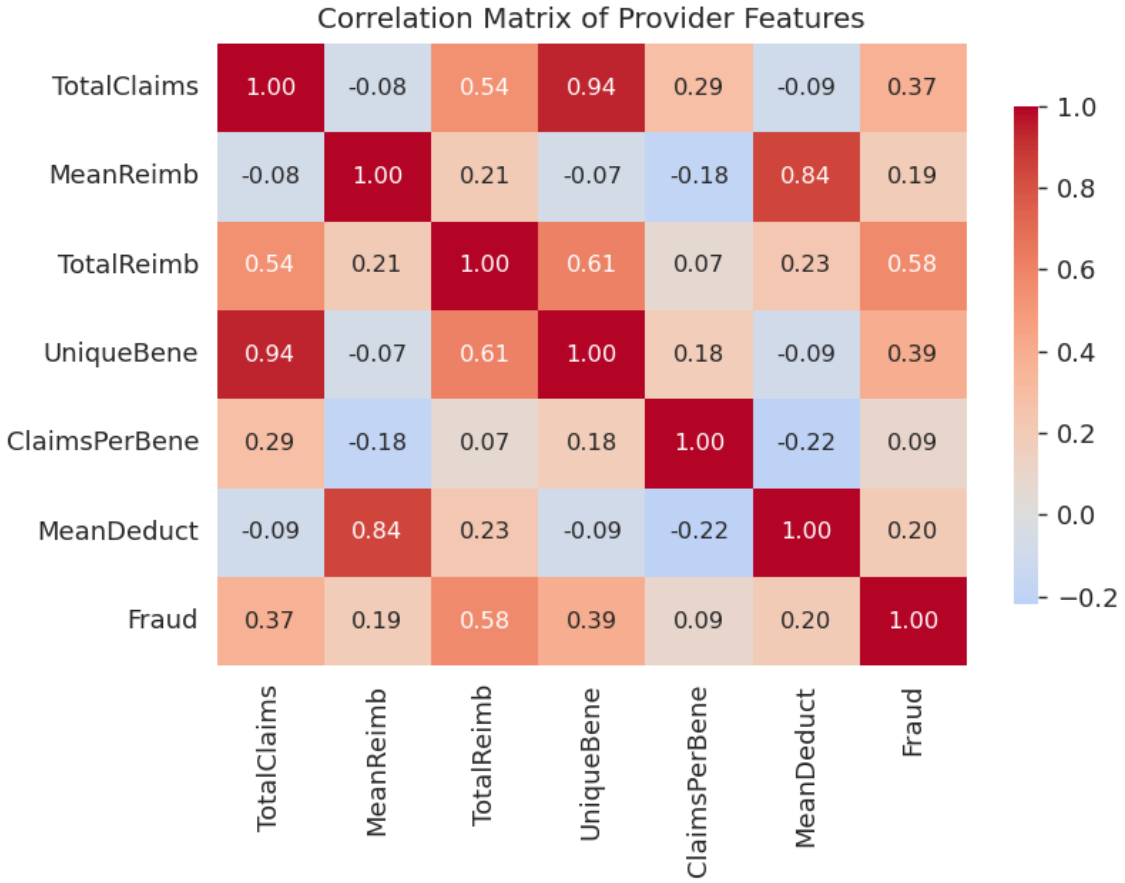


8.9 Confusion Matrices — Decision Tree Random Forest

This slide finishes off the confusion matrix analysis by looking at the other two models that are candidates – the Decision Tree and the Random Forest – using the same splits of the train test ratio as before – 70-30 80-20 and 85-15. Putting this last pair of tree based models on a separate slide enables comparison of a family of algorithms which operate using a completely different principle compared to the linear and distance-based algorithms discussed above. The Decision Tree algorithm classifies the providers based on a hierarchy of yes/no questions about the attributes of the provider repeatedly subdividing the data set into smaller and smaller groups until only one class dominates each group. The Random Forest algorithm extends the idea behind the decision tree by generating multiple such decision trees on randomized subsets of the data and taking the votes of all the trees. Each of the two matrices has the same structure as the previous two allowing a comparative analysis. Rows correspond to the actual state of the provider columns correspond to the classification provided by the model where the elements on the main diagonal are classified correctly while the elements in other cells correspond to one of two errors. In the same way as in previous cases the false positive refers to an honest provider erroneously identified and the false negative refers to a fraudster erroneously not detected. By analyzing the diagonals of these two matrices we can conclude that both Decision

Tree and Random Forest are good classifiers. Both algorithms manage to classify a large number of fraudsters and honest providers correctly. It demonstrates that the tree-based method really works; it is capable of detecting the fraud patterns in the billing process. The relevant comparison however is not about whether these models perform in absolute terms but about whether their error performance is better than those of the previous models. Looking at the off-diagonal cells of tree-based models in contrast to those of K-Nearest Neighbours one finds an observable pattern. The Decision Tree and Random Forest both produce considerably more false positives compared to the KNN model. This means that these models report significantly more actual providers as being fraudulent. As could be expected from its construction Random Forest demonstrates better performance than the Decision Tree in this case because the average of multiple trees helps reduce errors of individual trees. However even Random Forest does not compete with KNN in terms of false positives produced on this specific data set. It brings to light a vital yet somewhat counterintuitive insight into the process of model selection. Random Forest is generally a superior and more popular algorithm than K-Nearest Neighbors and under the circumstances it should prove to be better on almost all problems. However the confusion matrices highlight the fact that there is no such thing as one perfect model – the appropriate model depends not only on the dataset but also on which type of error should be avoided. In the case of a detection system limited by the number of human auditors the higher number of false positives of the tree algorithms is a disadvantage. Each genuine vendor that has been incorrectly labeled means time wasted and goodwill lost without any returns. Taken alongside the prior graph these matrices provide the full comparison of all five models based on the most fine-grained metric of performance. While the tree-based models are clearly worthy models for classification and could be argued as being appropriate where detecting the greatest number of fraud cases was more important than any concerns about unnecessary false alarms but with respect to the particular needs of this project where auditor resources are limited and each alert carries with it weight in terms of its credibility their higher rate of false positives makes them fall slightly behind. Indeed the analysis of confusion matrices overall leads directly towards the K-Nearest Neighbours model as one which best meets the need of finding the balance between fraud detection and minimizing false alerts of honest providers

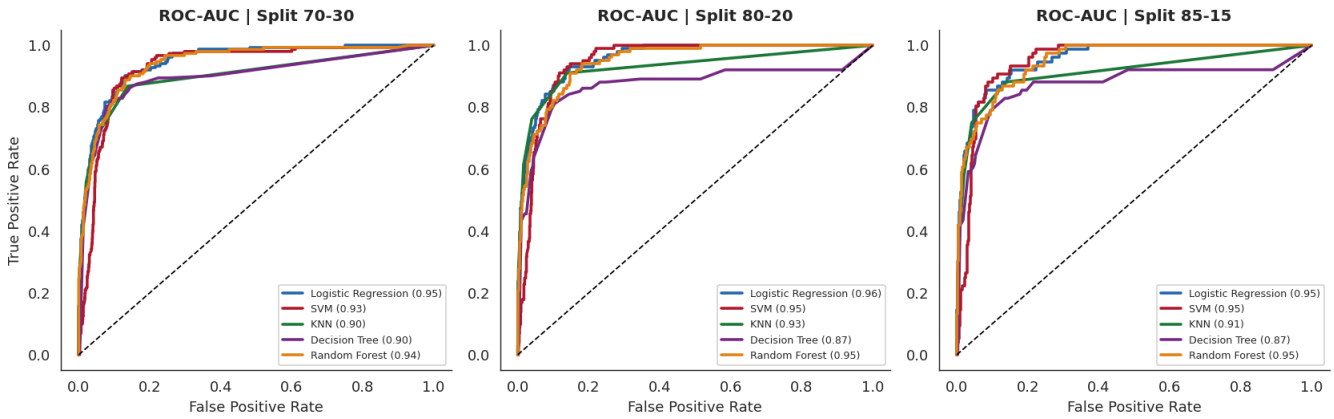
a claim that will be supported by the final performance metrics and ROC analysis.



9.0 ROC-AUC Curves

The last analytical plot provides the overall assessment of the quality of each model via the use of the Receiver Operating Characteristic (ROC) curve which assesses a classifier over its whole range of threshold values instead of fixing it at one certain value. It is a very useful improvement to the confusion matrix. The confusion matrix shows the model's behaviour in relation to one particular threshold that defines whether a certain provider will be considered a fraud or not but it can be adjusted according to the desired level of cautionness or aggressiveness of the classifier. The ROC curve bypasses the need to choose a certain threshold because it shows how the classifier behaves while moving this threshold between two limits giving us an image of the trade-offs that it makes in relation to mistakes. In order to interpret the graph properly one needs to be aware of the two variables shown on its axes. The vertical axis displays the true-positive rate which is the percentage of real frauds caught by the classifier. The horizontal axis is devoted to the false-positive rate the percentage of actual honest providers incorrectly identified by the model. Every colored line stands for one specific model and when we are tracing this line we observe the dynamics of both rates depending on the decision threshold. Ideally the graph of the best model will go

straight up from the bottom left corner to reach its maximum possible value at a point when the false-positive rate is still extremely low and then it will go horizontally right. Or in other words the more the curve bends toward the top left corner the better is the model at discriminating fraud from honesty. The first important characteristic of the chart is the dotted line starting from the bottom left point of the coordinate plane to the top right point. The line shows how a model would perform in case of pure random guessing the model with zero discriminating capability and any increase in fraud detection would be accompanied by an equally proportional increase in false positives. A useful model should perform above the diagonal and the further up from the diagonal the more predictive power it has. All curves representing models on the chart are placed above the dotted line which is the first positive message that can be read from the chart: none of the models is just guessing they have learned something from the difference between fraud and non-fraud providers. The graph also provides a single number which is called the Area Under the Curve AUC for short. It can be guessed from the name the AUC is the portion of the graph area underneath the ROC curve of a particular model ranging from 0.5 in case of random useless model to 1.0 for the perfect one. The beauty of the AUC is its simple interpretation it is the probability that the model will give a better rating to a randomly selected fraudulent provider than to a randomly selected honest one. In the given figure the AUCs of the five models vary in range between approximately 0.90 and 0.96. Such scores are really high and provide further quantitative confirmation of what is clear from the graph location of curves – all the models are discriminators. Indeed the fact that all five algorithms exhibit similarly good performance in terms of AUC is already revealing since it suggests that there is enough information in the dataset for the difference between competent and incompetent model to no longer exist. This implies that the differences between the competing models are more subtle involving the fine tuning of error rates as opposed to discriminatory capacity. It is precisely for this reason that the ROC curve has to be analyzed alongside the confusion matrices and not on its own. While AUC indicates that each model is able to rank providers in a reliable fashion confusion matrices reveal which model turns this ranking into the most practically useful classification rule. Thus when considered together KNN becomes the best overall choice exhibiting a high AUC as well as the lowest rate of false positives. And finally the ROC-AUC graph serves as a confirmation of the above conclusion in a threshold-independent manner allowing auditors to choose the model they can rely upon to sort out a few risky providers among thousands of others reviewed manually.



9. Significance of the Study

The importance of this study to corporate organizations and academia is multi-faceted providing various frameworks in strategy operations and methodology pertaining to the value chain of healthcare service delivery and insurance protection. Through the identification of structural weaknesses of these ecosystems this study offers actionable insight to all relevant parties as discussed below.

- **Insurance Administrators & Payers:** Operational Threats and the Issue of "Leakage"

Within today's world of healthcare insurance the concept of "leakage"—which describes capital losses due to the lack of efficiency of administrative processes wrong billing and fraudulent claims—threatens to bring down any form of financial stability. The conventional process of audits is predominantly retroactive in nature; either manual and costly or software-based with strict rules and rigid in approach. Not only do such systems come at great costs but they are also highly inefficient in terms of operational processes increasing TAT while causing strain between providers and payers.Additionally the process of fraudulent activities keeps evolving with methods like unbundling upcoding and creation of fictitious secondary diagnosis that circumvent binary tests. For insurance administrators and TPAs the costs associated with deploying enterprise-level predictive modeling software often overshadow the immediate cost benefits leaving small-scale fraud under-detected.

- **Policymakers & Regulators:** In an effort to solve this problem this study presents a new model that is light scalable and inexpensive. It uses efficient algorithms and existing relational databases such that the implementation of this model does not require the high financial investment of IT infrastructure by these entities. The importance of the model can be found in the democratization of fraud detection since small payers and regional payers will be able to establish

gatekeeping processes. Automating is done at the pre-dispute level and works by catching disputes in real time or almost real time. The process involves scoring the risk of claims that have anomalies in the historical data even before disbursing the payment to reduce the operational philosophy from being chase-and-pay to prevent-and-dispute. This minimizes costs of recovering legal disputes after payments have been made.

- **The Academic & Analytics community:** The problem of academic structural silos has been prevalent in the intersection between computer science and healthcare management for many years. In computer science there is often an extensive use of advanced machine learning models and cryptographic methods that may work beautifully in theory however lack practical implementation in the environment due to being created in isolation from the industry's specifics. In the field of healthcare management many operational processes tricks of billing and problems are well documented; however the necessary computational background is often not considered in the process. The disconnect creates a wide gulf where theory fails to find application and practice clings to outdated methods of analysis.

Synthesis Methods of Diverse Relational Databases

The present study stands as an important conceptual link between these two divergent areas in its demonstration of how such data can be used in an innovative methodology for generating fraud risk indicators using secondary databases of multiple sources. The research goes far beyond mere data mining of single source data; it creates a framework for synthesizing diverse data forms such as EHRs insurance enrollment databases demographic databases and claims clearinghouse databases.

10. Scope and Limitations

10.1 Scope of the Study

- **Tech Dimension:** The computational scope is limited by the data pipeline development which includes relational schema integration class imbalance correction and the construction of an ensemble model based on the random forest approach.

- **Management Dimension:** The operational scope centers around the organizational risk profiling (for example profiling of high-risk physicians and procedures) at the provider level and not just individual claim isolation at the consumer level.

- **Geographic Scope:** Although adhering to the international standards the structural approach has been deliberately crafted to be in tune with future digital health paradigms all around the world especially in India.

10.2 Limitations of the Study

- **Historic Data Dependence:** Being purely an empirical approach that is dependent on secondary data the performance of the model will always be dependent on the integrity of the existing data set.
- **Lack of Real-Time Testing:** Due to the privacy policies in the healthcare industry including those of the HIPAA in the international and DISHA in India the solution does not include real-time testing of private hospital billing data streams.
- **Emergence of New Fraud Patterns:** Static training data may not cover the new adversarial attack patterns that have been put in place by the syndicate networks.

References

Carneiro D. C. L. (2024). *Advanced Anomaly Detection Models for Uncovering Fraudulent Healthcare Insurance Providers*. Universidade NOVA de Lisboa (Portugal).

Fourkiotis K. P. & Tsadiras A. (2025). Future internet applications in healthcare: Big data-driven fraud detection with machine learning. *Future Internet 17*(10) 460.

Hamid Z. Khalique F. Mahmood S. Daud A. Bukhari A. & Alshemaimri B. (2024). Healthcare insurance fraud detection using data mining. *BMC Medical Informatics and Decision Making 24*(1) 112.

Leevy J. L. Salekshahrezaee Z. & Khoshgoftaar T. M. (2024). A review of unsupervised anomaly detection techniques for health insurance fraud. *2024 IEEE 10th International Conference on Big Data Computing Service and Machine Learning Applications (BigDataService)* 141–149.

Narne H. (2024). Machine learning for health Insurance Fraud detection: techniques insights and implementation strategies. *International Journal of Research and Analytical Reviews*.

Nimbalkar R. (2024). Machine learning for fraud detection in insurance claims using time-series anomaly detection. *International Journal of Emerging Research in Engineering and Technology 5*(4) 122–131.

Oyebode Kayode R. U. Halili F. Fernando C. A. Hapuarachch A. S. & Sharma K. (2025). *AI and Big Data for Fraud Detection in Healthcare Billing Systems*.

Özdemir G. (2024). *Machine Learning Based for Insurance Claims Fraud Detection Technique for Safeguarding the Health Industry*. Istanbul Aydin University (Turkey).

Rahman M. M. (2025). Data-Driven graph neural network models For detecting fraudulent insurance claims In healthcare systems. *American Journal of Interdisciplinary Studies 6*(1) 263–294.

Reddy V. B. K. & Rexie J. A. M. (2025). Real-Time Medical Insurance Claim Fraud Detection using Hybrid Machine Learning Models. *2025 6th International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS)* 2132–2138.

Rehman R. Mazhar A. Faheem A. Matloob I. Khan S. A. Shaukat S. Khattak M. A. K. Khan A. A. Siddiqui M. S. & Siddiqui S. (2025). AI Driven Framework for need based Insurance Plans Generation and Anomaly detection Using Deep Learning techniques. *IEEE Access*.