A Machine Learning Methodology for Detecting Dishonest Billing by Healthcare Providers to Federal Healthcare Programs

Ayomide Ayeni
Department of Computer Science
Atlanta, GA, USA
ayeniayomide5@gmail.com

Abstract—Fraud in healthcare billing is a persistent threat to the fiscal sustainability of Medicare and Medicaid, with improper payments estimated at approximately $94 billion annually. This paper presents a complete, reproducible methodology for detecting potentially fraudulent billing patterns at the healthcare provider level using supervised machine learning. The approach encompasses end to end model development: structured data preprocessing, clinically informed feature engineering, comparative model training using Random Forest, Support Vector Machine, and K Nearest Neighbors classifiers, and rigorous evaluation on held out Medicare claims data. The best performing model, a Random Forest classifier configured with maximum depth 25 and 1,000 estimators, achieved 86.88% accuracy, an F1 score of 0.8187, and an area under the receiver operating characteristic curve of 0.94. Beyond model performance, the methodology is designed for institutional adoption: a published feature dictionary specifies every measurement and its computation; a staged validation protocol enables any programme integrity unit to test the method on its own claims data without external data transfers; and the reference implementation is released under an open license. This paper describes the technical foundations, experimental results, deployment considerations, and regulatory context of the complete detection framework.

Keywords—healthcare fraud detection, machine learning, random forest, provider level anomaly detection, Medicare, Medicaid, claims data analysis, imbalanced classification, feature engineering, open methodology

I. INTRODUCTION

A. Background and Motivation

Healthcare fraud is a pervasive challenge confronting public and private healthcare systems worldwide. It involves intentional deception or misrepresentation by healthcare providers, patients, or other entities to obtain unauthorized benefits from insurance programmes, government funded healthcare, or private payers. The complexity of healthcare billing systems, the sheer volume of transactions processed daily, and the intricate nature of medical coding create conditions in which fraudulent activities can persist undetected for extended periods.

Federal healthcare programmes, principally Medicare and Medicaid, represent the largest targets. The U.S. Department of Health and Human Services estimated approximately $94 billion in improper payments across federal healthcare programmes in recent reporting years [1]. Fraudulent schemes take many forms: phantom billing, where services are charged but never rendered; upcoding, where the billed procedure code reflects a higher cost intervention than the one actually performed; unbundling, where services that should be billed as a package are separated to inflate reimbursement; illegal kickback arrangements; falsification of medical records; and prescription fraud [2].

Existing detection infrastructure relies heavily on expert defined rules and manual audit processes. While these approaches provide interpretability and regulatory familiarity, they struggle to keep pace with increasingly sophisticated evasion strategies. A billing entity that ensures each individual claim is defensible in isolation can maintain an aggregate pattern that is implausible without any single claim triggering a rule. Machine learning offers a complementary capability: learning complex, nonlinear patterns from historical data that rules cannot anticipate [3], [4].

B. Scope and Objectives

This study presents a complete, reproducible methodology for detecting potentially fraudulent healthcare provider billing. The specific objectives are as follows.

First, to construct a clinically informed feature engineering pipeline that transforms raw Medicare claims data into measurements with defensible fraud risk rationales. Second, to train and comparatively evaluate three supervised classifiers, namely Random Forest, Support Vector Machine (SVM), and K Nearest Neighbors (KNN), on structured claims data. Third, to publish the full methodology, including the reference implementation, feature dictionary, and institutional validation protocol, under open access terms so that any qualified programme integrity unit can retrieve, inspect, reproduce, and test the approach without purchasing a licence or entering into a contract with the author. Fourth, to contextualise the work within the regulatory framework governing healthcare fraud enforcement, including the Health Insurance Portability and Accountability Act (HIPAA), the False Claims Act (FCA), and the Anti Kickback Statute (AKS).

C. Contributions

This paper makes four principal contributions to the field:

(1) A practical supervised learning framework for healthcare provider fraud detection, demonstrated on structured Medicare claims data with full reproducibility. (2) A documented feature engineering pipeline encompassing age derivation, chronic condition complexity indexing, survival status encoding, and coverage tenure segmentation, each with an explicit fraud risk rationale. (3) A comparative evaluation of Random Forest, SVM, and KNN classifiers that provides empirical guidance on algorithm suitability for imbalanced fraud detection tasks. (4) An open distribution model comprising a reference implementation, feature dictionary, and staged validation protocol, designed to enable institutional adoption without dependence on the author.

II. RELATED WORK

A. Traditional Fraud Detection Approaches

Rule based systems represent the earliest and most widely deployed approach to healthcare fraud detection. These systems apply predefined thresholds and business rules to flag claims that exceed policy limits, duplicate prior submissions, or carry procedure codes inconsistent with the patient's documented condition. The CMS Fraud Prevention System (FPS) uses predictive analytics to screen claims in real time, while Zone Programme Integrity Contractors (ZPICs) are tasked with investigating billing anomalies [1]. While interpretable and readily auditable, rule based systems require continuous manual updating and generate high false positive rates as fraud strategies evolve [3].

B. Machine Learning for Fraud Detection

Supervised learning methods have demonstrated substantial promise when labelled fraud data is available. Bauder and Khoshgoftaar [5] applied Random Forest with class imbalance mitigation on Medicare data and reported strong discrimination. He et al. [6] proposed a clustering model incorporating geolocation features. Liyanage et al. [7] explored data mining techniques for health insurance fraud. More recently, hierarchical attention networks and graph based methods have been applied to capture inter provider referral patterns [8]. However, a persistent gap in the literature is the absence of fully specified, openly published methodologies that enable institutional replication. Most published studies report model performance on proprietary or synthetic datasets without providing the feature definitions, validation procedures, or deployment protocols necessary for a programme integrity unit to test the approach on its own data. This study addresses that gap directly.

C. Legal and Regulatory Context

Healthcare fraud detection operates within a dense regulatory framework. The False Claims Act (31 U.S.C. §§ 3729–3733) prohibits the knowing submission of false claims for federal payment and has been instrumental in recovering billions in settlements [9]. The Anti Kickback Statute (42 U.S.C. § 1320a 7b(b)) criminalises referral incentives tied to federal programme business [10]. The Stark Law (42 U.S.C. § 1395nn) restricts physician self referrals for designated health services payable by Medicare or Medicaid [11]. HIPAA (Pub. L. 104 191) established the Health Care Fraud and Abuse Control Programme (HCFAC) and mandates that fraud detection systems respect patient confidentiality and data protection standards [12]. Any operational deployment of the methodology described in this paper must comply with these statutes and with the institution's own governance framework.

III. DATASET AND PREPROCESSING

A. Dataset Overview

This study utilises a structured dataset sourced from publicly available Medicare claims records that mirrors the schema of real world beneficiary and billing data. The dataset encompasses inpatient and outpatient billing records, patient demographics, chronic condition indicators, coverage duration, and reimbursement amounts. The complete dataset contains over 138,000 beneficiary level records with known fraud labels, providing a foundation for supervised classification [16].

The dataset contains a significant class imbalance: approximately 9% of records are labelled as fraudulent against 91% that are legitimate. This ratio reflects the operational reality of healthcare fraud and requires explicit mitigation strategies, discussed in Section IV.C.

B. Feature Categories

The raw feature space comprises 23 columns organised into four categories. Table I summarises the complete feature set.

TABLE I. DATASET FEATURE CATEGORIES

Category	Features	Count	Type
Identifiers and Demographics	BeneID, DOB, DOD, Gender, Race, State, County	7	Mixed
Coverage and Enrolment	NoOfMonths_PartACov, NoOfMonths_PartBCov	2	Integer
Clinical Indicators	ChronicCond_* (10 columns), RenalDiseaseIndicator	11	Binary/Coded
Financial	IPAnnualReimbursementAmt, IPAnnualDeductibleAmt, OPAnnualReimbursementAmt, OPAnnualDeductibleAmt	4	Currency

Patient demographic features include unique beneficiary identifiers, dates of birth and death, gender, race, and geographic codes for state and county. The date of death feature is particularly valuable for identifying posthumous claims, a well documented fraud signal. Coverage duration features indicate enrolment length under Medicare Parts A and B, where short coverage paired with high cost claims signals front loaded billing schemes. Ten binary columns indicate the presence of specific chronic conditions, enabling cross referencing of claimed services against documented diagnoses. Financial features capture annual inpatient and outpatient reimbursement and deductible amounts, which serve as primary quantitative fraud indicators.

C. Data Cleaning

Real world healthcare data contains inconsistent formats, missing entries, and text based categorical variables that must be transformed for machine learning compatibility. Several targeted cleaning operations were performed.

The RenalDiseaseIndicator column initially contained inconsistent string values: the string 'Y' mixed with the string '0'. The former was replaced with the integer 1 to signify end stage renal disease, and the latter was retained as 0. The column was then cast to integer type. This is the first cleaning step in the pipeline because the mixed encoding blocks all downstream numeric operations [13].

Dates of birth and death were converted to datetime objects. Missing date of death values, indicating living patients, were filled with the latest observed date of death to enable consistent age computation. This fill must occur after the derivation of the Alive binary feature (described in Section III.D), because reversing the order produces a constant, useless survival indicator and no element of the pipeline will raise an error [13].

Categorical features, including Gender, Race, RenalDiseaseIndicator, State, and County, were encoded using label encoding for tree based models. One hot encoding was applied where distance based or margin based models required binary feature splits. Importantly, the encoder was fit on training data only and applied via transform to held out data. Fitting independently on each split produces inconsistent code assignments and a silently corrupted evaluation. Redundant columns, specifically the raw date of death and derived BirthYear, were dropped after feature derivation to prevent data leakage.

D. Feature Engineering

Four features were engineered from the raw data to enhance predictive power. Each is documented in the companion feature dictionary [13] with its exact computation, missing value handling, and fraud risk rationale.

The Alive feature is a binary integer (1 = living, 0 = deceased) derived from the presence or absence of a recorded date of death: df['Alive'] = df['DOD'].isna().astype(int). This encodes the mortality signal as a clean binary, avoiding the complexity of null valued date columns and enabling the posthumous billing check. The participant facing form of this measurement is the guidance to continue reviewing a deceased relative's Medicare Summary Notices for at least one year, because the fraud scheme depends on nobody opening the mail.

The Age feature is computed in whole years: age at death for deceased beneficiaries and age at the extract reference date for living ones. It supports detection of age inconsistent procedures and provides a demographic reference for peer comparison, since expected cost varies strongly with age.

The ChronicDiseaseIndex collapses ten binary chronic condition flags into a single ordinal score representing health complexity. High reimbursement against a high index is expected; high reimbursement against an index of zero or one is the discrepancy worth investigating. The computation uses an explicit assertion on column count to prevent a prefix selector from silently matching nine or eleven columns, which would produce an index on a different scale.

The BirthYear feature was derived for exploratory analysis and then dropped before model training, as it carries no information beyond Age and risks being misread as a proxy for data extract vintage.

TABLE II. ENGINEERED FEATURES

ID	Feature	Type	Role	Status
E1	Alive	Binary	Model input	Implemented
E2	Age	Integer	Model input	Implemented
E3	ChronicDiseaseIndex	Integer (0 to 10)	Model input	Implemented
E4	BirthYear	Integer	Exploratory only	Dropped

IV. METHODOLOGY

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 measurable accuracy.

B. Classification Algorithms

Three supervised classifiers were selected to represent distinct algorithmic families, enabling a comparative analysis of approach suitability for this domain.

Random Forest (RF): An ensemble of decision trees where each tree is trained on a bootstrapped sample of the data. Random Forest handles high dimensional feature spaces without extensive preprocessing, provides built in feature importance analysis through Gini impurity or permutation based measures, and is robust to overfitting through ensemble averaging. It is particularly suited to detecting complex, nonlinear interaction patterns while maintaining partial interpretability [14]. In this study, tree depth was progressively increased from 4 to 25, and the number of estimators was varied between 500 and 1,000.

Support Vector Machine (SVM): SVM constructs a hyperplane that maximally separates the fraud class from the legitimate class in feature space. It is effective in high dimensional spaces and capable of handling nonlinear decision boundaries through kernel functions. However, SVM is computationally expensive on large datasets and less flexible than ensemble methods when dealing with severe class imbalance [15].

K Nearest Neighbors (KNN): KNN predicts the class label of a record based on the majority class among the k nearest neighbours in feature space. It provides an intuitive proximity based classification but is sensitive to the curse of dimensionality in high feature datasets and to class imbalance without explicit distance weighting [6].

C. Class Imbalance Mitigation

In this dataset, fraudulent claims represent approximately 9% of all records. Without mitigation, classifiers bias predictions toward the majority class, producing high overall accuracy but critically poor recall on the minority fraud class. Cost sensitive learning was implemented by assigning higher misclassification penalties to the fraud class via the class_weight='balanced'

parameter in the Random Forest classifier. This adjusts the learning objective so that the model penalises failure to detect fraud more heavily, without altering the underlying data distribution [5].

D. Training and Validation Protocol

The dataset was split into training and testing partitions using an 80:20 ratio. The training partition (80%) was used to fit each model and tune hyperparameters through internal cross validation folds. The held out test set (20%) was used exclusively for final performance evaluation, providing an unbiased estimate of generalisation capability. Feature standardisation was applied for KNN and SVM models given their sensitivity to feature magnitudes; tree based models do not require scaling.

E. Evaluation Metrics

Given the class imbalanced nature of the dataset, accuracy alone is insufficient and potentially misleading. This study evaluates models using four complementary metrics. Precision measures the fraction of predicted fraud cases that are genuinely fraudulent, directly relevant to the investigative workload a programme integrity unit would bear. Recall measures the fraction of actual fraud cases correctly identified, reflecting the system's coverage. F1 Score is the harmonic mean of precision and recall, balancing both concerns. The Area Under the Receiver Operating Characteristic Curve (AUC ROC) reflects discrimination capability across all classification thresholds, with a value of 1.0 indicating perfect separation and 0.5 indicating random performance. In healthcare fraud detection, high recall is especially critical because a missed fraud case (false negative) represents ongoing financial loss, while a false positive generates investigative cost but not direct harm [17].

V. 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. Tree depth was progressively increased from 4 to 25, and the number of estimators was varied to evaluate the impact on classification performance. Table III presents the complete results across all model configurations.

TABLE III. COMPARATIVE PERFORMANCE OF ALL MODELS

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	N/A	N/A	62.98	0.577	0.108	0.181	N/A
KNN	N/A	N/A	56.40	0.413	0.341	0.373	N/A

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

B. Progressive Depth Analysis

At depth 4 (Model 1), the Random Forest underfits the data, capturing only the most basic fraud signals and achieving an AUC of 0.66 and an F1 of 0.433. Increasing depth to 10 (Model 2) yields a significant improvement, with AUC rising to 0.77 and F1 to 0.612, as the model gains capacity to represent conditional patterns such as the interaction between coverage duration and reimbursement amounts. Model 3 (depth 15) reaches an AUC of 0.85 and F1 of 0.699, reflecting the model's enhanced ability to learn complex multi way interaction patterns across clinical, demographic, and financial features.

Model 4 (depth 25, 500 trees) represents a substantial performance leap, achieving an AUC of 0.94 and F1 of 0.818. This configuration is sufficient to capture the full feature interaction space without overfitting, as evidenced by the stable AUC. Increasing the number of estimators to 1,000 in Model 5 yields only marginal further gains (F1 = 0.819), confirming diminishing returns and establishing this as the optimal configuration.

C. Best Performing Model Analysis

Model 5, a Random Forest with max_depth=25 and n_estimators=1,000 configured with class_weight='balanced', achieved the best overall performance: accuracy of 86.88%, precision of 0.865, recall of 0.777, F1 score of 0.8187, and AUC

ROC of 0.94. An AUC of 0.94 means that the model correctly ranks a randomly chosen fraud case above a randomly chosen legitimate case 94% of the time, which indicates strong discrimination suitability for deployment as a prioritisation instrument.

Feature importance analysis revealed that the financial features (IPAnnualReimbursementAmt and OPAnnualReimbursementAmt) carry the most discriminative weight, followed by the ChronicDiseaseIndex and coverage tenure features. This is consistent with domain expectations: reimbursement amounts out of line with the beneficiary's documented condition profile represent the primary quantitative fraud signal, and the interaction between cost and clinical complexity is where the model's signal concentrates.

D. SVM and KNN Performance

The SVM classifier achieved accuracy of 62.98%, precision of 0.577, recall of 0.108, and F1 of 0.181. While the accuracy figure appears moderate, it is misleading under class imbalance. The critically low recall of 10.8% means the model detects barely one in ten actual fraud cases. SVM's underperformance on this task is attributable to its sensitivity to class imbalance without adequate kernel optimisation, and to computational limitations on the 138,000+ row dataset relative to ensemble methods.

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

VI. PROVIDER LEVEL AGGREGATION DESIGN

The unit of analysis in the completed beneficiary level work described above is the individual claim record. The design objective of the broader endeavour extends this to the billing entity, the provider. This distinction is consequential: a rule evaluating one claim at a time cannot observe that a provider's overall billing pattern is implausible when every individual claim is defensible in isolation. Aggregating to the provider level makes that aggregate pattern visible.

The feature dictionary [13] specifies ten provider level aggregation features (F1 through F10) that are defined but not yet implemented in the reference implementation. These include: claim volume (prov_claim_count), beneficiary breadth (prov_distinct_benes), billing intensity (prov_claims_per_bene), cost profile (prov_mean_reimb), peer deviation score (prov_reimb_peer_z), posthumous billing rate (prov_posthumous_rate), front loading rate (prov_short_tenure_rate), clinical implausibility rate (prov_cond_mismatch_rate), and cross provider beneficiary overlap (prov_bene_overlap).

Each is stated in full in the feature dictionary so that the specification is fixed before the code exists, rather than reverse engineered from it afterwards. This is intentional transparency: the dictionary records that these measurements are specified but not implemented, and any statement elsewhere that this detection work computes provider level aggregates today would be inaccurate.

VII. INSTITUTIONAL VALIDATION FRAMEWORK

A detection method that cannot be reproduced on another institution's data is a result, not a capability. The companion validation protocol [18] defines a 12 stage, 8 gate self test that a programme integrity unit runs on its own claims data, inside its own environment, without contacting the author and without sending data anywhere.

The protocol is organised into four phases. Phase 1 (Stages 0 through 2) establishes whether the test can be run at all: authority and governance, extract assembly, and data integrity verification. Phase 2 (Stages 3 through 6) tests whether the methodology reproduces on the institution's data: feature construction reproducibility, baseline measurement including the institution's existing rule set, model training, and measured performance against baselines. Phase 3 (Stages 7 through 9) determines whether the method is worth an investigator's time: operating point selection based on investigative capacity, precision and lift measurement at the top of the working queue, and a fairness audit confirming that scoring does not concentrate on a demographic group without cause. Phase 4 (Stages 10 through 12) is field validation: blind investigator review of flagged entities against controls, a documented adoption decision, and ongoing monitoring and revalidation.

Stage 10, the blind investigator review, dominates the schedule at six to twelve weeks, and this is the honest shape of the work. Stages 1 through 9 can be completed in approximately two weeks by one analyst. At the end of Stage 9, an institution will know whether the method is worth a pilot. It will not know whether it works, because nothing before Stage 10 involves an investigator opening a case.

A model that clears every gate has demonstrated that it ranks billing entities in a way that puts more genuine problems near the top of an investigator's queue than the institution's current method. That is the entire claim. It is not a finding that any flagged entity committed fraud, it is not admissible as evidence, and it does not license contacting an entity on the strength of a score. The output is a prioritisation instrument and should be described that way.

VIII. DISCUSSION

A. Key Findings

The experimental results confirm that ensemble tree based classifiers substantially outperform traditional classifiers (SVM, KNN) on imbalanced healthcare fraud datasets. The Random Forest's capacity to model complex feature interactions, combined with cost sensitive learning, produces strong discrimination (AUC = 0.94) at operationally acceptable precision levels. Feature engineering, particularly the derivation of the ChronicDiseaseIndex and the Alive survival indicator, meaningfully enhanced detection capability by providing the model with clinically grounded measurements rather than raw data fields.

B. Operational Considerations

Managing the precision recall tradeoff is a central operational challenge. Excessive false positives can overwhelm human investigators, delay reimbursement for honest providers, and erode trust in the detection system. Excessive false negatives allow fraud to persist undetected. The validation protocol addresses this directly by requiring the institution to set its operating point based on available investigative capacity before examining model scores, and by measuring precision and lift only at the top of the queue the institution would actually work.

The Race feature presents a specific fairness concern. Although present in the source data, it carries no causal mechanism by which it would indicate fraudulent billing behaviour. However, it can correlate with a fraud label if enforcement history was itself unevenly distributed, in which case the model learns the enforcement pattern rather than the fraud pattern. The feature dictionary recommends excluding Race from any deployed scoring model and using it only as a diagnostic fairness audit: computing score distributions across race groups to verify that the model is not reproducing historical enforcement bias [13].

C. Limitations

This study has several acknowledged limitations. First, the dataset is synthetic, and while it mirrors the schema of real world Medicare claims data, it does not contain the missingness patterns, coding drift, and provider behaviour present in production environments. The validation protocol exists precisely to bridge this gap. Second, the current implementation operates at the beneficiary level; the provider level aggregation layer (Section VI) is specified but not yet implemented. Third, the chronic condition column coding ambiguity described in the feature dictionary represents a silent inversion risk that must be resolved against each institution's own extract before deployment.

D. Privacy and Ethical Considerations

Healthcare fraud detection systems necessarily process sensitive Protected Health Information (PHI) regulated by HIPAA. The validation protocol addresses this by design: all processing occurs inside the institution's environment, no data leaves at any stage, beneficiary and provider identifiers are pseudonymised before analysis, and the minimum necessary standard is applied to the extract specification. The protocol's no contact commitment during validation (Stages 1 through 10) ensures that no entity is referred, audited, or contacted on the basis of a model score during the testing phase.

IX. FUTURE DIRECTIONS

Several extensions of this work are planned or underway. First, the provider level aggregation layer specified in Section F of the feature dictionary will be implemented, enabling the model to flag billing entities rather than individual claim records. Second, cross institutional validation through federated learning approaches would enable consortium based detection without centralising sensitive data. Third, temporal modelling using sequential claims data could capture emerging fraud patterns that static snapshots miss. Fourth, real time integration with electronic health record systems and claims processing pipelines would enable flag before payment workflows, shifting from retrospective detection to prospective prevention. Fifth, the methodology will be extended to incorporate unsupervised anomaly detection techniques, including Isolation Forests and autoencoders, to complement the supervised approach and detect novel fraud patterns for which no labelled examples exist.

X. CONCLUSION

This paper presents a complete, reproducible machine learning methodology for detecting healthcare provider fraud in federal healthcare programmes. Using structured Medicare claims data, three supervised classifiers were trained and comparatively evaluated. The Random Forest classifier with maximum depth 25 and 1,000 estimators achieved 86.88% accuracy, F1 score of 0.8187, and AUC ROC of 0.94, substantially outperforming SVM and KNN on this imbalanced task.

The methodology is designed for adoption, not merely for publication. The reference implementation, feature dictionary, and institutional validation protocol are published under open access terms. The validation protocol defines a staged process that enables any qualified programme integrity unit to test the method on its own claims data, inside its own environment, and to reach a documented adoption decision based on field results rather than published numbers. The feature dictionary specifies every measurement, its exact computation, its fraud risk rationale, and its known ambiguities. Together, these artefacts transform a research result into an institutional capability.

The broader contribution of this work is a demonstration that rigorous, openly published, and independently testable fraud detection methodologies can be developed and distributed at no cost to the institutions that need them most. Federal and state healthcare programme integrity units serve a critical function in protecting public funds and patient welfare. Providing these units with tools they can inspect, reproduce, and validate on their own data, rather than tools they must purchase or accept on trust, is a direct contribution to that mission.

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 and Medicaid Services (CMS), "Medicare Fraud and 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] R. A. Bauder and T. M. Khoshgoftaar, "Medicare fraud detection using random forest with class imbalanced big data," in Proc. IEEE Int. Conf. Big Data, 2019.

[6] 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.

[7] 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.

[8] Y. Yang, J. Xie, Z. Wang, and M. Lin, "Medical insurance fraud detection via hierarchical attention networks with multi head self attention," in Proc. ACM CIKM, 2022.

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

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

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

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

[13] A. Ayeni, "Feature Dictionary: Provider Level Fraud Detection for Federal Healthcare Programmes," Open methodology artefact, published May 30, 2026. Available: www.ayomideayeni.com

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

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

[16] A. Ayeni, "Healthcare Provider Fraud Detection: Reference Implementation," Open source artefact, published May 30, 2026. Available: www.ayomideayeni.com

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

[18] A. Ayeni, "Validation Protocol: Provider Level Fraud Detection for Federal Healthcare Programmes," Open methodology artefact, published May 30, 2026. Available: www.ayomideayeni.com