IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

International Journal of Innovative Research and Scientific Studies
ONLINE ISSN : 2617-6548

Investigating the impact of feature engineering and machine learning model selection for real-world fraud detection systems in healthcare insurance claims

Mohamed F. Abouelenein[1*], Hatem M. Noaman[2], Gaber Sallam Salem Abdalla Al Salmany[3]

[1]Department of Insurance and Risk Management, College of Business, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Riyadh, Saudi Arabia.
[2]Department of Computer Science, Faculty of Computer Science and Artificial Intelligence, Beni-Suef University, Beni-Suef 62511, Egypt.
[3]Department of Mathematics and Insurance, Faculty of Commerce, Beni-Suef University, Beni-Suef 62511, Egypt.

Corresponding author: Mohamed F. Abouelenein (*Email: mabouelenein@imamu.edu.sa*)

Abstract

The objective of this study is to create and assess an extensive machine learning framework for identifying healthcare fraud in the National Health Insurance Scheme (NHIS) claims, targeting the significant financial losses and degradation of patient care resulting from fraudulent practices. This work examined 20,388 NHIS medical claim data exhibiting phantom billing, incorrect diagnoses, and ghost enrollee fraud trends. A systematic feature engineering approach increased 8 initial characteristics to 27 engineered features, encompassing temporal patterns, financial abnormalities, medical classifications, and indicators of patient behavior. Six machine learning algorithms were assessed: Random Forest, Logistic Regression, Gradient Boosting, XGBoost, Support Vector Machine, and Neural Network, utilizing extensive performance criteria such as accuracy, AUC, calibration quality, and demographic fairness analysis. Gradient Boosting attained the highest test AUC of 0.9213 with an accuracy of 80.11%, whilst XGBoost exhibited superior computational efficiency (0.71 seconds training time) alongside competitive performance (AUC: 0.9187, accuracy: 80.48%). Financial variables predominantly influenced fraud detection judgments, with daily billing rates (AMOUNT_PER_DAY: 0.55) and total billed amounts (0.36) contributing to 91% of model predictions. Significant calibration difficulties were detected across models, with minor demographic bias noted. Ensemble tree-based algorithms routinely surpass alternative approaches in the identification of healthcare fraud. Nevertheless, the primary dependence on financial attributes can cause vulnerabilities to sophisticated fraud schemes that keep accurate billing amounts while capitalizing on weaknesses in medical coding. This research offers healthcare administrators actionable insights for the implementation of real-time fraud detection systems, emphasizing the necessity of balancing detection accuracy with computational efficiency and the enhancement of medical coding analysis capabilities.

Keywords: Feature engineering, Healthcare insurance claims, Machine learning Model selection, Real-world fraud detection.

DOI: 10.53894/ijirss.v8i5.8858
Funding: This study received no specific financial support.
History: Received: 6 June 2025 / **Revised:** 10 July 2025 / **Accepted:** 14 July 2025 / **Published:** 25 July 2025

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Publisher: Innovative Research Publishing

1. Introduction

1.1. Background

Healthcare fraud is a significant and growing risk to global healthcare systems, leading to substantial financial burdens and compromising the integrity of healthcare services and the quality of patient care. Estimates indicate that 3% to 10% of total healthcare expenditure is wasted due to fraudulent activities [1]. In the United States, healthcare insurance fraud, especially Medicare and Medicaid fraud, is the most financially costly form of insurance fraud, with legal healthcare fraud settlements and judgments under the False Claims Act exceeding $1.32 trillion annually, representing 30% of total healthcare spending [2]. These significant financial losses result in elevated healthcare expenses for patients, increased insurance rates, and reduced access to vital medical treatments, especially for those most vulnerable who rely on public health insurance systems [3].

Healthcare fraud schemes have become increasingly complex and sophisticated, involving various fraudulent activities that exploit weaknesses in healthcare systems [3]. Common types of fraud include billing for services not provided, utilizing real patient information often obtained through identity theft, charging for unnecessary medical services, misrepresenting diagnoses to justify more expensive treatments, and unbundling or "exploding" charges [3]. Fraudulent activities within the healthcare ecosystem are carried out by various actors, including healthcare providers, patients, and health insurance companies, each exploiting different system aspects for financial gain [4]. Healthcare fraud presents significant challenges due to the intentional deception employed by healthcare providers who perform unnecessary treatments or services for greater revenue or payments. Additionally, fraudulent actions by patients or groups of individuals pretending to be ill to exploit healthcare funds further complicate this problem [4].

1.2. Problem Statement

The detection and prevention of healthcare fraud involve numerous challenges that distinguish this field from other fraud detection applications. Detecting healthcare fraud presents significant challenges, including a small number of fraudulent cases, inconsistent data, a lack of data standardization and integration, privacy issues, and a limited availability of labeled fraudulent cases for model training [1]. The issue of healthcare fraud presents a fundamental imbalance, characterized by a limited number of identified fraudulent providers compared to a larger population of non-fraudulent providers. This disparity leads to bias in machine learning algorithms, which typically favor the majority class, ultimately resulting in suboptimal classification performance for the fraudulent class [5]. Healthcare data exhibit high dimensionality and mixed data types, including categorical variables that necessitate specialized handling. Furthermore, the complexity of medical coding systems complicates pattern recognition for conventional rule-based detection systems [5].

Research on NHIS implementations has identified patterns of fraud, such as the repetition of NHIS registration numbers, overbilling for medications, discrepancies in drug prescriptions, excessive prescribing of treatments, and duplication of client records [6]. Challenges to the financial sustainability of national health insurance schemes include fraud and corruption, misuse of gatekeeper systems, and broad benefits packages that encourage exploitation [7].

The use of machine learning in healthcare fraud detection has evolved from basic rule-based systems to advanced machine learning methods that can identify complex patterns and relationships in vast claims data [1]. Recent studies highlight the significance of ensemble machine learning methods and interpretability in fraud detection models, enabling healthcare administrators to better understand the influence of individual features on predictions through approaches like partial dependence plots, SHAP, and LIME [8]. The effectiveness of these approaches relies significantly on the quality and comprehensiveness of feature engineering processes that convert raw healthcare claims data into meaningful predictive indicators[5]. Feature engineering is a vital aspect of developing effective healthcare fraud detection systems, as it directly impacts the ability of machine learning algorithms to recognize fraudulent patterns [5]. In healthcare, effective feature engineering should identify temporal patterns, including anomalies in service duration, financial irregularities such as billing rate outliers, medical classification patterns through diagnostic code analysis, and patient behavior indicators such as encounter frequency and repeat visits [9]. The systematic development of engineered features enables machine learning algorithms to identify subtle indicators of fraud that may not be evident through conventional rule-based methods [1].

1.3. Research Gap and Objectives

This research focuses on addressing significant gaps by creating and assessing a comprehensive machine learning framework for detecting healthcare fraud, utilizing real-world claims data from a National Health Insurance Scheme [10].

This research enhances prior studies in several key areas: first, it systematically compares six machine learning algorithms (Random Forest, Logistic Regression, Gradient Boosting, XGBoost, Support Vector Machine, and Neural Network) for healthcare fraud detection using a dataset of 20,388 medical claim records; second, it suggests a feature engineering methodology that expands 8 original features into 27, capturing temporal patterns, financial anomalies, medical classifications, and patient behavior indicators; third, it provides a detailed analysis of model performance, including accuracy, AUC, calibration quality, and fairness across demographic groups; and fourth, it offers practical insights for healthcare administrators on the deployment and operational considerations of machine learning-based fraud detection systems.

This paper is organized as follows: The Materials and Methods section outlines the characteristics of the dataset, the feature engineering methodology, and the machine learning approaches utilized in this evaluation. The Results and Discussion section details the evaluation results, including performance metrics, confusion matrix analysis, misclassification patterns, model calibration analysis, and feature importance findings. The Conclusion section synthesizes key findings, discusses limitations, and offers recommendations for future research and practical implementation of machine learning-based healthcare fraud detection systems.

2. Literature Review

2.1. Supervised Learning Approaches in Healthcare Fraud Detection

2.1.1. Tree-based Methods

Tree-based approaches have achieved extensive utilization in healthcare fraud detection due to their balance of efficacy and interpretability. Random Forest, which generates numerous decision trees and averages their predictions, demonstrates superior performance on healthcare fraud datasets while maintaining acceptable interpretability [8]. The method's tolerance to overfitting and capacity to manage missing values render it especially appropriate for the unique features of healthcare data. Gradient boosting approaches exhibit remarkable efficacy in fraud detection, with XGBoost distinguished for its efficiency and regularization strategies that construct trees consecutively to mitigate overfitting while maintaining high performance [11].

LightGBM enhances computational efficiency through its leaf-wise growth technique and Exclusive Feature Bundling, making it particularly effective for datasets characterized by numerous sparse features, as often encountered in healthcare billing data [8]. CatBoost utilizes ordered boosting and inherently incorporates categorical features, minimizing preprocessing requirements and reducing the likelihood of target leakage, which is especially advantageous in healthcare contexts where categorical variables are common [8]. Support Vector Machines have been utilized in healthcare fraud detection due to their ability to manage high-dimensional data and non-linear correlations via kernel approaches; nonetheless, they encounter scaling issues with extensive datasets [12].

Several studies have proven the efficacy of Random Forest algorithms in fraud detection. Varmedja et al. [9] determined that RF algorithms had outstanding results in accuracy, recall, and precision metrics. Likewise, Severino and Peng [13] showed that RF models demonstrated enhanced performance in terms of accuracy metrics when employed on Brazilian real-world insurance claims data. The ensemble learning method of Random Forest, which generates several decision trees and outputs the mode of classes, has demonstrated notable efficacy in mitigating the overfitting biases common to decision tree models.

2.1.2. Linear and Statistical Methods

Logistic regression has been extensively employed as a predictive modelling method for fraud detection. Sheffali and Deepa [14] attained 92% accuracy in their fraud detection classification study utilizing logistic regression. Sumalatha and Prabha [15] devised a model utilizing logistic regression that successfully identified fraud in claims with an accuracy of 83.35%. The technique's ability to examine correlations between categorical dependent and independent variables makes it particularly suitable for binary fraud/no-fraud classifications.

2.1.3. Neural Network Approaches

Neural networks and deep learning methodologies demonstrated effectiveness in healthcare fraud detection, as evidenced by Nabrawi and Alanazi [16] who illustrated the application of artificial neural networks for the identification of fraudulent activity in healthcare insurance claims. Recent comparison studies indicate that ensemble methods frequently surpass single neural network approaches, especially when interpretability is essential [8]. The complex nature of neural networks often results in "black box" models that are challenging to fully understand, which presents difficulties in healthcare applications where decision transparency is crucial for regulatory compliance and stakeholder confidence. Deep learning methodologies employing artificial neural networks have demonstrated promising results in the detection of healthcare fraud. Shamitha and Ilango [17] created a real-time artificial neural network model that attained 85.3% accuracy, 97% precision, and 73% recall. The multilayer perceptron (MLP) methodology employed in artificial neural networks (ANNs) facilitates the differentiation, processing, and examination of datasets in manners unachievable by traditional systems.

2.2. Unsupervised Learning and Anomaly Detection

Unsupervised learning methods have been utilized in healthcare fraud detection, especially for uncovering previously unrecognized fraud tendencies. Clustering techniques, such as K-means clustering, have been employed to discern atypical patterns in billing behavior, whereas DBSCAN is effective for identifying outliers in healthcare transaction data [12].

Hierarchical clustering has been utilized to categorize analogous fraud patterns, facilitating the detection of fraud schemes that may remain obscured using supervised learning methods. Anomaly detection methodologies have gained importance in healthcare fraud detection owing to their capacity to identify uncommon patterns without requiring tagged instances of fraud. One-class SVM discovers outliers in standard billing patterns, whereas Isolation Forest effectively identifies abnormalities in high-dimensional healthcare data. Autoencoders, like neural network methodologies, have been utilized for identifying anomalous patterns by learning to recreate typical billing behavior and signaling instances that cannot be precisely reconstructed [18].

2.3. Graph-Based and Network Analysis Approaches

Graph-based approaches to healthcare fraud detection have received considerable attention in recent years, especially for identifying collective fraud schemes and examining relationships among healthcare entities. Work by Nabrawi and Alanazi [16] employed convolutional neural networks and graph convolutional networks to model and integrate networks of patient-doctor relationships for enhanced fraud inference. This method acknowledges that behavior during medical encounters inherently mirrors the relationships between patients and physicians, which can be modeled using graph structures.

Zhou et al. [19] examined collective medical fraud cases by developing visitation networks to illustrate spatiotemporal relationships among patients. They utilized the Louvain community detection algorithm, grounded in modularity optimization, to identify potentially suspicious groups. Work by Yoo et al. [20] had developed graph structures and employed Graph Neural Network algorithms for model training, while also incorporating traditional machine learning techniques to extract relationships from datasets and generate bipartite graphs as input features.

The dynamic nature of medical data poses challenges for graph-based methodologies, necessitating adaptable graph structures to effectively represent the evolving relationships within healthcare systems. Dynamic networks have yet to see extensive application in the fraud detection industry, indicating a potential area for future research development [8]. Utilizing graph theory facilitates the detection of fraud patterns that are often obscured in conventional feature-based methods, especially in scenarios involving coordinated fraudulent actions among various entities.

3. Materials and Methods

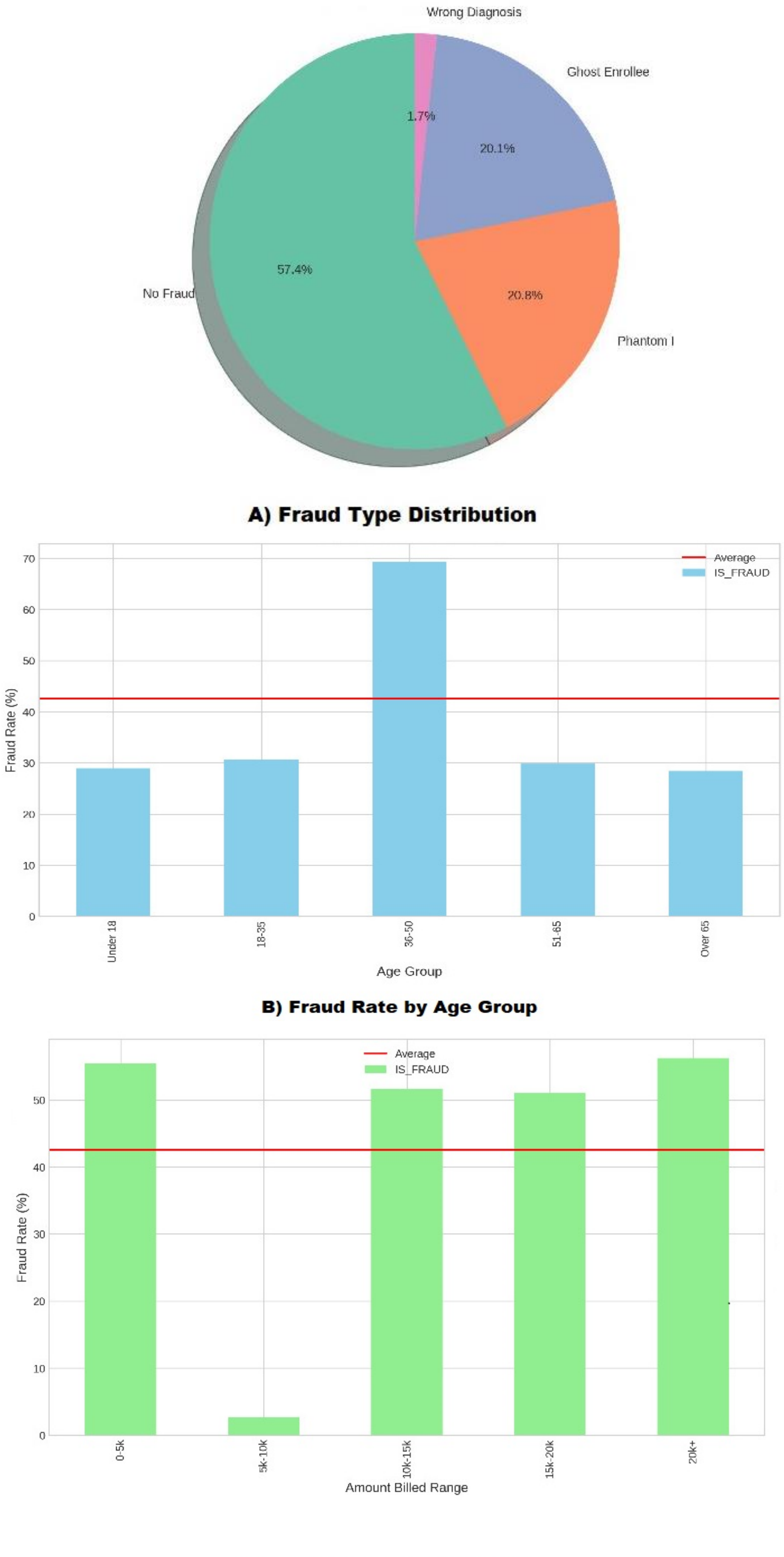
3.1. Dataset

This paper used synthetic fraud patterns in conjunction with a comprehensive healthcare fraud detection dataset gathered from the National Health Insurance Scheme (NHIS) [10]. The dataset comprises 20,388 medical claim records with eight key attributes that capture essential information about patient interactions, billing procedures, and healthcare fraud. It includes patient data such as gender classifications, age demographics ranging from young children to the elderly, and unique patient identifiers. Encounter dates, which mark the beginning of patient-provider interactions, and discharge dates, indicating the conclusion of services, provide temporal features that trace the entire patient journey. The total amount paid for services reflects financial data and is the primary target of deceptive inflation techniques. According to standard medical coding guidelines, medical data also includes diagnosis codes and related descriptions associated with each claim. Features in the dataset are explained in Table 1, along with details regarding their data types, value ranges, and relevance to fraud detection patterns.

Table 1.
Dataset Feature Descriptions.

Feature Name	Data Type	Description	Value Range/Examples
Patient ID	Integer	Unique identifier for each patient in the healthcare system	1-20,388
AGE	Float	Patient's age in years at time of encounter	0.0-120.0
GENDER	String	Patient's gender classification	M, F
DATE OF ENCOUNTER	Date	Timestamp of initial patient-provider interaction	YYYY-MM-DD format
DATE OF DISCHARGE	Date	Timestamp of patient discharge or service completion	YYYY-MM-DD format
Amount Billed	Float	Total monetary amount claimed for healthcare services	$0.00-$50,000+
DIAGNOSIS	String	Medical diagnosis codes and descriptions	ICD codes, medical terms
FRAUD_TYPE	String	Target variable indicating fraud classification	No Fraud, Phantom Billing, Wrong Diagnosis, Ghost Enrollee

The target variable is a list of classes for types of fraud, including "No Fraud" for legitimate healthcare claims, "Phantom Billing" for inflated charges or unnecessary services, "Wrong Diagnosis" for deliberate misclassification of medical conditions to justify higher reimbursements, and "Ghost Enrollee" for claims submitted for non-existent or deceased patients.



A) Fraud Type Distribution

B) Fraud Rate by Age Group

C) Fraud Rate by Amount Range

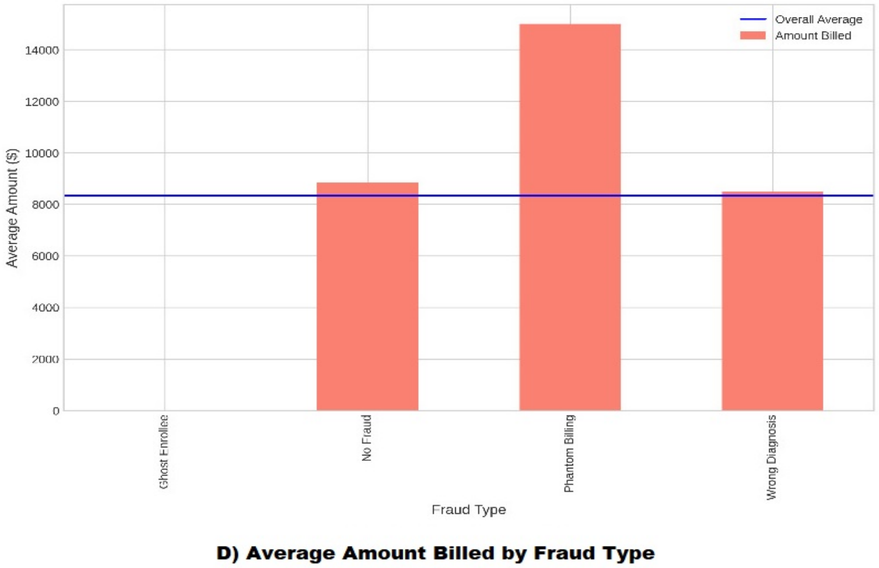


Figure 1.
Exploratory Data Analysis of Healthcare Fraud Patterns in NHIS Dataset. (A) Fraud Type Distribution. (B) Fraud Rate by Age Group. (C) Fraud Rate by Amount Range. (D) Average Amount Billed by Fraud Type.

Figure 1 presents a comprehensive exploratory analysis of fraud patterns in the dataset, providing crucial insights and characteristics of different fraud types. As shown by Figure 1 (A), "No Fraud" cases make up the majority, at 57.6% of all records. "Phantom Billing" accounts for 20.9% of cases, making it the most common type of fraud, followed by "Ghost Enrollee" at 20.1%. "Wrong Diagnosis" is the smallest fraud category, at 1.7%. This indicates that phantom billing and ghost enrollee fraud are the main issues in the medical system in question.

Figure 1 (B) shows that there are clear patterns of fraud by age group. The 30-45 age group has the highest fraud rate, at about 70%, which is much higher than the overall average fraud rate, shown by the horizontal red line at about 42%. The 19–30 and 46–65 age groups have moderate fraud rates of about 30%, while the youngest (under 18) and oldest (over 65) groups have the lowest fraud rates of about 28%. This suggests that middle-aged adults are more likely to be targeted or involved in fraudulent healthcare activities. Figure 1 (C) shows the fraud rate by amount range. It indicates that the fraud rate remains high across most billing amount ranges, with rates around 50–55% for amounts over $4,000. The lowest amount range (0–5K) exhibits a different pattern, with a significantly higher percentage of legitimate claims. This suggests that smaller billing amounts are less likely to be fraudulent. Conversely, the consistent fraud rates across higher amount ranges imply that fraudsters operate across a spectrum of costs rather than targeting specific high-value claims. Figure 1 (D) illustrates that the average billed amount by fraud type analysis reveals different billing patterns for various types of fraud. "Phantom Billing" has the highest average amount, approximately $15,000, which is substantially higher than the overall average of about $8,000 shown by the horizontal blue line. The average amounts for "Wrong Diagnosis" and "No Fraud" cases are similar, around $8,000 to $8,500. This indicates that wrong diagnosis fraud does not always involve inflated billing amounts but may involve inappropriate coding for services actually provided. Meanwhile, "Ghost Enrollee" fraud has moderate average billing amounts, suggesting that this type of fraud may focus more on volume than on high-value individual claims. Collectively, these patterns suggest that phantom billing is the most common and most valuable type of fraud.

3.2. Feature Engineering and Data Preprocessing

To identify patterns linked to fraudulent activity, the feature engineering methodology attached a significant amount of focus on temporal feature extraction. By computing the difference between the discharge and encounter dates, length of stay calculations were obtained, offering information on anomalies in service duration. Off-hours service patterns, weekend admission patterns that could be signs of fraud, binary flags for impossible negative length of stay scenarios, and seasonal fraud pattern detection using encounter month analysis were all part of the temporal anomaly detection process. Financial feature engineering focused on cost analysis variables such as per-day cost anomaly identification, statistical identification of outliers using two standard deviations from the mean, and daily billing rate computations.

Frequency of patient encounters and binary indicators for repeat visits that could indicate possible fraud patterns were included in the analysis of patient behavior patterns. To identify unusual medical patterns, medical classification processing included frequency encoding for rare diagnoses, numeric diagnostic code extraction, and primary diagnosis classification extraction. Also, demographic patterns can be linked to fraudulent activities; demographic classification generated age-based risk categorization groups spanning 0–18, 19–30, 31–45, 46–65, and 65+ years, along with corresponding frequency encoding.

The preprocessing pipeline included categorical encoding techniques, such as label encoding reserved for ordinal variables, frequency encoding for high-cardinality categorical features, and one-hot encoding for nominal categorical variables.

Numerical preprocessing used median imputation to handle missing values, z-score normalization to standardize, and statistical techniques to detect outliers using standard deviation and interquartile range. The full set of engineered features developed during the feature engineering and data preprocessing stage is shown in Table 2, which illustrates how raw information was converted into useful fraud detection markers.

Table 2.
Engineered Feature Descriptions

Feature Name	Data Type	Description	Calculation Method
LENGTH_OF_STAY	Integer	Duration of patient stay in days	DATE_OF_DISCHARGE - DATE_OF_ENCOUNTER
NEGATIVE_STAY	Binary	Flag for impossible negative length of stay	1 if LENGTH_OF_STAY ≤ 0, else 0
ENCOUNTER_HOUR	Integer	Hour of day when the encounter occurred	Extracted from DATE_OF_ENCOUNTER (0-23)
ENCOUNTER_DAY_OF_WEEK	Integer	Day of week for encounter	Extracted from DATE_OF_ENCOUNTER (0-6)
ENCOUNTER_MONTH	Integer	The month when the encounter occurred	Extracted from DATE_OF_ENCOUNTER (1-12)
DISCHARGE_HOUR	Integer	Hour of day when discharge occurred	Extracted from DATE_OF_DISCHARGE (0-23)
DISCHARGE_DAY_OF_WEEK	Integer	Day of week for discharge	Extracted from DATE_OF_DISCHARGE (0-6)
IS_WEEKEND_ENCOUNTER	Binary	Flag for weekend encounters	1 if ENCOUNTER_DAY_OF_WEEK in [5,6], else 0
IS_NIGHT_ENCOUNTER	Binary	Flag for nighttime encounters	1 if ENCOUNTER_HOUR < 6 or > 22, else 0
IS_WEEKEND_DISCHARGE	Binary	Flag for weekend discharges	1 if DISCHARGE_DAY_OF_WEEK in [5,6], else 0
AMOUNT_PER_DAY	Float	Daily billing rate	Amount Billed / max(1, LENGTH_OF_STAY)
HIGH_AMOUNT	Binary	Flag for billing amount outliers	1 if Amount Billed > $\mu \pm 2\sigma$, else 0
HIGH_AMOUNT_PER_DAY	Binary	Flag for daily rate outliers	1 if AMOUNT_PER_DAY > $\mu \pm 2\sigma$, else 0
DIAGNOSIS_CATEGORY	String	Primary diagnosis classification	Extracted the first alphabetical part from DIAGNOSIS
DIAGNOSIS_CODE	Integer	Numeric diagnostic code	Extracted numeric part from DIAGNOSIS
AGE_GROUP	Categorical	Age-based risk categories	Binned AGE into [0-18, 19-30, 31-45, 46-65, 65+]
PATIENT_VISIT_COUNT	Integer	Number of visits per patient	Count of records per Patient ID
MULTIPLE_VISITS	Binary	Flag for patients with multiple visits	1 if PATIENT_VISIT_COUNT > 1, else 0
DIAGNOSIS_CATEGORY_FREQ	Float	Frequency of diagnosis category	Proportion of records with the same DIAGNOSIS_CATEGORY
AGE_GROUP_FREQ	Float	Frequency of age group	Proportion of records in the same AGE_GROUP

3.3. Machine Learning Approaches for Fraud Detection

Figure 1 shows a detailed flowchart illustrating the systematic approach utilized for healthcare fraud detection through machine learning approaches. The approach starts with the NHIS dataset, containing 20,388 healthcare records and 8 original features, which undergoes thorough data preprocessing, including data cleaning and missing value handling, to guarantee data quality and consistency. The methodology subsequently advances to feature engineering, in which 19 supplementary features are systematically derived from the original data, resulting in a total of 27 features that encapsulate temporal patterns, financial anomalies, medical classifications, and indicators of patient behavior relevant to fraud detection. The augmented dataset is then divided through divided sampling into training (80%), validation (10%), and test

(10%) sets to preserve the distribution of fraud types across all segments. Six distinct machine learning algorithms are trained: Random Forest (RF), Logistic Regression (LR), Gradient Boosting (GB), XGBoost (XGB), Support Vector Machine (SVM), and Neural Network (NN), each providing unique advantages for pattern detection in healthcare fraud applications. After model training, extensive evaluation is performed utilizing several performance measures to assess the effectiveness of each approach. The methodology contains a decision point for selecting the optimal model based on evaluation results; if no definitive winner is identified, the process involves a model comparison phase utilizing statistical testing and ensemble analysis before revisiting the selection decision. Following successful model selection, the end result is an improved fraud detection method prepared for implementation in healthcare environments to identify phantom billing, erroneous diagnoses, and ghost enrollee fraud behaviors.

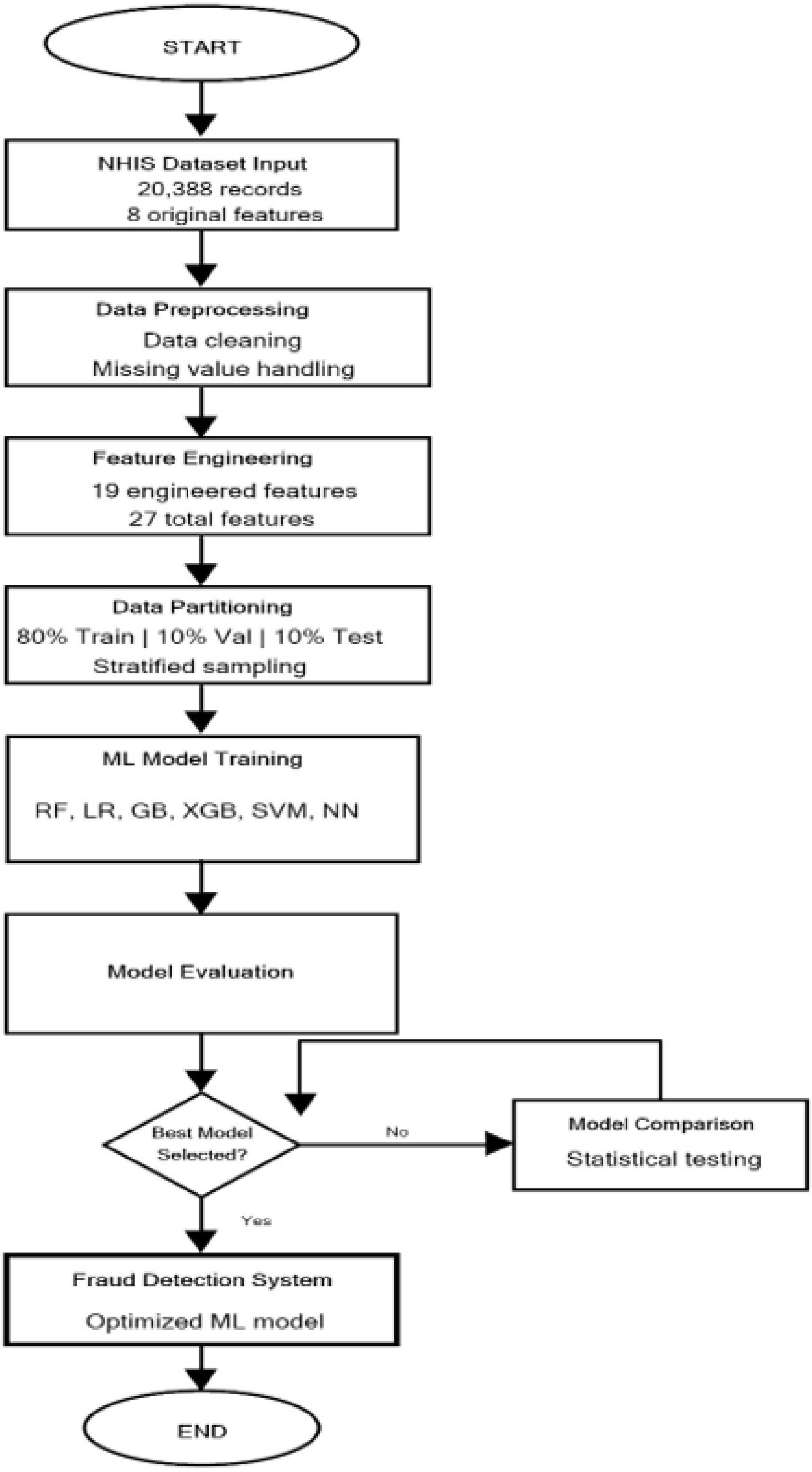


Figure 2.
Flowchart of the proposed healthcare fraud detection methodology.

We applied six different machine learning methods to classify healthcare fraud. Each approach used a different algorithm to capture various types of fraudulent behavior in healthcare claims data. We chose these methods because we needed to examine a wide range of machine learning techniques, from simple linear models that are easy to interpret to more complex methods. This approach allowed us to cover all aspects of the fraud detection problem space while also addressing the unique challenges associated with healthcare data, such as mixed data types, class imbalance, and the necessity for models that can be understood in clinical settings.

The first approach used was Random Forest Ensemble Learning, which employed bootstrap aggregating to develop robust fraud detection models by combining several decision trees trained on different parts of the data. This ensemble method is highly effective for identifying healthcare fraud because it can handle various data types, such as categorical diagnosis codes and continuous billing amounts, without requiring extensive preprocessing. It also tends not to overfit, making it suitable for complex healthcare datasets where relationships between features and fraud labels may not be linear or straightforward. The Random Forest algorithm's feature importance mechanism assists healthcare administrators in understanding the most significant indicators of fraud by highlighting which patient traits, time patterns, and billing behaviors are most likely associated with fraudulent activity. This enables targeted strategies to prevent fraud. The bootstrap sampling method ensures that each tree views the data from a unique perspective, increasing diversity within the ensemble and enhancing its generalization capabilities. The majority voting method further ensures that errors from individual trees have a limited impact.

Logistic Regression Linear Classifier was the baseline method, providing healthcare professionals with clear decision boundaries that are easy to understand, verify, and incorporate into their current clinical workflows. This linear approach transforms the problem of detecting fraud into a weighted combination of input features. It then employs the sigmoid function to estimate probabilities, indicating how likely each healthcare claim is to be fraudulent. The primary advantage of this method is its interpretability. The coefficients reveal how each feature influences the likelihood of fraud, enabling healthcare administrators to identify specific risk factors such as unusual billing patterns, suspicious timing, demographic anomalies, or a combination of these. Logistic regression is particularly useful for real-time fraud detection systems because it can process claims rapidly without delaying patient care. Its probabilistic output provides confidence estimates that assist in deciding which claims require manual review and which can be automatically approved. The third method discussed was Gradient Boosting Sequential Learning. This technique constructs models iteratively, with each new learner aiming to correct the errors of previous ones by focusing on misclassified cases. This makes it especially effective at uncovering complex fraud patterns that are difficult to detect with single-pass algorithms. Gradient boosting is valuable for identifying healthcare fraud because it can detect intricate relationships between features, such as combinations of patient age, billing amounts, time patterns, and diagnosis codes that collectively indicate a higher risk of fraud, even if individual indicators are weak or ambiguous. The algorithm is well-suited for healthcare datasets because it can handle various data types and missing values without extensive preprocessing. This flexibility is crucial given the variability in data quality across different providers, systems, and collection methods. The Support Vector Machine Geometric Approach incorporates kernel-based pattern recognition, transforming healthcare claims data into higher-dimensional spaces where fraud patterns are more discernible and separable. SVM seeks optimal decision boundaries that maximize the distinction between fraudulent and legitimate cases. This enables accurate classification even when fraud patterns are complex, overlapping, or resemble legitimate healthcare activities. The kernel technique allows SVM to identify non-linear relationships among features without requiring explicit high-dimensional transformations, making it suitable for healthcare fraud detection scenarios involving numerous variables such as patient demographics, provider characteristics, temporal patterns, and billing behaviors.

4. Results and Discussion

Table 3 presents a comprehensive evaluation of machine learning models for healthcare fraud detection, which reveals several critical insights about model performance, efficiency, and practical deployment considerations. Results indicate that Gradient Boosting is the most effective model, achieving the highest test AUC of 0.9213 and a competitive test accuracy of 80.11%, establishing it as the most dependable predictor for identifying fraudulent and real healthcare claims. XGBoost achieved a test AUC of 0.9187 and the greatest test accuracy of 80.48%, indicating that ensemble tree-based methods routinely surpass other algorithmic techniques in this field. The Random Forest model attained a commendable test AUC of 0.9148 but showed significant overfitting, achieving a flawless training accuracy of 99.99% in contrast to its test performance, suggesting inadequate generalization ability. Random Forest's excellent training accuracy (99.99%) against test performance (79.77%) is a typical example of extreme overfitting, in which the model memorized training patterns instead of learning generalizable fraud indications.

Table 3.
Model Performance Comparison.

Model	Train Accuracy	Validation Accuracy	Test Accuracy	Train AUC	Validation AUC	Test AUC	PR-AUC (Test)	Training Time (s)
Gradient Boosting	0.8258	0.7907	0.8011	0.9331	0.9106	0.9213	0.8953	15.26
XGBoost	0.8489	0.7939	0.8048	0.9461	0.9111	0.9187	0.8988	0.71
Random Forest	0.9999	0.7920	0.7977	1.0000	0.9111	0.9148	0.8954	4.06
Neural Network	0.8448	0.7789	0.7940	0.9460	0.8984	0.9091	0.8853	18.29
SVM	0.8048	0.7810	0.7916	0.9079	0.8844	0.8952	0.8759	156.39
Logistic Regression	0.7756	0.7452	0.7560	0.8727	0.8357	0.8483	0.8278	19.51

Concerning training efficiency and scalability, XGBoost demonstrated outstanding computational efficiency with a training time of 0.71 seconds while maintaining top-tier performance, making it suitable for real-time fraud detection systems that require frequent model updates. Random Forest provided an acceptable balance, with a training duration of 4.06 seconds; however, Gradient Boosting required 15.26 seconds to achieve better results. The Neural Network showed competitive prediction ability (AUC 0.9091) but took 18.29 seconds to train, indicating a less favorable efficiency profile. Support Vector Machine was the least practical, with a costly training time of 156.39 seconds despite respectable performance (AUC 0.8952), making it unsuitable for operational deployment.

4.1. Confusion Matrix Analysis

The confusion matrices, shown in Figure 3, demonstrate varying classification behaviors among the machine learning models, highlighting notable differences in their approaches for balancing sensitivity and specificity in healthcare fraud detection. Logistic Regression showed the most conservative approach to fraud detection, accurately identifying 2,040 legitimate cases (the highest true negatives) and producing only 301 false positives. This suggests a strategy focused on minimizing false alarms, albeit at the expense of overlooking some fraudulent cases (694 false negatives). This conservative approach yielded high precision at the expense of recall, rendering it appropriate for situations where the repercussions of false fraud accusations are significant.

XGBoost and Neural Network demonstrated enhanced fraud detection capabilities, with XGBoost identifying 1,377 true fraud cases while effectively managing false positives at 436 cases. The Neural Network demonstrated the highest sensitivity, identifying 1,445 fraud cases correctly; however, this resulted in 548 false positives, indicating the most flexible classification threshold among the models evaluated. This approach enhances fraud detection but elevates the operational burden associated with investigating false alarms.

Gradient Boosting and Random Forest exhibited comparable performance metrics, with Gradient Boosting accurately classifying 2,015 legitimate cases and 1,252 fraud cases, whereas Random Forest recorded similar results with 1,989 legitimate cases and 1,264 fraud cases. Both models exhibited moderate false positive rates (326 and 352), indicating their capacity to balance fraud detection with false alarm management effectively. The Support Vector Machine demonstrated balanced performance across both classes, yielding 2,020 true negatives and 1,208 true positives, alongside 321 false positives and 529 false negatives. Nonetheless, its computational inefficiency renders this balanced performance less appealing for practical application, despite its adequate classification attributes. The analysis indicates that XGBoost attained an optimal balance for effective fraud detection systems, accurately identifying 1,377 fraudulent cases (79.3% recall) while maintaining acceptable false positive rates (18.6% of predicted fraud cases were false alarms). The performance profile suggests that XGBoost effectively identifies a significant proportion of fraud cases while maintaining manageable investigation costs, positioning it as the most operationally viable solution among the assessed models. The differing false positive rates among models, which range from 292 to 548, highlight the necessity of selecting models in accordance with an organization's tolerance for false alarms and the resources allocated for fraud investigation.

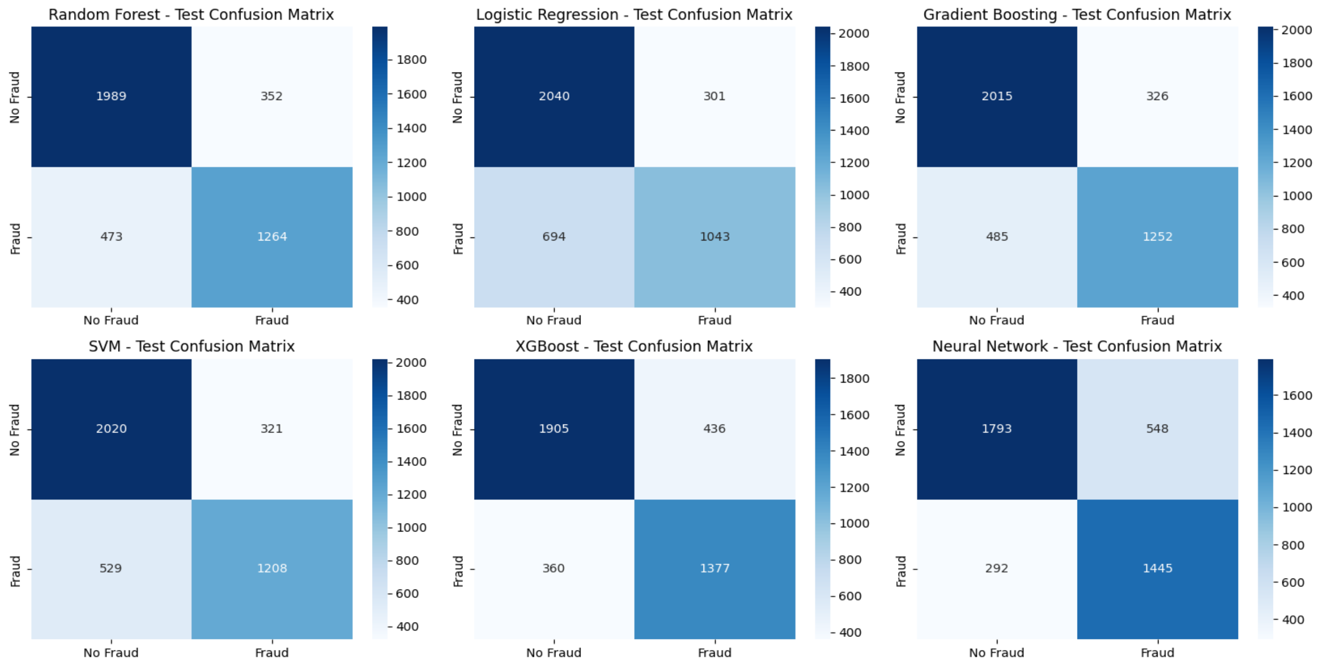


Figure 3.
Test Set Confusion Matrices for Machine Learning Models in Healthcare Fraud Detection.

4.2. Fraud Detection Misclassification Analysis

The analysis of misclassification distribution indicates a distinct pattern of agreement and disagreement among models in healthcare fraud detection, with most cases exhibiting moderate classification difficulties instead of universal consensus or total disagreement. The distribution shown by Figure 4 indicates that 393 instances were misclassified by three models, marking the highest classification difficulty. This is followed by 372 cases misclassified by one model and 333 cases misclassified by two models. This pattern indicates that the majority of challenging cases reside in a moderate difficulty range, where approximately half of the models encounter difficulties in classification, reflecting an intrinsic ambiguity in the fraud patterns associated with these cases. Only 85 cases were misclassified by all six models, indicating the most challenging instances for all machine learning approaches. In contrast, 214 cases were misclassified by five models, suggesting a subset of cases with subtle fraud indicators that only the most advanced models can accurately detect.

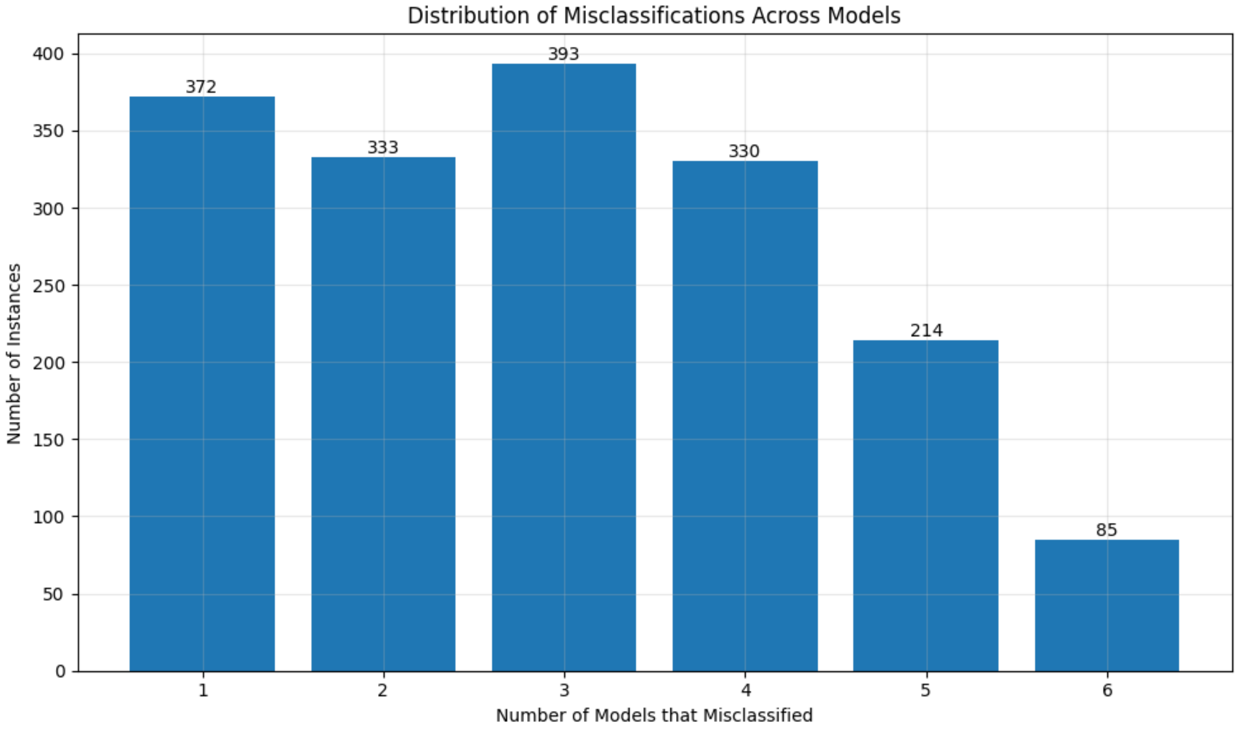


Figure 4.
Distribution of Misclassification Patterns Across Machine Learning Models in Healthcare Fraud Detection

Figure 5 illustrates the misclassification patterns for demographic analysis, demonstrating a notably uniform distribution across both gender and age groups. This suggests that the machine learning models do not display significant demographic bias in their fraud detection capabilities. Gender-based analysis reveals comparable average misclassification counts of approximately 3.0 for both males (n=831) and females (n=896), indicating that the classification challenges of the models are not systematically associated with patient gender. The analysis of age groups reveals minor differences in misclassification rates. Younger patients (0-18 years, n=279) and young adults (19-30 years, n=459) exhibit slightly elevated average misclassification counts of around 3.2-3.3. In contrast, middle-aged and older patients demonstrate progressively lower misclassification rates, with the 46-65 age group (n=480) recording the lowest rate at approximately 2.6. These differences are modest and may indicate the complexity of healthcare patterns in pediatric and young adult populations rather than systematic bias. This suggests that the models exhibit reasonable fairness across demographic groups while potentially facing challenges with age-related healthcare utilization patterns that are more variable in younger populations.

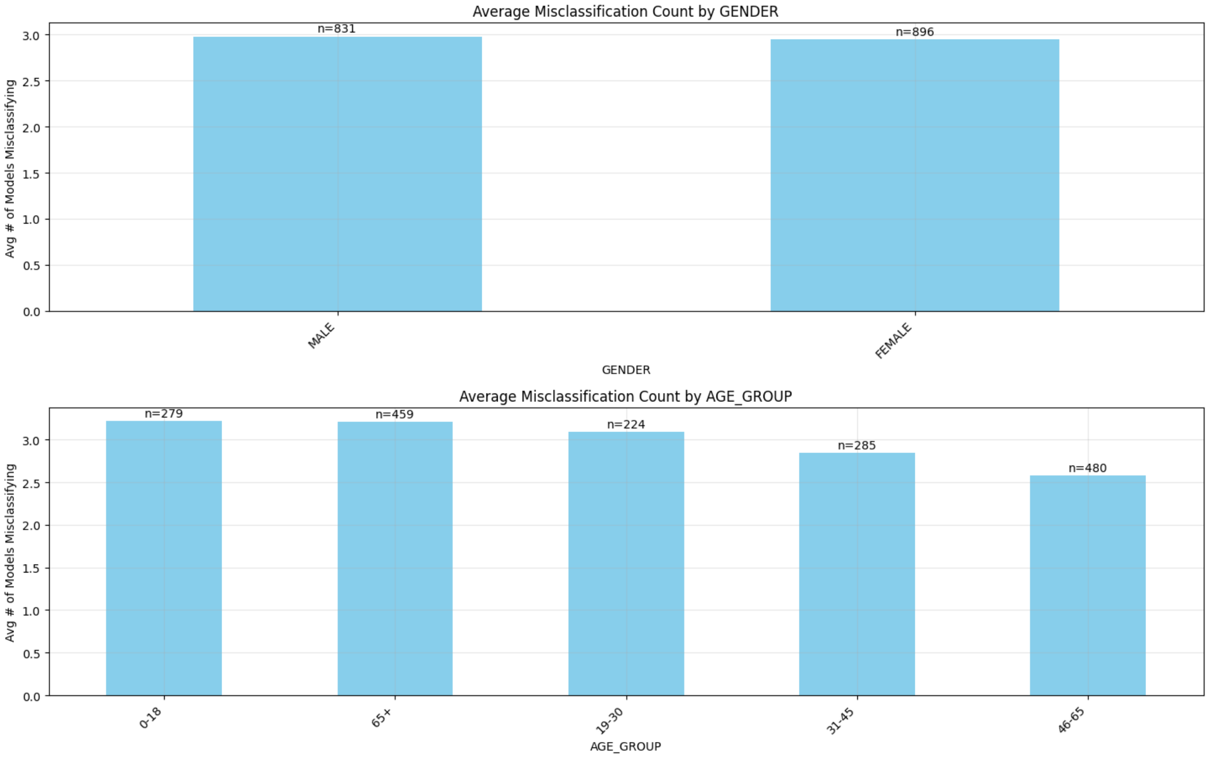


Figure 5.
Average Misclassification Count by Demographic Groups.

4.3. Model Calibration Analysis

The calibration curves indicate notable variations in the reliability of probability estimates among the six machine learning models, which have critical implications for operational fraud detection systems dependent on prediction confidence levels. Gradient Boosting exhibits significant calibration issues, including underconfidence in the low to moderate probability ranges (0.0-0.4), where predicted probabilities significantly underestimate actual fraud rates. This is followed by noticeable overcorrection, achieving near-perfect calibration only at very high predicted probabilities (>0.8). This pattern indicates that the probability estimates from Gradient Boosting are not reliable for establishing operational thresholds, as an instance predicted with a 40% fraud probability demonstrates actual fraud rates close to 50%. Random Forest and Logistic Regression show consistent calibration patterns, with Random Forest demonstrating moderate underconfidence across various probability ranges while maintaining a relatively smooth trajectory toward the ideal calibration line. Logistic Regression shows improved calibration within the middle ranges (0.2-0.6) but still displays patterns of overconfidence at moderate probabilities. XGBoost and SVM demonstrate moderate calibration quality, suffering from balanced deviations from ideal calibration. In contrast, the Neural Network displays inconsistent calibration behavior, characterized by overconfidence in specific probability ranges and underconfidence in others. Calibration differences are essential for practical deployment, as inadequately calibrated models may lead to misplaced confidence in fraud predictions. This can result in either excessive false alarms or overlooked fraud cases when decision thresholds rely on predicted probabilities instead of optimized classification boundaries.

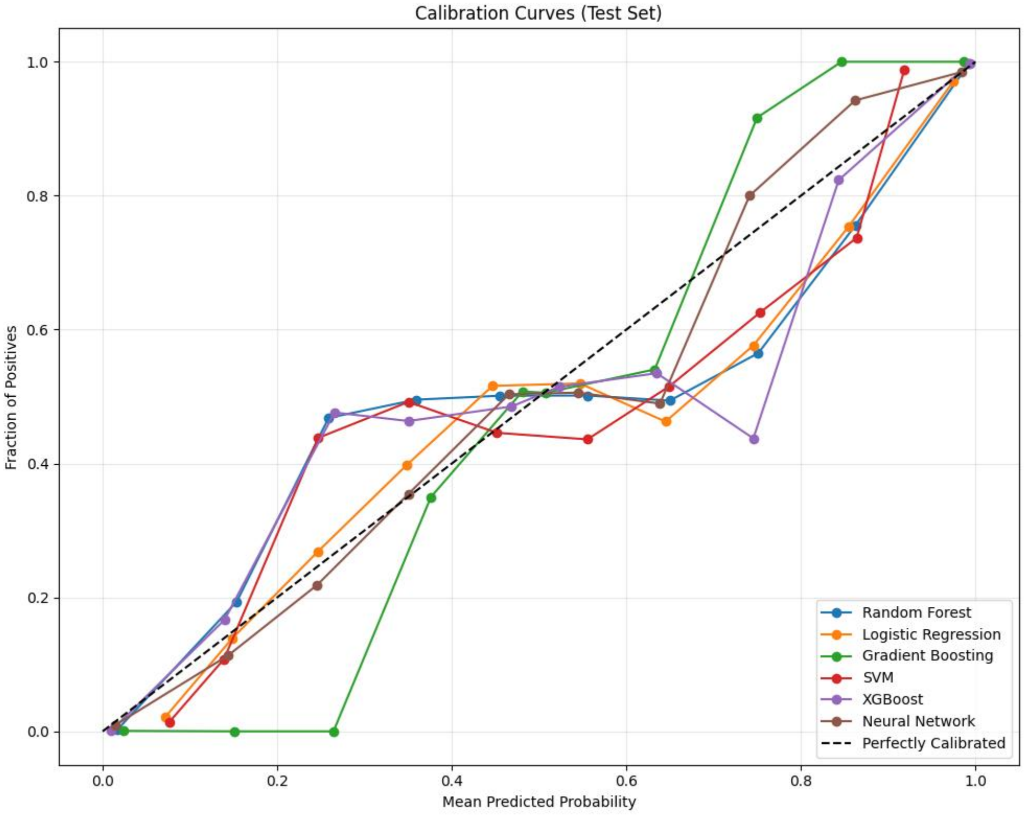


Figure 6.
Calibration Curves Comparing Predicted vs. Actual Fraud Probabilities on Test Set. The dashed line represents perfect calibration, where predicted probabilities match observed fraud rates.

4.4. Gradient Boosting Model Feature Importance Analysis

The feature importance analysis from the Gradient Boosting model, as shown in Figure 7, indicates a significant imbalance in predictive power, with financial variables predominantly influencing the fraud detection process, overshadowing medical and demographic factors. AMOUNT_PER_DAY is identified as the most significant feature, with an importance score of 0.55, while Amount Billed follows at 0.36. Together, these two features contribute to approximately 91% of the model's decision-making process. The intense focus on billing-related features indicates that the existing fraud detection methodology predominantly recognizes instances of atypical financial patterns, rather than addressing advanced medical coding fraud or intricate healthcare scheme manipulation. The third most significant feature, DIAGNOSIS_CATEGORY_CYESIS (pregnancy-related diagnoses), contributes merely 0.033 to the model's decisions, indicating a substantial decline that underscores the limited impact of medical diagnostic information. The 17 features remaining in the top 20 list exhibit negligible individual importance scores, all below 0.02. This suggests that demographic factors such as GENDER_MALE (0.016) and AGE (0.002), along with various diagnostic categories, have minimal influence within the current fraud detection framework. This pattern highlights serious concerns regarding the model's capacity to identify complex fraud schemes that exploit diagnostic codes, procedure classifications, or patient eligibility while preserving plausible billing amounts. The significant dependence on financial metrics may render the system susceptible to fraudsters who understand these patterns and can devise schemes that evade amount-based detection algorithms while taking advantage of deficiencies in medical coding oversight.

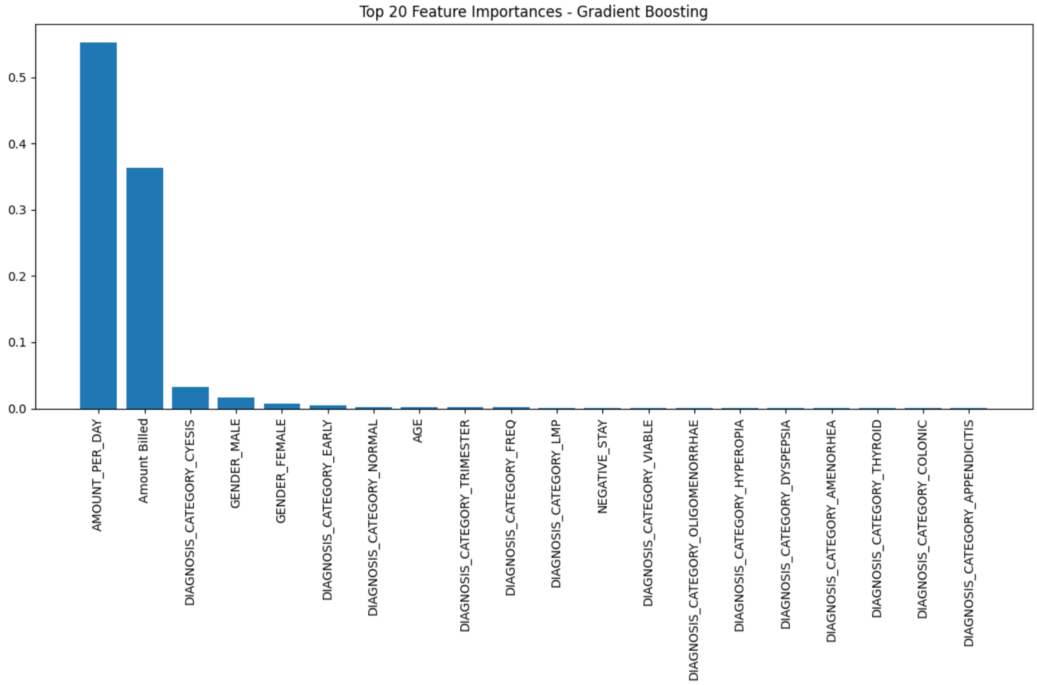


Figure 7.
Top 20 Feature Importance Scores from Gradient Boosting Model for Healthcare Fraud Detection.

6. Conclusion

This study demonstrates the effectiveness of machine learning approaches in healthcare fraud detection, revealing insights about model performance, efficiency, and deployment considerations. The assessment of six different machine learning algorithms on the NHIS dataset, which comprises 20,388 healthcare records, offers significant insights for healthcare administrators aiming to establish automated fraud detection systems.

The study demonstrates that ensemble tree-based methods, specifically Gradient Boosting and XGBoost, consistently outperformed other approaches in the context of healthcare fraud detection. Gradient Boosting demonstrated the highest test AUC of 0.9213, establishing it as the most reliable predictor. In contrast, XGBoost provided an optimal balance of performance and computational efficiency, making it particularly suitable for real-time operational deployment, as evidenced by its exceptional training time of 0.71 seconds.

The study identifies notable limitations in existing fraud detection methodologies. The significant prevalence of financial features in model decision-making processes, specifically AMOUNT_PER_DAY and Amount Billed, accounting for 91% of predictions, suggests that current methodologies mainly detect anomalous billing patterns rather than advanced medical coding fraud or intricate healthcare scheme manipulation. This financial-focused strategy may render healthcare systems vulnerable to fraudsters who recognize these patterns and can develop schemes that maintain reliable billing amounts while taking advantage of shortages in medical coding supervision.

The calibration analysis revealed significant reliability issues in probability estimates across all models, with Gradient Boosting demonstrating notable underconfidence in low-to-moderate probability ranges. The calibration issues present significant challenges for operational deployment, as poorly calibrated models may create dropped confidence in fraud predictions, leading to either an increase in false alarms or the failure to detect actual fraud when decision thresholds depend on predicted probabilities.

Despite these limitations, the research indicates that machine learning models display acceptable fairness among demographic groups, characterized by minimal gender bias and only slight age-related variations in misclassification patterns. The misclassification analysis indicates that the most challenging cases are mainly situated within moderate difficulty ranges, implying a fundamental ambiguity in fraud patterns rather than systematic failures of the model.

References

[1] A. Du Preez, S. Bhattacharya, P. Beling, and E. Bowen, "Fraud detection in healthcare claims using machine learning: A systematic review," *Artificial Intelligence in Medicine,* vol. 160, p. 103061, 2025.

[2] W. Fraud, R and i. H. Abuse, "R fraud, waste, and abuse in healthcare," 2024. https://www.goinvo.com/vision/fraud-waste-abuse-in-healthcare/

[3] National Healthcare Anti-Fraud Association, "The challenge of health care fraud," National Healthcare Anti-Fraud Association, 2024. https://www.nhcaa.org/tools-insights/about-health-care-fraud/the-challenge-of-health-care-fraud/

[4] C. Zhang, X. Xiao, and C. Wu, "Medical fraud and abuse detection system based on machine learning," *International Journal of Environmental Research and Public Health,* vol. 17, no. 19, p. 7265, 2020. https://doi.org/10.3390/ijerph17197265

[5] J. M. Johnson and T. M. Khoshgoftaar, "Data-centric ai for healthcare fraud detection," *SN Computer Science,* vol. 4, no. 4, p. 389, 2023.

[6] J. D. Kittoe and S. K. Asiedu-Addo, "Exploring fraud and abuse in national health insurance scheme (NHIS) using data mining technique as a statistical model," *African Journal of Educational Studies in Mathematics and Sciences*, vol. 13, pp. 13-31, 2017.

[7] R. K. Alhassan, E. Nketiah-Amponsah, and D. K. Arhinful, "A review of the national health insurance scheme in Ghana: What are the sustainability threats and prospects?," *PloS One*, vol. 11, no. 11, p. e0165151, 2016. https://doi.org/10.1371/journal.pone.0165151

[8] Z. Wang, X. Chen, Y. Wu, L. Jiang, S. Lin, and G. Qiu, "A robust and interpretable ensemble machine learning model for predicting healthcare insurance fraud," *Scientific Reports*, vol. 15, no. 1, p. 218, 2025.

[9] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla, "Credit card fraud detection-machine learning methods," presented at the International Symposium Infoteh-Jahorina (Infoteh), IEEE, 2019, pp. 1-5, 2019.

[10] B. Chosen, "NHIS healthcare claims and fraud dataset," Kaggle, 2024. https://www.kaggle.com/datasets/bonifacechosen/nhis-healthcare-claims-and-fraud-dataset

[11] J. T. Hancock, R. A. Bauder, H. Wang, and T. M. Khoshgoftaar, "Explainable machine learning models for Medicare fraud detection," *Journal of Big Data*, vol. 10, no. 1, p. 154, 2023.

[12] K. G. Al-Hashedi and P. Magalingam, "Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019," *Computer Science Review*, vol. 40, p. 100402, 2021. https://doi.org/10.1016/j.cosrev.2021.100402

[13] M. K. Severino and Y. Peng, "Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata," *Machine Learning with Applications*, vol. 5, p. 100074, 2021. https://doi.org/10.1016/j.mlwa.2021.100074

[14] S. Sheffali and D. Deepa, "Effective fraud detection in healthcare domain using popular classification modeling techniques," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 11, pp. 579-583, 2019.

[15] M. R. Sumalatha and M. Prabha, "Mediclaim fraud detection and management using predictive analytics," presented at the International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), IEEE, 2019, pp. 517-522, 2019.

[16] E. Nabrawi and A. Alanazi, "Fraud detection in healthcare insurance claims using machine learning," *Risks*, vol. 11, no. 9, p. 160. https://doi.org/10.3390/risks11090160

[17] S. K. Shamitha and V. Ilango, "A time-efficient model for detecting fraudulent health insurance claims using artificial neural networks," presented at the International Conference on System, Computation, Automation and Networking (ICSCAN), IEEE, 2020, pp. 1-6, 2020.

[18] J. Chen, X. Hu, D. Yi, M. Alazab, and J. Li, "A variational autoencoder-based relational model for cost-effective automatic medical fraud detection," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 4, pp. 3408-3420, 2022.

[19] J. Zhou *et al.*, "Fraud auditor: A visual analytics approach for collusive fraud in health insurance," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 6, pp. 2849-2861, 2023.

[20] Y. Yoo, J. Shin, and S. Kyeong, "Medicare fraud detection using graph analysis: a comparative study of machine learning and graph neural networks," *IEEE Access*, vol. 11, pp. 88278-88294, 2023.