Healthcare Provider Fraud Detection

Ayomide Ayeni
Department of Computer Science, Georgia State University, Atlanta, GA, USA

Abstract—*Fraud in healthcare is a direct threat to patient trust, medical integrity, and the sustainability of healthcare systems. With billions lost annually to deceptive billing, unnecessary treatments, and provider kickback schemes, fraudulent practices are becoming increasingly sophisticated. This study takes a proactive approach to healthcare provider fraud detection by leveraging machine learning. Rather than relying on rigid rule-based systems, this research employs supervised learning models—Random Forest, Support Vector Machine, and K-Nearest Neighbors—to expose hidden fraud patterns in medical claims data. The best-performing model, a Random Forest classifier with depth 25 and 1,000 estimators, achieved 86.88% accuracy, an F1-score of 0.819, and an AUC-ROC of 0.94. The study also critically examines data accessibility, regulatory compliance (HIPAA, FCA, Anti-Kickback Statute), and the trade-offs between privacy and fraud prevention, proposing an industry-ready framework that balances accuracy, interpretability, and real-time fraud mitigation.*

Keywords—*healthcare fraud detection, machine learning, random forest, anomaly detection, Medicare fraud, claims data, HIPAA, false claims act, imbalanced classification*

I. INTRODUCTION

A. Background of Healthcare Fraud

Healthcare fraud is a pervasive issue affecting public and private healthcare systems worldwide. It involves intentional deception or misrepresentation by healthcare providers, patients, or other entities to gain unauthorized benefits from insurance programs, government-funded healthcare, or private payers. The complexity of healthcare billing systems, the volume of transactions, and the intricate nature of medical procedures create an environment where fraudulent activities can thrive undetected for extended periods.

Fraudulent schemes manifest in various forms, including phantom billing—charging for procedures that never occurred; upcoding—inflating the cost of services; unbundling—separating bundled services to maximize reimbursement; kickback and referral fraud; falsification of medical records; and prescription fraud. The financial impact results in billions of dollars in annual losses, increasing insurance premiums and healthcare costs for both providers and patients.

As healthcare systems become increasingly digital, with electronic health records (EHRs), telemedicine, and AI-driven diagnostics, the landscape of fraud is also evolving. The need for more intelligent fraud detection has never been greater, requiring a balance between effective fraud mitigation and the protection of patient rights and data privacy.

B. Importance of Fraud Detection in Healthcare

Healthcare fraud is not merely a financial crime—it compromises patient care, inflates costs, and undermines trust in medical institutions. Fraudulent billing costs the industry hundreds of billions of dollars every year, diverting essential resources from legitimate patient care. By implementing effective fraud detection systems, healthcare organizations can prevent unnecessary financial losses and allocate resources more efficiently.

Fraudulent medical practices can also result in harmful or unnecessary treatments, exposing patients to incorrect diagnoses, surgical procedures, or medications they do not need. Prescription fraud, especially in the case of opioid abuse and controlled substances, poses a significant public health risk. Advanced fraud detection mechanisms help prevent legal violations and safeguard institutional credibility while reducing the investigative burden through automated risk assessments.

C. Objectives of the Study

The primary objective of this study is to develop an intelligent fraud detection model that identifies and mitigates fraudulent activities among healthcare providers. Specific objectives are:

- Identify common patterns and techniques used in healthcare provider fraud, including phantom billing, upcoding, unbundling, and kickbacks.
- Develop a data-driven fraud detection model capable of distinguishing fraudulent from legitimate claims using supervised and unsupervised learning.
- Evaluate the role of Big Data and AI in fraud detection through EHRs, insurance claims data, and billing records.
- Ensure compliance with regulatory and ethical standards including HIPAA, FCA, and Anti-Kickback Statutes.

- Test the proposed model on unseen health datasets to measure accuracy, efficiency, and false positive rates.

D. Research Questions

This study seeks to answer the following key questions:

- What are the most common fraudulent schemes used by healthcare providers, and how do they impact patients and payers?
- What patterns and anomalies in claims data indicate potential fraudulent activity?
- Which machine learning techniques are most effective for healthcare fraud detection?
- How can the accuracy and reliability of fraud detection models be improved while minimizing false positives?

II. HEALTHCARE SYSTEMS AND FRAUD TYPES

A. The Healthcare Ecosystem

The healthcare ecosystem is a complex network of interconnected entities including healthcare providers (hospitals, clinics, physicians, and specialists), payers (insurance companies, government programs like Medicare and Medicaid), patients, pharmaceutical companies, regulators, and technology vendors. The ecosystem relies heavily on administrative processes—medical billing, coding, claim submissions, and reimbursements—all regulated by governmental and private institutions.

The sheer volume of transactions, combined with varying levels of oversight and complex coding systems such as CPT and ICD-10, opens numerous vulnerabilities for fraudulent activities. Understanding these foundational elements is crucial for designing effective fraud detection strategies.

B. Payment Models and Fraud Risk

Payment models define how healthcare providers are compensated and directly impact fraud risk by determining financial incentives and reporting requirements.

1) Fee-for-Service (FFS)

Under the traditional FFS model, providers are paid for each service or procedure. This creates opportunities for overutilization and billing fraud such as performing unnecessary tests or splitting services to bill separately (unbundling). Fraud risk is high, with susceptibility to phantom billing, upcoding, and overbilling.

2) Value-Based Care (VBC)

VBC rewards providers based on patient outcomes and care quality. While it encourages efficiency and preventative care, providers may avoid high-risk patients to maintain favorable metrics or underreport adverse events. Fraud risk is moderate, with susceptibility to data manipulation and outcome misreporting.

3) Capitation

Providers receive a fixed amount per patient regardless of care volume. Fraud may involve inflating patient rosters or failing to provide adequate treatment. Risks include 'ghost patients' and service avoidance.

C. Categories of Healthcare Fraud

Healthcare fraud can be committed by providers, patients, or insurers, each exploiting different system vulnerabilities. Table I summarizes the most prevalent fraud schemes and their detection signals.

TABLE I. COMMON HEALTHCARE FRAUD SCHEMES AND DETECTION SIGNALS

Fraud Scheme	Description	Detection Signal
Phantom Billing	Services billed but never rendered	No corresponding EHR event
Upcoding	Higher-cost code than procedure performed	CPT code mismatch vs. clinical notes
Unbundling	Separate billing for bundled procedures	Code-pair frequency anomalies
Kickbacks	Illegal referral incentives	Unusual referral network clusters
Unnecessary Services	Medically unjustified procedures	Volume vs. patient-condition mismatch
Identity Theft	Claims using stolen patient identity	Claims post-DOD; duplicate BeneIDs
Prescription Fraud	Opioid overprescription / doctor shopping	Multi-provider prescription overlap
Telemedicine Fraud	Virtual visits never conducted	Billing without telehealth platform log

III. LEGAL AND REGULATORY FRAMEWORK

A. False Claims Act (FCA)

The False Claims Act is one of the most powerful tools used by the U.S. government to fight fraud, particularly in government-funded programs like Medicare and Medicaid. It prohibits individuals and organizations from knowingly submitting false or fraudulent claims for payment to the federal government and has been instrumental in recovering billions in healthcare fraud settlements.

B. Anti-Kickback Statute (AKS)

The AKS criminalizes the offering, solicitation, payment, or receipt of anything of value in exchange for referrals or for generating federal healthcare program business. Violating the AKS is a felony offense and may result in fines, imprisonment, and exclusion from federal healthcare programs. Even if a claim is medically necessary, if it was influenced by a kickback, it is considered fraudulent.

C. Stark Law

The Stark Law (Physician Self-Referral Law) prohibits physicians from referring patients for certain designated health services payable by Medicare or Medicaid to entities with which the physician has a financial relationship. The law is civil rather than criminal but can lead to denied payments, required refunds, and civil monetary penalties.

D. HIPAA and Privacy Enforcement

While HIPAA is often associated with privacy and data security, it also includes provisions related to fraud enforcement. It established the Health Care Fraud and Abuse Control Program (HCFAC), which coordinates federal, state, and local efforts to combat fraud, created new criminal penalties for healthcare fraud, and expanded investigative powers. HIPAA also ensures that fraud detection systems respect patient confidentiality and data protection standards.

E. Medicare and Medicaid Fraud Prevention

As large government-funded programs, Medicare and Medicaid are primary targets for fraud. The Medicare Integrity Program (MIP) and Medicaid Integrity Program fund audits, data analysis, and fraud investigations. The CMS Fraud Prevention System (FPS) uses predictive analytics to screen claims in real-time, while Zone Program Integrity Contractors (ZPICs) are tasked with detecting billing anomalies and investigating suspicious claims.

IV. DATA SOURCES AND FEATURE ENGINEERING

A. Dataset Overview

Effective fraud detection depends heavily on the quality, diversity, and granularity of available data. This study utilizes a synthetic but realistic dataset sourced from Kaggle that mirrors the structure of real-world Medicare claims data. The dataset encompasses inpatient and outpatient billing records, patient demographics, chronic condition indicators, coverage duration, and reimbursement amounts, providing a comprehensive foundation for detecting fraudulent billing patterns.

B. Feature Categories

The dataset comprises several feature categories essential for fraud detection:

1) Patient Demographics

Patient demographic features include unique beneficiary identifiers (BeneID), dates of birth and death, gender (binary-coded), race (categorical), and geographic codes for state and county. The date of death feature is particularly valuable for identifying posthumous claims, a classic fraud signal.

2) Coverage and Insurance Information

Coverage duration features (NoOfMonths_PartACov and NoOfMonths_PartBCov) indicate how long a beneficiary has been enrolled under Medicare Parts A and B respectively. Short coverage duration associated with high-cost claims is a significant indicator of front-loaded fraud schemes.

3) Chronic Condition Indicators

Ten binary columns indicate the presence of specific chronic conditions including Alzheimer's disease, congestive heart failure, kidney disease, cancer, COPD, depression, diabetes, ischemic heart disease, osteoporosis, and rheumatoid arthritis. Cross-referencing these indicators with claimed services helps identify misrepresented diagnoses and unjustified procedures.

4) Financial Features

Annual inpatient and outpatient reimbursement and deductible amounts provide quantifiable indicators of service cost. Unusually high reimbursement values relative to documented patient conditions serve as primary fraud indicators.

TABLE II. DATASET FEATURE SUMMARY

Feature	Description
BeneID	Unique beneficiary identifier
DOB / DOD	Birth and death dates; flags posthumous claims
Gender	Coded 0/1; detects gender-inconsistent procedures
Race	Categorical; supports population-level analysis
State / County	Geographic codes for regional fraud mapping
NoOfMonths_PartACov	Medicare Part A coverage duration
NoOfMonths_PartBCov	Medicare Part B coverage duration
ChronicCond_* (10)	Binary chronic condition flags (Alzheimer's, Diabetes…)
ChronicDiseaseIndex	Engineered feature: total chronic conditions per patient
RenalDiseaseIndicator	End-Stage Renal Disease flag (Y/N → 1/0)
IPAnnualReimbursementAmt	Total annual inpatient reimbursements
OPAnnualReimbursementAmt	Total annual outpatient reimbursements
IPAnnualDeductibleAmt	Annual inpatient deductible paid by patient
OPAnnualDeductibleAmt	Annual outpatient deductible paid by patient
Age (engineered)	Derived from DOB/DOD; flags age-inconsistent procedures
Alive (engineered)	Binary survival status derived from DOD

C. Engineered Features

Several new features were engineered to enhance predictive power. The ChronicDiseaseIndex represents the total count of chronic conditions per patient, providing a health complexity score. The Alive binary column explicitly distinguishes living from deceased patients, avoiding the complexity of working with null values in the DOD column. Patient Age was derived from DOB and DOD to enable age-inconsistency detection and demographic cohort analysis.

V. METHODS AND TECHNIQUES

A. Rule-Based Systems

Rule-based systems are among the earliest fraud detection tools in healthcare. They apply predefined rules to identify suspicious claims, such as billing for services after a patient's date of death, repeated claims for the same procedure within an unreasonably short time frame, or claims that exceed policy limits. While interpretable and easy to implement, these systems struggle with adaptability against novel fraud strategies and often generate high false-positive rates requiring continuous manual updating.

B. Supervised Learning

Supervised learning is used when labeled data is available—where past claims are tagged as fraudulent or legitimate. The model learns discriminating patterns and applies this knowledge to new, unseen data. This study evaluates three supervised classifiers:

1) Random Forest

An ensemble of decision trees where each tree is trained on a bootstrapped sample. Random Forest handles high-dimensional feature spaces without extensive preprocessing, provides built-in feature importance analysis, and is robust to overfitting through ensemble averaging. It is particularly effective for detecting complex fraud patterns while maintaining partial interpretability.

2) Support Vector Machine (SVM)

SVM constructs a hyperplane that maximally separates fraud from non-fraud instances. It is effective in high-dimensional spaces and capable of handling non-linear patterns through kernel functions. However, it is computationally expensive on large datasets and less flexible than ensemble methods when dealing with class imbalance.

3) K-Nearest Neighbors (KNN)

KNN predicts class labels based on the majority class among the k nearest neighbors in feature space. It offers an intuitive understanding of proximity-based fraud clustering but suffers from the curse of dimensionality in high-feature datasets and is sensitive to class imbalance without additional weighting.

C. Unsupervised Learning for Anomaly Detection

Unsupervised learning is critical when fraud labels are missing or unreliable—a common scenario in real-world healthcare data. Clustering algorithms such as K-Means and DBSCAN group claims and identify those that do not fit any cluster well. Isolation Forests efficiently isolate rare fraud cases by exploiting their low-density structure. Autoencoders—neural networks trained to reconstruct normal claims—identify fraud through large reconstruction errors.

D. Handling Class Imbalance

A fundamental challenge in fraud detection is class imbalance. In this study's dataset, fraudulent claims represent approximately 9% of all records against 91% legitimate claims. Without mitigation, models bias predictions toward the majority class, resulting in high overall accuracy but poor fraud recall. Cost-sensitive learning was implemented by assigning higher weights to the fraud class via the class_weight='balanced' parameter in the Random Forest classifier. This adjusts the learning objective so that the model penalizes misclassification of fraud more heavily without altering the data distribution.

VI. IMPLEMENTATION OF THE FRAUD DETECTION MODEL

A. Problem Formulation

The core problem is formulated as a binary classification task: given a healthcare claim record, classify it as fraudulent (1) or legitimate (0). The working hypothesis is that statistically learnable patterns exist within healthcare billing data—encompassing demographic, clinical, and financial variables—that differentiate fraudulent claims from genuine ones, and that machine learning models trained on these patterns can detect such anomalies with high accuracy.

B. Data Preprocessing

1) Data Cleaning

Several targeted cleaning operations were conducted to standardize the data and reduce noise. The RenalDiseaseIndicator column, which contained inconsistent string values ('Y' and '0'), was normalized by converting 'Y' to 1 and '0' to 0, then cast to integer type. DOB and DOD columns were converted to datetime objects. Missing DOD values—indicating living patients—were filled with the latest known DOD to enable consistent lifespan estimates. An Alive binary column was derived (1 = living, 0 = deceased).

2) Feature Selection and Transformation

Coverage duration features were grouped to identify outliers and segment patients into short-term versus long-term coverage groups. The ChronicDiseaseIndex was computed by summing binary chronic condition flags per patient. Categorical features—Gender, Race, RenalDiseaseIndicator, State, and County—were encoded using label encoding for tree-based models and one-hot encoding where appropriate. Redundant columns (raw DOD and BirthYear) were dropped after feature derivation to prevent data leakage.

C. Model Training and Validation

The dataset was split into training and testing sets using an 80:20 ratio. The training partition (80%) was used to fit each model and tune hyperparameters via internal validation folds. The test set (20%) was held out entirely and used only for final performance evaluation, providing an unbiased estimate of generalization capability. Feature scaling (standardization) was applied for KNN and SVM models given their sensitivity to feature magnitudes.

D. Performance Metrics

Given the class-imbalanced nature of the dataset, accuracy alone is misleading. This study evaluates models using: Precision—the fraction of predicted fraud cases that are genuinely fraudulent; Recall—the fraction of actual fraud cases correctly identified; F1-Score—the harmonic mean of precision and recall; and AUC-ROC—the area under the receiver operating characteristic curve, reflecting discrimination capability across all classification thresholds. High recall is especially critical in healthcare fraud detection, where missing a fraud case (false negative) is costlier than a false alarm.

VII. EXPERIMENTAL RESULTS AND ANALYSIS

A. Random Forest Progressive Evaluation

A series of Random Forest classifiers were trained with class_weight='balanced' to address class skew. Table III summarizes performance across all model configurations. Tree depth was progressively increased from 4 to 25, and estimator count from 500 to 1,000, to evaluate the impact on classification performance.

TABLE III. COMPARATIVE PERFORMANCE OF ALL MODELS (* = Best Model)

Model	Depth	Est.	Acc.%	Prec.	Recall	F1	AUC
RF – Model 1	4	500	64.70	0.558	0.354	0.433	0.66

RF – Model 2	10	500	72.54	0.663	0.569	0.612	0.77
RF – Model 3	15	500	78.13	0.735	0.666	0.699	0.85
RF – Model 4	25	500	86.80	0.864	0.776	0.818	0.94
RF – Model 5*	25	1000	86.88	0.865	0.777	0.819	0.94
SVM	—	—	62.98	0.577	0.108	0.181	—
KNN	—	—	56.40	0.413	0.341	0.373	—

** RF Model 5 (depth=25, n_estimators=1000) — Best performing configuration*

At depth 4 (Model 1), the Random Forest underfits the data, capturing only basic fraud signals (AUC = 0.66, F1 = 0.433). Increasing depth to 10 (Model 2) yields a significant improvement (AUC = 0.77, F1 = 0.612). Model 3 (depth 15) achieves AUC = 0.85 and F1 = 0.699, reflecting the model's enhanced capacity to learn complex interaction patterns. Model 4 (depth 25, 500 trees) represents a substantial leap to AUC = 0.94 and F1 = 0.818. Increasing estimators to 1,000 (Model 5) yields marginal further gains (F1 = 0.819), confirming diminishing returns beyond this configuration.

B. Best-Performing Model Analysis

Model 5—Random Forest with max_depth=25 and n_estimators=1,000—achieved the best overall performance: accuracy of 86.88%, precision of 0.865, recall of 0.777, F1-score of 0.819, and AUC-ROC of 0.94. The ROC curve for this model is significantly skewed toward the top-left corner, demonstrating exceptional sensitivity even at very low false-positive rates. An AUC of 0.94 implies the model correctly ranks a randomly chosen fraud case above a legitimate one 94% of the time, making it suitable for real-world deployment.

C. SVM Performance

The SVM classifier achieved accuracy of 62.98%, precision of 0.577, recall of 0.108, and F1-score of 0.181. While the accuracy figure appears moderate, it is misleading due to class imbalance. The critically low recall of 10.8% means the model detects only a small fraction of actual fraud cases. SVM's underperformance can be attributed to its sensitivity to class imbalance without adequate kernel tuning, and to its computational limitations on the 138,000+ row dataset relative to ensemble methods.

D. KNN Performance

KNN achieved accuracy of 56.4%, precision of 0.413, recall of 0.341, and F1-score of 0.373. The instance-based nature of KNN, combined with the curse of dimensionality in the 25+ dimensional feature space, rendered distance metrics unreliable. KNN's inherent bias toward the majority class in imbalanced datasets further suppressed its recall, making it unsuitable as a standalone fraud classifier.

VIII. CHALLENGES AND LIMITATIONS

A. Data Quality and Standardization

Healthcare data is collected from numerous sources—hospitals, insurers, pharmacies, and providers—each using different formats, terminologies, and data entry standards. Missing values, unstructured text fields, and ambiguous encoding (e.g., binary values represented as 'Y' or '1') hinder model training and interpretation. Without rigorous data cleaning, normalization, and mapping to standard ontologies (ICD-10, CPT codes), the accuracy of fraud detection systems is severely compromised.

B. Privacy and Ethical Concerns

Healthcare fraud detection systems rely on sensitive Protected Health Information (PHI) regulated by HIPAA in the U.S. and GDPR in Europe. Key ethical concerns include patient consent—using data beyond its original intent may violate ethical boundaries—and transparency—many machine learning models, especially deep learning, are black-box in nature, making it difficult to explain why a claim is flagged, raising concerns about due process and accountability for providers falsely accused.

C. Balancing False Positives and False Negatives

Managing the precision-recall trade-off is a critical operational challenge. Excessive false positives can overwhelm human investigators, delay reimbursement for honest providers, and erode trust in the system. Excessive false negatives allow fraud to persist undetected, costing billions annually. High-precision models are preferred when investigative resources are limited; high-recall models are essential when casting a wide net for manual review. Multi-stage screening systems, where low-confidence predictions are passed to human experts, offer a practical compromise.

IX. FUTURE DIRECTIONS AND RECOMMENDATIONS

A. Real-Time Integration

Most fraud detection models today operate in batch or retrospective modes, flagging suspicious activity after it has occurred. The future lies in embedding fraud detection directly into real-time healthcare systems—EHR platforms, claims processing systems, and pharmacy benefit managers—enabling flag-before-payment workflows, interruption of high-risk billing during submission, and real-time alerts for pattern-based anomalies in patient-provider interactions.

B. Cross-Institutional Data Sharing

A significant gap in fraud detection arises from data silos where each insurer, hospital, or government agency holds data in isolation. Federated learning—training shared models across institutions without centralizing sensitive data—offers a promising approach for consortium-based detection. National fraud registries and cross-institutional case studies could drastically improve fraud signal detection while preserving patient privacy.

C. Policy Recommendations

Current regulatory frameworks often lag behind emerging fraud tactics, particularly in telemedicine fraud, AI-generated synthetic patient identities, and cross-border digital claims. Regulators must update statutes to include digital forensics requirements, mandate traceability of health data, and require audit trails in electronic claims processing. Institutions that proactively detect and report fraud internally should receive policy protection and financial incentives to foster a culture of compliance.

X. CONCLUSION

A. Summary of Key Findings

This study explored the complex and evolving landscape of healthcare provider fraud and demonstrated how data-driven machine learning techniques can effectively detect fraudulent behavior. Using a structured Medicare claims dataset, three supervised classifiers—Random Forest, SVM, and KNN—were implemented and evaluated. The Random Forest classifier with max_depth=25 and n_estimators=1,000 emerged as the most effective model, achieving 86.88% accuracy, F1-score of 0.819, and AUC-ROC of 0.94. Key preprocessing steps—missing value imputation, categorical encoding, chronic disease indexing, and cost-sensitive learning—were critical to model performance.

The study also confirmed that ensemble methods substantially outperform traditional classifiers (SVM, KNN) on imbalanced healthcare fraud datasets. Feature engineering, particularly the derivation of patient survival status and chronic disease complexity scores, meaningfully enhanced detection capability.

B. Contributions

This project makes the following contributions to the field:

- A practical supervised learning framework for healthcare fraud detection, demonstrating real-world implementation using structured Medicare claims data.
- A reproducible feature engineering pipeline for healthcare datasets, including age derivation, chronic condition indexing, and survival status encoding.
- A comparative evaluation of Random Forest, SVM, and KNN models providing clarity on algorithm suitability for imbalanced fraud detection tasks.
- Policy-aware contextualization integrating regulatory frameworks, ethical considerations, and actionable recommendations for real-time deployment.

References.

REFERENCES

[1] U.S. Department of Justice, "Health Care Fraud and Abuse Control Program Annual Report for Fiscal Year 2022," U.S. Dept. of Health and Human Services, Washington, DC, 2023.

[2] Centers for Medicare & Medicaid Services (CMS), "Medicare Fraud & Abuse: Prevent, Detect, Report," CMS, Baltimore, MD, 2023.

[3] L. Bauder, T. A. Khoshgoftaar, and N. Seliya, "A survey on the state of the art in healthcare fraud investigation," Health Informatics J., vol. 26, no. 1, pp. 555–573, 2020.

[4] D. Herland, T. M. Khoshgoftaar, and R. Wald, "A review of data mining using big data in health informatics," J. Big Data, vol. 1, no. 2, pp. 1–35, 2014.

[5] Z. He, S. Zhang, J. Lv, J. Li, and Q. Chen, "Healthcare fraud detection: A survey and a clustering model incorporating geo-location information," in Proc. IEEE Int. Conf. Big Data, 2020, pp. 1474–1479.

[6] S. Liyanage, U. Karunasekera, C. Leckie, K. Doss, and M. Palaniswami, "Fraud detection in health insurance using data mining techniques," in Proc. IEEE Int. Conf. Information Science and Technology, 2019.

[7] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[8] V. Vapnik, The Nature of Statistical Learning Theory, 2nd ed. New York, NY: Springer-Verlag, 2000.

[9] J. R. Quinlan, "Induction of Decision Trees," Machine Learning, vol. 1, no. 1, pp. 81–106, 1986.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[11] U.S. Congress, "False Claims Act, 31 U.S.C. §§ 3729–3733," 1863, as amended.

[12] U.S. Congress, "Anti-Kickback Statute, 42 U.S.C. § 1320a-7b(b)," 1972, as amended.

[13] U.S. Congress, "Stark Law (Physician Self-Referral Law), 42 U.S.C. § 1395nn," 1989, as amended.

[14] U.S. Congress, "Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191," 1996.

[15] T. Fawcett, "An introduction to ROC analysis," Pattern Recognit. Lett., vol. 27, no. 8, pp. 861–874, 2006.

[16] A. Ayeni, "Healthcare Provider Fraud Detection Dataset," Kaggle, 2024. [Online]. Available: https://www.kaggle.com