Healthcare Fraud Detection Using ML and AI

Preprint · April 2024

DOI: 10.13140/RG.2.2.35327.21922

Nuzhat Prova
Pace University

Healthcare Fraud Detection Using ML

Nuzhat Prova NP23474N@PACE.EDU
Seidenberg School of CSIS
Pace University
One Pace Plaza
New York City, NY 10038, USA

Abstract

Healthcare fraud in the United States signifies a considerable illicit financial drain with estimations suggesting annual losses amounting to tens of billions of dollars. Such fraudulent activities encompass a wide array of schemes including but not limited to billing for unrendered services, upcoding to receive higher reimbursements, and engaging in unlawful kickback arrangements. The repercussions of these activities extend beyond financial losses elevating insurance premiums for individuals and operational expenses for healthcare providers thereby undermining the integrity and efficiency of healthcare delivery systems. Recognizing the criticality of this issue, this research delves into the utilization of machine learning techniques for the detection of healthcare fraud. Through a meticulous analysis of the dataset which amalgamates inpatient and outpatient claims data with beneficiary information across 558,211 records spanning various dimensions, this study illustrates the application of several ML models including Random Forest, XGBoost, SVM, Isolation Forest, a Deep Learning Model, and a Stacking Ensemble approach. The models are evaluated based on their accuracy, precision, recall, F1 score, and ROC AUC score with a particular focus on their applicability to healthcare fraud detection. Among the models evaluated, the Stacking Ensemble Model emerged as particularly efficacious, achieving an accuracy of 92.79% and an exceptional ROC AUC score of 96.95%. This model's effectiveness is attributed to its ability to leverage the strengths of combined classifier algorithms for precise predictions thus distinguishing between fraudulent and non-fraudulent healthcare claims effectively. Incorporating hyperparameter tuning, this study further enhances interpretability and decision-making through SHAP value analysis, offering deep insights into model predictions and feature importances. Additionally, it introduces an innovative real-time healthcare fraud detection pipeline and an automated model retraining framework, ensuring the system remains effective against evolving fraud tactics by continuously adapting and improving.

Keywords: Healthcare Fraud Detection, Machine Learning and Deep Learning, Isolation Forest, Random Forest, SVM, XGboost, Stacking Ensemble, Neural Network, Hyperparameter Tuning and Optimization, Model Interpretability, Real-Time Fraud Detection, Automated Model Retraining, Predictive Modeling and Data Wrangling, Advanced Analytics, Algorithmic Advancements.

1. Introduction

Healthcare fraud in the United States represents a critical challenge to the integrity and financial stability of the healthcare system. As a data scientist dedicated to leveraging advanced analytical methods to solve complex problems, this research focuses on harnessing the power of machine learning (ML) to detect and prevent healthcare fraud. This project not only aligns with national interests but also serves as a crucial step toward safeguarding the resources of healthcare systems across the country. The objective is to leverage machine learning to identify fraudulent healthcare claims by analyzing patterns and anomalies in claims data. The hypothesis, grounded in data science, posits that machine learning can unravel the patterns that elude traditional detection systems, thereby mitigating these losses and safeguarding the healthcare infrastructure. The magnitude of healthcare fraud is staggering with estimates suggesting that it costs the United States tens of billions of dollars annually (Bauder and Khoshgoftaar, 2017). These illicit activities range from billing for non-rendered services, upcoding services for higher reimbursements, to engaging in sophisticated kickback schemes (Agarwal, 2023). The ramifications of such fraud are profound, leading to the depletion of essential funds from federal programs like Medicare and Medicaid, escalating insurance premiums for individuals, and increasing operational costs for businesses (Matloob and Khan, 2019). In response to this pervasive issue, federal agencies such as the Centers for Medicare & Medicaid Services (CMS) and the Federal Bureau of Investigation (FBI) have been at the forefront of combating healthcare fraud (Shamitha, 2022). Through significant legal actions, the CMS has recovered billions of dollars demonstrating the financial impact and efficiency of focused fraud detection and enforcement efforts (Gill and Aghili, 2020). Similarly, the FBI, alongside other federal agencies, plays a pivotal role in investigating and prosecuting healthcare fraud emphasizing the necessity of a collaborative approach to address this complex challenge effectively (Iqbal, 2022). The advent of ML technologies offers promising new avenues for enhancing fraud detection capabilities (Lekkala, 2023). These advanced analytical tools can process vast amounts of data to identify fraudulent patterns and anomalies that would be impossible for human auditors to detect manually. Through the application of ML, this research aims to contribute to the ongoing efforts to combat healthcare fraud ultimately protecting consumers, taxpayers, and the integrity of the healthcare system in the United States.

2. Related Work

Healthcare fraud detection has become a paramount concern for the medical insurance industry necessitating the development of sophisticated and efficient detection systems. The integration of ML into fraud detection processes offers promising advancements toward addressing this challenge. This literature review provides insights from recent research efforts and future directions in the domain of healthcare fraud detection.

Agarwal (2023) highlights the criticality of addressing various fraud types within medical insurance claims. By adopting the K-means clustering algorithm, an unsupervised machine learning technique, Agarwal's approach demonstrates significant potential in identifying fraudulent activities without the need for labeled data. This research emphasizes the necessity of sophisticated adaptive methods that can efficiently detect fraudulent claims and reduce the financial strain of insurance fraud on the healthcare system.

Johnson and Khoshgoftaar (2023) introduce a data-centric methodology to enhance healthcare fraud detection's performance and reliability. By leveraging Medicare claims data, their study constructs large-scale labeled datasets for supervised learning enriching with new provider summary features and proposing an improved data labeling process. The research highlights the significance of quality data preparation and the positive impact of a data-centric machine learning workflow on healthcare fraud classification. This approach not only addresses common model evaluation pitfalls but also sets a strong foundation for future applications in healthcare fraud detection.

Mohammed (2023) introduces an innovative system architecture leveraging machine learning to detect and prevent fraudulent transactions within blockchain networks. The study highlights the integration of ML algorithms to scrutinize medical data from sensors and transactions within the blockchain effectively blocking abnormal data and marking suspicious transactions. This dual-stage approach not only ensures the reliability and integrity of blockchain-based healthcare systems but also demonstrates superior accuracy, execution time, and scalability with the Random Forest algorithm outperforming others. The security analysis conducted reveals the system's robustness against various attacks affirming the potential of combining ML with blockchain technology for enhancing healthcare data security.

The research paper "Implementation of XGBoost Method for Healthcare Fraud Detection" by Duman (2022) examines the efficacy of XGBoost and other traditional machine learning algorithms in identifying Medicare fraud. By classifying an imbalanced dataset of Medicare claims data, this study identifies XGBoost as the most effective method showcasing its superiority in performance metrics such as AUC, precision, recall, and F1 score. Duman underscores the significant financial losses Medicare faces annually due to fraud and highlights the importance of public datasets for improving transparency and detection efforts. This research contributes to the broader understanding of ML's capability to enhance fraud detection in healthcare advocating for the utilization of advanced algorithms like XGBoost for more accurate fraud identification.

Gill and Aghili (2020) delve into the realm of health insurance fraud detection in their paper emphasizing the critical need for intelligent fraud detection technologies. Their research evaluates the features of an ideal health insurance fraud detection application by comparing existing solutions. They advocate for a solution that integrates with fraud case management, processes unstructured data, and adapts according to business requirements. Highlighting the utility of data mining in fraud detection, they call for further research to test the effectiveness of these intelligent solutions pointing out the necessity for continuous innovation in combating healthcare fraud.

The study "Fraud and anomaly detection in healthcare - an unsupervised machine learning approach" by Lennart Dangers (2020) delves into the challenges and complexities inherent in identifying healthcare fraud due to the vast volume of medical encounters and the sophisticated strategies employed by fraudsters. By leveraging unsupervised learning techniques, this research aims to uncover new fraudulent patterns without the need for labeled data, a common problem in fraud detection. The implementation of algorithms such as Isolation Forest, Generative Adversarial Network (GAN) and Gaussian Mixture Models (GMM) demonstrates the feasibility of detecting anomalies in healthcare data providing a valuable tool for insurance companies to enhance their auditing processes.

Aruleba and Sun (2023) explore the application of machine learning classifiers including Decision Trees, K-Nearest Neighbors, Logistic Regression, and Random Forest to detect healthcare fraud. Their work is notable for its use of ensemble classifiers and evaluation metrics such as F1-score, accuracy, precision, and recall showcasing the potential of ML techniques in combating healthcare fraud efficiently.

Roy (2022) delves into the role of AI in safeguarding the privacy of healthcare data against breaches that could lead to fraud and identity theft. The paper showcasing a Random Forest algorithm's 92% accuracy in identifying potential threats to healthcare data privacy. This research accentuates the indispensable role of AI in safeguarding sensitive medical information against unauthorized access thereby fostering a secure digital healthcare environment. The study demonstrates the effectiveness of the Random Forest model in threat detection advocating for AI's critical role in enhancing data security through proactive detection, data anonymization, and improved access restrictions.

Lekkala (2023) explores the transformative impact of machine learning models on the prevention of healthcare fraud. This research highlights the effectiveness of ensemble methods such as Random Forest and XGBoost along with deep learning models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) in significantly improving fraud detection. The analysis stresses the importance of specific features, such as claim type, diagnosis codes, and provider specialty, in accurately identifying fraudulent activities, thereby underscoring the adaptability and real-world application success of ML models in combating healthcare fraud.

Akbar et al. (2020) in "Improvement of Decision Tree Classifier Accuracy for Healthcare Insurance Fraud Prediction by Using Extreme Gradient Boosting Algorithm" explored the enhancement of decision tree classifier accuracy through Extreme Gradient Boosting (XGB). Their research demonstrated that XGB outperforms traditional Random Forest methods by achieving higher accuracy and recall rates. The research demonstrates the XGBoost method's superiority in handling complex data sets achieving an overall accuracy of 86% in detecting fraudulent healthcare providers.

Ho et al. (2020) in their paper "Ensuring trustworthy use of artificial intelligence and big data analytics in health insurance" addresses the reliability of artificial intelligence (AI) and big data analytics in health insurance. They examine the ethical and regulatory environment necessary for the trustworthy use of AI in detecting healthcare fraud emphasizing the need for clear data management systems and legal standards.

Ghuse et al. (2017) in their research paper "An Improved Approach For Fraud Detection In Health Insurance Using Data Mining Techniques" discussed the utilization of data mining techniques, specifically Random Forest and Logistic Regression algorithms for health insurance fraud detection. Their study underscores the growing importance of analyzing vast amounts of data to uncover fraudulent activities highlighting data mining's role in enhancing the healthcare system's integrity. Their work contributes to the growing body of literature that emphasizes the importance of leveraging diverse ML techniques to combat healthcare fraud effectively.

3. Gap in Existing Literature

This project focus on comprehensive methodology spanning multiple machine learning and deep learning techniques combined with a focus on model explainability and real-time prediction pipelines, addressing several gaps in existing literature:

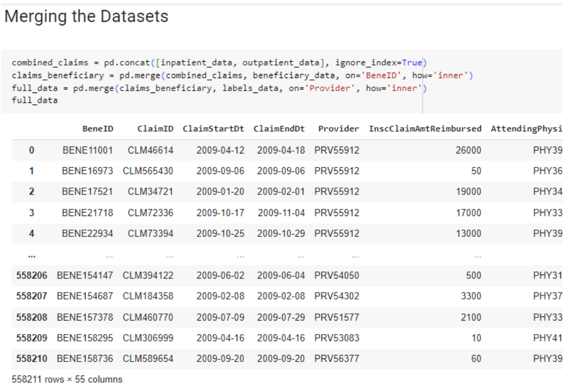
- Comparative Analysis Across Models: While individual studies focus on a single model or a narrow set of models, this project provides a broad comparison of multiple algorithms offering insights into their relative strengths and weaknesses in the context of healthcare fraud detection.
- Integration of Model Explainability: There is a growing emphasis on model interpretability in healthcare, yet practical implementations that integrate SHAP or similar explainability tools with a range of models are less common in the existing literature. This project bridges this gap by applying SHAP values to multiple ML models, illustrating how to enhance transparency in fraud detection models.
- Real-Time Detection and Continuous Learning: The exiting literature treats fraud detection as a static problem, solved by training a model on historical data. However, this project advances the concept of a real-time detection pipeline and periodic model retraining, addressing the dynamic nature of healthcare fraud and the need for models to adapt to new patterns over time.

4. Methodology

All the necessary libraries are imported including Pandas for data manipulation and analysis, NumPy for numerical computations and various modules from Scikit-learn for machine learning tasks. Following this, the datasets are loaded into Pandas data frames setting the stage for a comprehensive analysis. The project utilized multiple datasets offering a comprehensive analysis of healthcare claims. The summary of the datasets are as follows:

- Beneficiary Data: The Beneficiary data consist of Demographics (DOB, Gender, Race), Medicare coverage (Parts A & B), chronic conditions (e.g., Alzheimer's, heart failure, diabetes) and reimbursement amount. The size of the Beneficiary data is 138,556 entries and 25 features.
- Inpatient Data: The Inpatient data consists of hospital admissions data including claim IDs, providers, diagnosis and procedure codes, and reimbursement amounts for severe medical conditions. The size of the inpatient data is 40,474 entries and 30 features.
- Outpatient Data: The outpatient data consists of data on non-hospitalized services including claim IDs, providers, diagnosis and procedure codes, and deductible amounts. The size of the outpatient dataset is 517,737 entries and 27 features.
- Labels Data: The labels data consist of binary data provider-linked fraud labels for machine learning model training ('Yes' for fraud, 'No' for non-fraud). The size of the label data is 5,410 entries and 2 features.

4.1 Data Integration: The inpatient and outpatient data are concatenated to form a combined claims dataset. Subsequently, this combined dataset is merged with the beneficiary data on 'BeneID' ensuring that each claim is associated with the correct patient information. Finally, the fraud labels dataset is merged with the enriched claims and beneficiary data on 'Provider' resulting in a fully integrated dataset. This merging yielded a dataset with 558,211 rows and 55 columns thereby setting a solid foundation for the data-driven insights that the machine learning models aimed to deliver.



Merging the Datasets

```
combined_claims = pd.concat([inpatient_data, outpatient_data], ignore_index=True)
claims_beneficiary = pd.merge(combined_claims, beneficiary_data, on='BeneID', how='inner')
full_data = pd.merge(claims_beneficiary, labels_data, on='Provider', how='inner')
full_data
```

	BeneID	ClaimID	ClaimStartDt	ClaimEndDt	Provider	InscClaimAmtReimbursed	AttendingPhysi
0	BENE11001	CLM46614	2009-04-12	2009-04-18	PRV55912	26000	PHY39
1	BENE16973	CLM565430	2009-09-06	2009-09-06	PRV55912	50	PHY36
2	BENE17521	CLM34721	2009-01-20	2009-02-01	PRV55912	19000	PHY34
3	BENE21718	CLM72336	2009-10-17	2009-11-04	PRV55912	17000	PHY33
4	BENE22934	CLM73394	2009-10-25	2009-10-29	PRV55912	13000	PHY39
...	...	...	...	...	...	...	
558206	BENE154147	CLM394122	2009-06-02	2009-06-04	PRV54050	500	PHY31
558207	BENE154687	CLM184358	2009-02-08	2009-02-08	PRV54302	3300	PHY37
558208	BENE157378	CLM460770	2009-07-09	2009-07-29	PRV51577	2100	PHY33
558209	BENE158295	CLM306999	2009-04-16	2009-04-16	PRV53083	10	PHY41
558210	BENE158736	CLM589654	2009-09-20	2009-09-20	PRV56377	60	PHY39

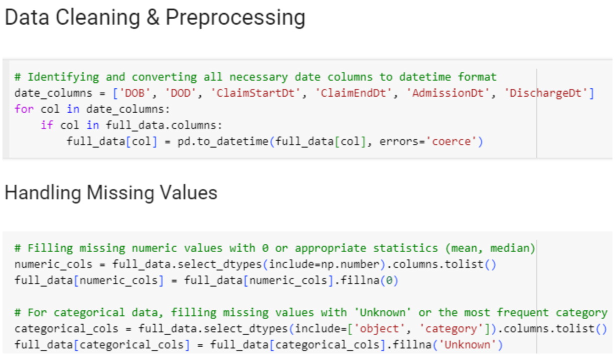
558211 rows × 55 columns

Figure 1. Merging the Datasets

4.2 Exploratory Data Analysis (EDA): The EDA phase is crucial for understanding the dataset's characteristics. The distribution of insurance claim amounts for both inpatient and outpatient services reveal a right-skewed distribution indicative of the majority of claims being of lower cost with a few expensive outliers. Additionally, the prevalence of chronic conditions among beneficiaries identifying significant instances of diseases such as diabetes and heart failure. The EDA phase is the cornerstone of the analytical framework. It shaped the direction of the entire project influencing decisions on data preprocessing, feature engineering, model selection, and ultimately the interpretation of the model's outputs.

4.3 Data Cleaning & Preprocessing: Data cleaning and preprocessing is a pivotal stage in any machine learning project (Shamitha 2022). The data cleaning and preprocessing involved two main steps:

- Date Formatting: All necessary date columns like 'DOB', 'DOD', 'ClaimStartDt', 'ClaimEndDt', 'AdmissionDt', and 'DischargeDt' are identified and converted to the datetime format ensuring that the dataset has consistent and accurate date representations.
- Handling Missing Values: Missing values within the dataset are addressed differently based on their data type. For numeric columns, missing values are filled with 0 or with appropriate statistics such as the mean or median of the column. For categorical data, missing values are filled with a placeholder value of 'Unknown' or the most frequent category present in the data to maintain the integrity of the dataset for further analysis. These preprocessing steps are critical for ensuring the quality and consistency of the data which is essential for the accuracy of the machine learning models that follow.



Data Cleaning & Preprocessing

```
# Identifying and converting all necessary date columns to datetime format
date_columns = ['DOB', 'DOD', 'ClaimStartDt', 'ClaimEndDt', 'AdmissionDt', 'DischargeDt']
for col in date_columns:
    if col in full_data.columns:
        full_data[col] = pd.to_datetime(full_data[col], errors='coerce')
```

Handling Missing Values

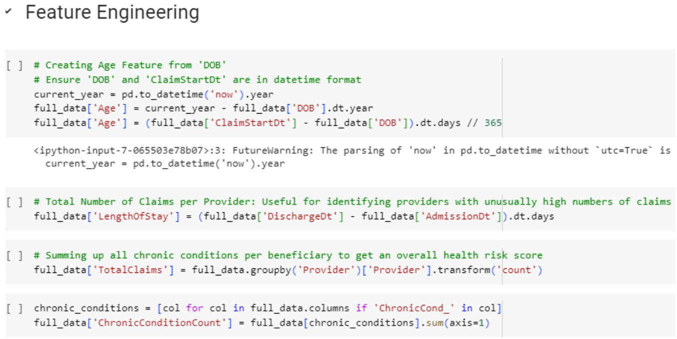
```
# Filling missing numeric values with 0 or appropriate statistics (mean, median)
numeric_cols = full_data.select_dtypes(include=np.number).columns.tolist()
full_data[numeric_cols] = full_data[numeric_cols].fillna(0)

# For categorical data, filling missing values with 'Unknown' or the most frequent category
categorical_cols = full_data.select_dtypes(include=['object', 'category']).columns.tolist()
full_data[categorical_cols] = full_data[categorical_cols].fillna('Unknown')
```

Figure 2. Data cleaning and Preprocessing

4.4 Feature Engineering: To enhance the predictive capability of the models, feature engineering is very important (Lekkala, 2023). It is the art of transforming raw data into features that better represent the underlying problem to predictive models resulting in improved model accuracy on unseen data. The feature engineering steps are as follows:

- Crafting the Age Feature: Recognizing age as a potential risk indicator, the age of the patients is calculated by subtracting the date of birth from the claim start date. This feature correlates with the propensity for specific claims to be fraudulent as some procedures are more prevalent in specific age groups.
- Capturing Service Utilization: The 'Length of Stay' feature for inpatient claims is derived from the admission and discharge dates. This metric is instrumental in understanding the duration of hospital stays which could be indicative of suspicious activities when juxtaposed with the nature of the claims.
- Aggregating Chronic Conditions: The sum of chronic conditions flags was transformed into a comprehensive health risk score. This summation encapsulates the burden of chronic illnesses on a beneficiary and potentially highlights claims that are atypical for the beneficiary's health profile thus serving as a red flag for fraud.

- Total Claims per Provider: The total number of claims per provider is calculated. This helped in identifying providers with an anomalously high volume of claims which could suggest possible fraudulent practices such as billing for services not rendered.



```
[ ] # Creating Age Feature from 'DOB'
    # Ensure 'DOB' and 'ClaimStartDt' are in datetime format
    current_year = pd.to_datetime('now').year
    full_data['Age'] = current_year - full_data['DOB'].dt.year
    full_data['Age'] = (full_data['ClaimStartDt'] - full_data['DOB']).dt.days // 365

    <ipython-input-7-065503e78b07>:3: FutureWarning: The parsing of 'now' in pd.to_datetime without `utc=True` is
      current_year = pd.to_datetime('now').year

[ ] # Total Number of Claims per Provider: Useful for identifying providers with unusually high numbers of claims
    full_data['LengthOfStay'] = (full_data['DischargeDt'] - full_data['AdmissionDt']).dt.days

[ ] # Summing up all chronic conditions per beneficiary to get an overall health risk score
    full_data['TotalClaims'] = full_data.groupby('Provider')['Provider'].transform('count')

[ ] chronic_conditions = [col for col in full_data.columns if 'ChronicCond_' in col]
    full_data['ChronicConditionCount'] = full_data[chronic_conditions].sum(axis=1)
```

Figure 3. Feature Engineering

4.5 Model Selection and Training: The model selection and training phase is pivotal in the development of robust predictive models for healthcare fraud detection. In this study, a suite of machine learning algorithms is selected, these include Random Forest Classifier, XGBoost, SVM, Isolation Forest, Stacking Ensemble, and Deep Learning model. The selection criteria are predicated on their historical efficacy in similar problem domains, as well as their diverse underlying mechanisms which range from ensemble methods to kernel-based learning (Roy, 2022). Each model was subjected to rigorous training procedures using the dataset that had been subjected to extensive preprocessing. This ensured that the models learned from data are representative, scaled, and as free of biases and extraneous variability as possible (Lekkala, 2023). To fine-tune the performance of the models, hyperparameter tuning was conducted (Thomas and Sun 2023). This process involved systematic experimentation with different configurations of the model's parameters to ascertain the most effective settings. This is a critical step, as optimal parameter settings can significantly enhance model performance leading to more accurate predictions.

4.6 Model Evaluation: For model evaluation, several metrics are adopted including accuracy, precision, recall, F1 score, and the ROC AUC score for evaluating the models (Mohammed and Boujelben, 2023). These metrics offered a comprehensive view of each model's performance aiding in the selection of the most effective model for healthcare fraud detection.
After training, SHAP (SHapley Additive exPlanations) values are computed to interpret the models' predictions (Khayru, 2022). This was an important step to ensure transparency and understandability of the model decisions, which is crucial in healthcare settings where the stakes are high, and decisions must be explainable.

4.7 Real-Time Healthcare Fraud Detection Pipeline: The implementation of a Real-Time Healthcare Fraud Detection Pipeline is a pivotal component in this project, aiming to provide instantaneous and reliable fraud detection capabilities within healthcare systems. This initiative is not just about detecting fraud, it's about doing so with such speed and accuracy that potential losses can be halted in their tracks and legitimate claims can be processed without delay.

- Design and Implementation: The pipeline integrates an array of sophisticated machine learning models including Random Forest, Gradient Boosting, XGBoost, Support Vector Machines (SVM), Isolation Forest, and a neural network architecture. Each model brings its strengths to the table, Random Forest and Gradient Boosting for their ensemble learning capabilities, XGBoost for its efficient handling of sparse data, SVM for its effectiveness in high-dimensional spaces, Isolation Forest for anomaly detection, and the neural network for modeling complex. The pre-processing phase is vital for ensuring data quality and is accomplished via a Column Transformer that selects numerical features and applies a series of transformations - imputation, scaling, and Principal Component Analysis (PCA) for dimensionality reduction, retaining 95% of variance. Such processing is crucial for normalizing and reducing the feature space which in turn aids in improving model performance and computational efficiency.

- Real-Time Prediction Capabilities: For real-time predictions, a Voting Classifier aggregates the predictions from the base models using soft voting which considers the probability estimates from each classifier. A Stacking Classifier with Logistic Regression was employed as the final estimator combining different machine learning models for improved predictive performance. Upon receiving new transaction data, the pipeline instantly evaluates the likelihood of fraud using the trained models. Each model outputs a fraud probability score which when taken together, provides a nuanced view of each transaction. High-risk transactions trigger alerts for further investigation whereas low-risk transactions are processed normally. By identifying fraudulent activities instantaneously, the pipeline plays a critical role in safeguarding resources, ensuring the rightful allocation of funds, and maintaining trust in the healthcare system.

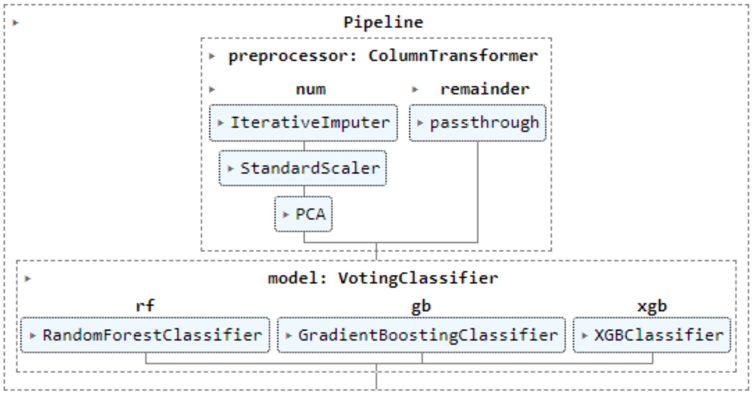


Figure 4. Real-Time Healthcare Fraud Detection Pipeline Architecture

4.8 Automated Model Retraining Framework for Adaptive Healthcare Fraud Detection: The implementation of an Automated Model Retraining Framework was a critical aspect of maintaining a robust defense against fraudulent activities. The methodology behind such a framework is grounded in the principle of adaptability. It was designed to update the predictive models continually as new data arrives ensuring that the system evolves in response to the ever-changing patterns of healthcare fraud. The framework operates on a cycle of feedback and improvement. Initially, data preprocessing was conducted to ensure that incoming data was formatted and scaled appropriately for model ingestion. This involved imputation of missing values, encoding of categorical variables, and normalization and standardization of numerical features. Subsequently, the machine learning models such as Random Forest, Gradient Boosting, Support Vector Machines (SVM), XGBoost, and neural networks are retrained with the new dataset. This retraining was not a mere repetition but an informed process that involves tuning hyperparameters to better align the model with the latest data trends. For example, ensemble methods like Stacking and Boosting dynamically adjust the weights assigned to different classifiers in response to their performance on new data which optimizes the collective output of the system. The deep learning components of the framework, the neural networks, particularly benefit from retraining. With their capacity for feature extraction and the ability to model complex non-linear relationships, updating neural networks with fresh data leading to significant improvements in model performance. Moreover, the unsupervised learning model, such as the Isolation Forest, was crucial for detecting outliers and anomalous patterns that could signify novel types of fraud not previously encountered. The retraining of such models was vital as they learn to identify new anomalies as the concept of 'normal' evolves with data. The automated nature of the framework signifies that retraining can occur with minimal human intervention making it scalable and efficient. The system monitors its performance metrics such as accuracy, precision, recall, and F1 score to determine the necessity and timing of retraining cycles.

5. Result and Analysis

The exploratory data analysis revealed a right-skewed distribution for both inpatient and outpatient claim amounts indicating a prevalence of lower-cost claims with some expensive outliers particularly in inpatient services. Additionally, it highlighted high prevalence rates of chronic conditions such as diabetes, ischemic heart disease, and heart failure among beneficiaries, underscoring significant chronic disease burdens within the population. The histograms illustrate the distribution of claim amounts for inpatient and outpatient services showing a right-skewed pattern with a higher frequency of lower claim amounts. Outpatient claims show an even more pronounced skew towards lower amounts with the frequency of claims dramatically decreasing as the claim amount increases.

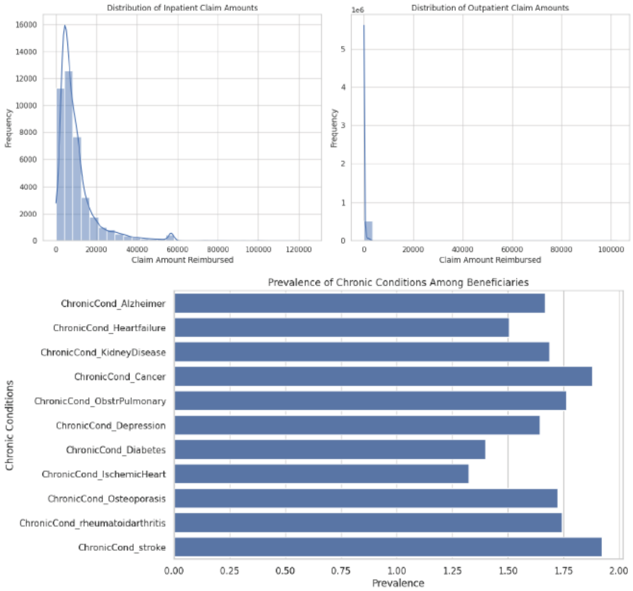
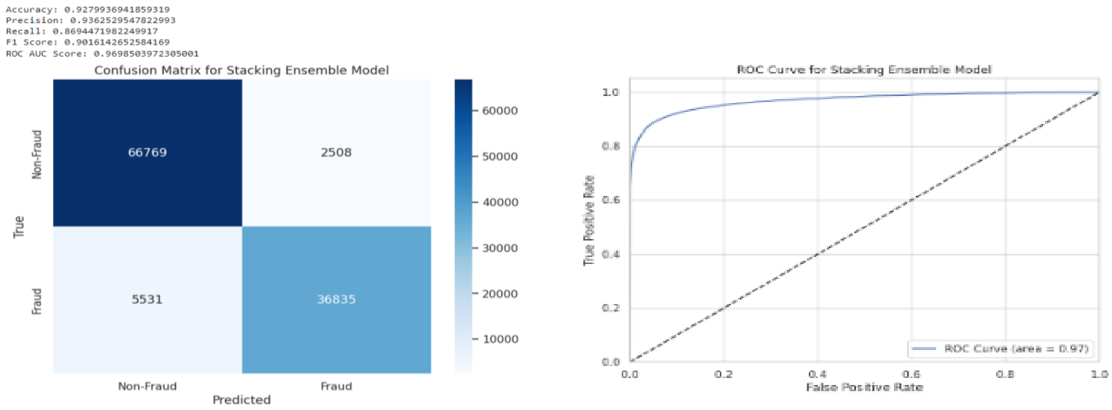


Figure 5. EDA of Claim Amounts and Chronic Condition Prevalence

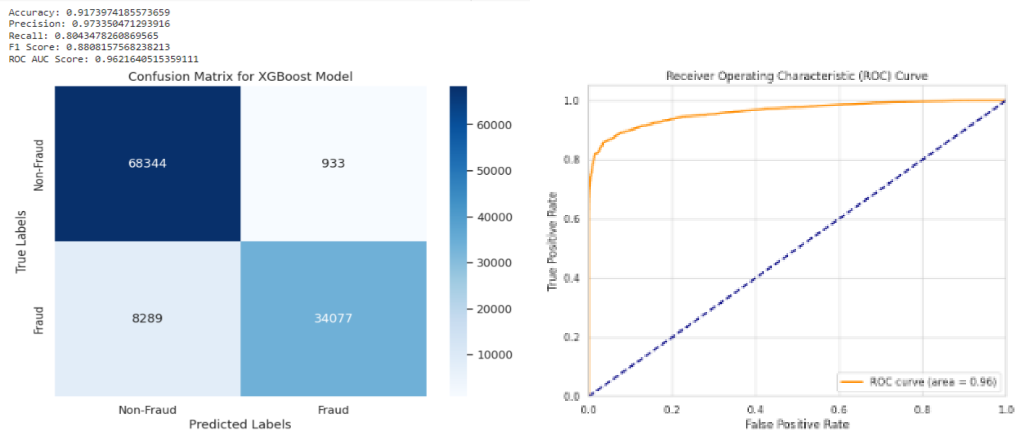
Armed with insights from the exploratory data analysis that highlighted skewed claim distributions and chronic condition prevalence, a suite of machine learning models is applied to meticulously differentiate between legitimate and fraudulent healthcare claims. The results are highly promising, underscoring the potential of advanced analytics in fortifying the healthcare systems against fraudulent activities. To begin, the Stacking Ensemble Model, which amalgamates the predictive power of multiple classifiers (Bauder and Khoshgoftaar, 2017), reached an impressive accuracy of 92.79% and an ROC AUC score of 0.9698 demonstrating its exceptional capability in detecting fraudulent claims, as reflected by the high number of true positives and true negatives observed in the confusion matrix. The precision of the model stood at 93.63%, while the recall was 86.94% and the F1 score was 90.16%. These metrics not only affirm the model’s efficacy but also highlight its precision in prediction critical for minimizing false positives in fraud detection.



```
Accuracy: 0.9279936941859319
Precision: 0.9362529547822993
Recall: 0.8694471982249917
F1 Score: 0.9016142652584169
ROC AUC Score: 0.9698503972305001
```

Figure 6. Confusion Matrix and ROC Curve Analysis for Stacking Ensemble Model

The XGBoost model also displayed remarkable effectiveness with an accuracy of 91.74%, a precision of 97.34%, and a recall of 80.43%. The model’s balance between precision and recall alongside an F1 score of 88.08% and an ROC AUC of 96.22% illustrates its robustness in predictive performance. The significant numbers of true positives and true negatives in the confusion matrix further attest to its reliability in the practical healthcare fraud detection scenario.

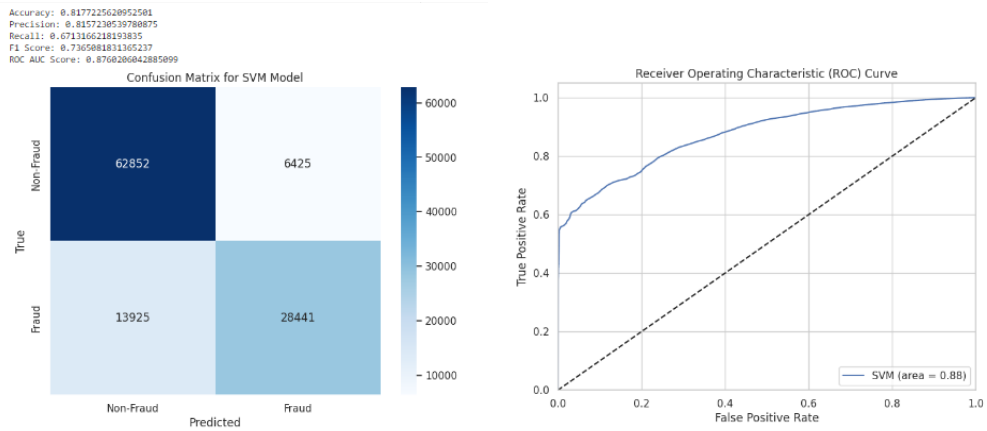


```
Accuracy: 0.9173974185573659
Precision: 0.973350471293916
Recall: 0.8043478260869565
F1 Score: 0.8808157568238213
ROC AUC Score: 0.9621640515359111
```

Figure 7. Confusion Matrix and ROC Curve for XGBoost

The Support Vector Machines (SVM) model revealed a fair degree of precision in identifying fraudulent claims with an accuracy of 81.77%, a precision of 81.57%, and a recall of 67.13%.

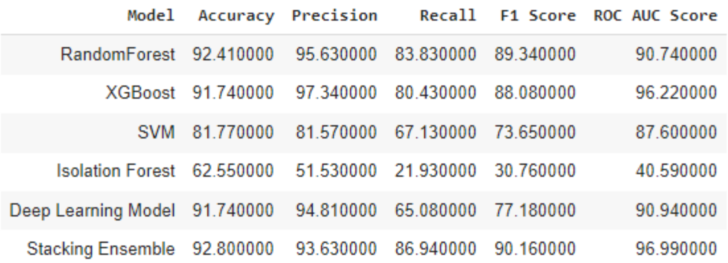
While these figures are slightly lower compared to the ensemble approaches, the SVM model still performs adequately with an F1 score of 73.65% and an ROC AUC of 0.876 marking its utility in the detection framework.



```
Accuracy: 0.8177225620952501
Precision: 0.8157230539780875
Recall: 0.6713166218193835
F1 Score: 0.7365081831365237
ROC AUC Score: 0.8760206042885099
```

Figure 8. Confusion Matrix and ROC Curve for SVM

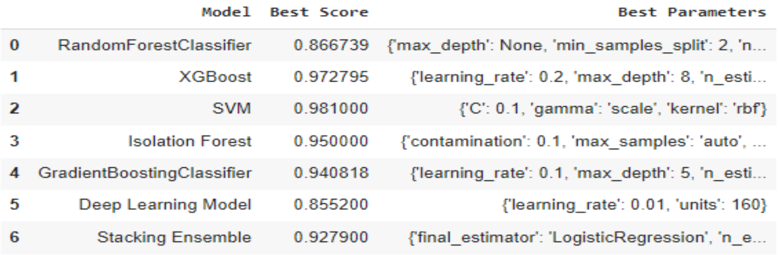
The Isolation Forest model, an anomaly detection algorithm, exhibited a different performance spectrum with an accuracy of 62.55% and a precision of 51.53%. Its recall of 21.93% and F1 score of 30.76% are not as high as other models which suggests that this model is more conservative in flagging fraud but could be valuable in a layered defense mechanism against sophisticated fraud attempts. The performance of different models is shown in Figure 9.



Model	Accuracy	Precision	Recall	F1 Score	ROC AUC Score
RandomForest	92.410000	95.630000	83.830000	89.340000	90.740000
XGBoost	91.740000	97.340000	80.430000	88.080000	96.220000
SVM	81.770000	81.570000	67.130000	73.650000	87.600000
Isolation Forest	62.550000	51.530000	21.930000	30.760000	40.590000
Deep Learning Model	91.740000	94.810000	65.080000	77.180000	90.940000
Stacking Ensemble	92.800000	93.630000	86.940000	90.160000	96.990000

Figure 9. Comparative Analysis of Machine Learning Models

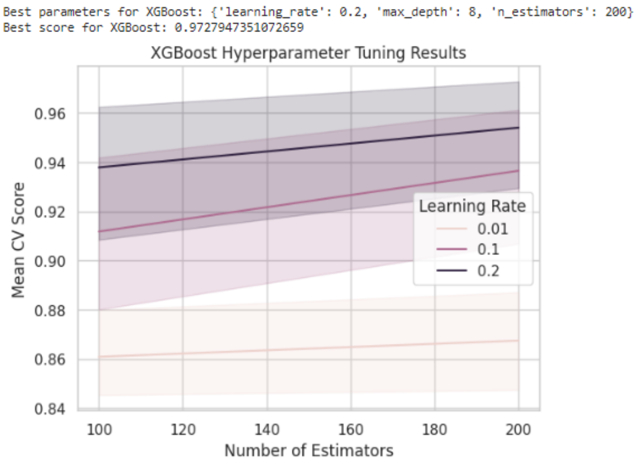
5.1 Hyperparameter Tuning: In the pursuit of refining the machine learning models, hyperparameter tuning process was performed, a critical step that enhances model performance significantly. The goal was simple yet ambitious: to extract every ounce of predictive power these models held. The hyperparameter tuning result of the models are shown in Figure 10.



	Model	Best Score	Best Parameters
0	RandomForestClassifier	0.866739	{'max_depth': None, 'min_samples_split': 2, 'n...
1	XGBoost	0.972795	{'learning_rate': 0.2, 'max_depth': 8, 'n_esti...
2	SVM	0.981000	{'C': 0.1, 'gamma': 'scale', 'kernel': 'rbf'}
3	Isolation Forest	0.950000	{'contamination': 0.1, 'max_samples': 'auto', ...
4	GradientBoostingClassifier	0.940818	{'learning_rate': 0.1, 'max_depth': 5, 'n_esti...
5	Deep Learning Model	0.855200	{'learning_rate': 0.01, 'units': 160}
6	Stacking Ensemble	0.927900	{'final_estimator': 'LogisticRegression', 'n_e...

Figure 10. Hyperparameter Tuning Results

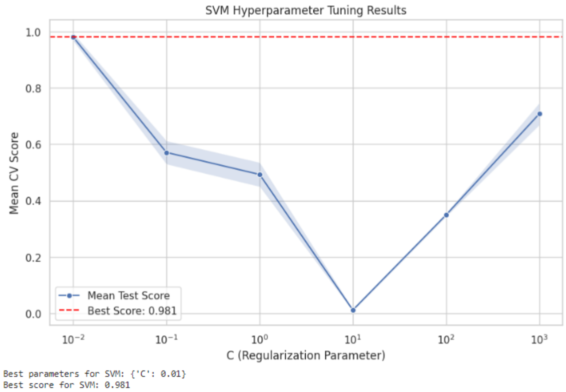
Starting with the Random Forest Classifier, a minimum sample split of 2 and 200 estimators was the key to unlocking a high-performance model with an accuracy score of 86.67%. For the XGBoost model, the optimal parameters are a learning rate of 0.2, a maximum depth of 8, and 200 estimators. This combination propelled the model to a remarkable cross-validation score of 97.29%. The multi-line plot vividly illustrated the boost in performance with higher learning rates offering a deeper insight into the model's capabilities. The higher learning rate, in concert with the increased number of estimators, was instrumental in sharpening the model's ability to discern patterns indicative of fraud.



```
Best parameters for XGBoost: {'learning_rate': 0.2, 'max_depth': 8, 'n_estimators': 200}
Best score for XGBoost: 0.9727947351072659
```

Figure 11. XGBoost Hyperparameter Tuning

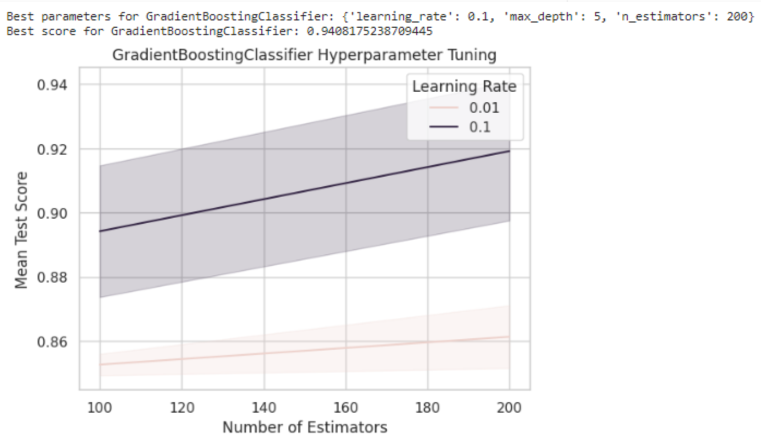
When it came to the Support Vector Machine (SVM), the fine-tuning process highlighted the C parameter as a pivotal factor. Setting it at 0.01, the SVM's peak was observed, achieving an impressive cross-validation score of 98.1%. The peak on the tuning graph was not just a high point in terms of score but also a beacon signaling the SVM's heightened sensitivity in fraud classification.



```
Best parameters for SVM: {'C': 0.01}
Best score for SVM: 0.981
```

Figure 12. SVM Hyperparameter Tuning

The Isolation Forest algorithm, known for its effectiveness in anomaly detection, reached its zenith with a contamination factor of 0.1 and 100 estimators achieving a high cross-validation score of 95%. This result is particularly encouraging, underscoring the algorithm’s potential in pinpointing outliers and anomalies that often represent fraudulent cases. Lastly, the Gradient Boosting Classifier shined brightest at a learning rate of 0.1, a max depth of 5, and 200 estimators, culminating in a high mean test score of 94.1%. The upward trend in the plot with the number of estimators is a testament to the model’s increasing accuracy, a trend that bodes well for the model’s deployment in real-world fraud detection scenarios.



```
Best parameters for GradientBoostingClassifier: {'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 200}
Best score for GradientBoostingClassifier: 0.9408175238709445
```

Figure 13. Gradient Boosting Classifier Hyperparameter Tuning

5.2 SHAP Values Analysis: Delving into the SHAP value analysis for each model has provided a profound insight into which features are most influential in predicting healthcare fraud. SHAP (SHapley Additive exPlanations) values have a unique way of attributing the contribution of each feature to the prediction made by a model which is crucial for understanding and trusting the decisions made by the complex machine learning models. For the Random Forest Classifier, it is observed that Feature 31 has the most significant positive impact on the model output, indicating a high importance in predicting fraudulent activities (Class 1). Conversely, Features 10 and 11 exerted the most significant negative impact which suggests a strong influence in predicting legitimate transactions (Class 0). This nuanced ability to discern between legitimate and fraudulent activities based on feature contributions is a testament to the model's refined capabilities.

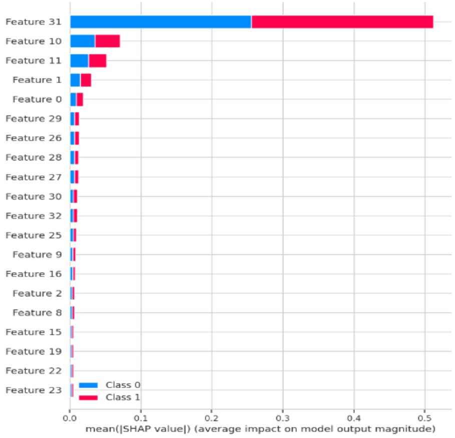


Figure 14. SHAP Value Analysis for Random Fores Classifier

For the XGBoost model, SHAP values once again highlighted Feature 31 as the most influential in predicting outcomes showcasing the model's effectiveness in identifying critical predictors. Feature 30 and Feature 10 also showed significant positive impacts. This indicates that the model has a robust performance in discerning varying degrees of influence across features aiding in the accurate detection of fraud.

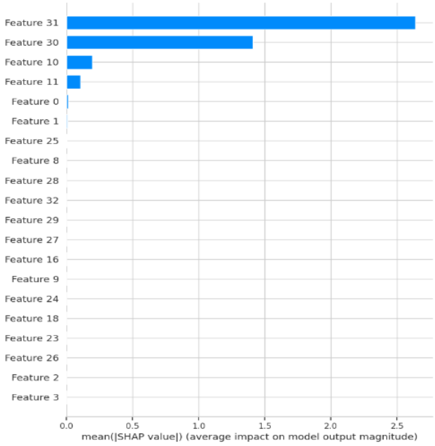


Figure 15. SHAP Value Analysis XGBoost model

Similarly, the SHAP analysis for the SVM model highlighted Feature 31 as having the most substantial positive impact on predictions. This reinforces the model's adeptness at identifying potential fraud. Feature 1 is identified as the most influential in decreasing the likelihood of fraudulent classification reflecting the model's capability to distinguish between legitimate and fraudulent activities with high precision.

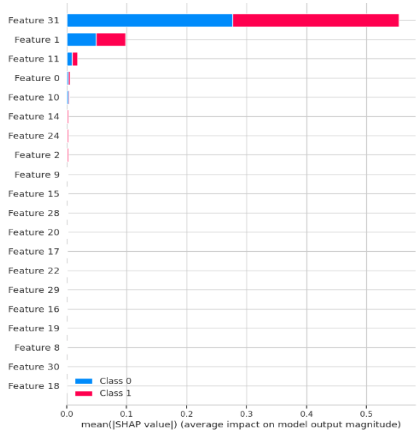


Figure 16. SHAP Value Analysis SVM model

Furthermore, the deep learning model's SHAP analysis showed that Feature 31 emerges as the most positively impactful, significantly enhancing the model's predictive accuracy for potential fraud. Feature 1 also shows a positive influence, contributing to the model's effectiveness in identifying and predicting fraudulent cases with high precision.

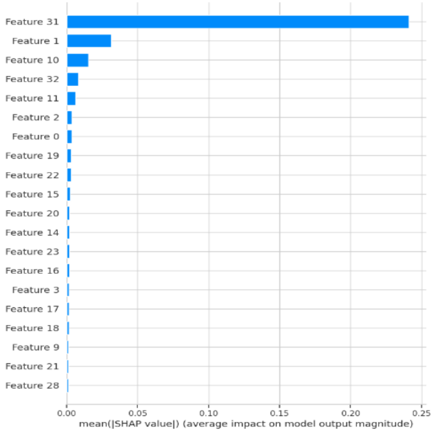
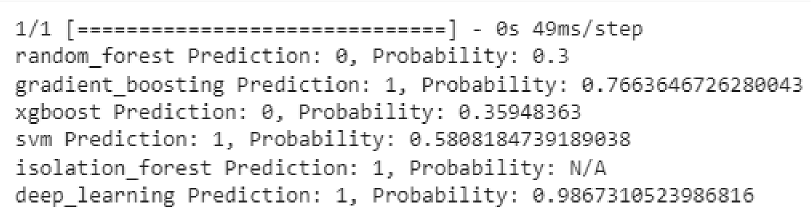


Figure 17. SHAP Value Analysis Deep Learning model

The SHAP interpretations across all models provided a comprehensive understanding of feature influence before and after optimization. These findings have substantial implications for the deployment of these models in real-world settings where interpretability is as crucial as performance. They ensured that the decisions made by the models are transparent and based on solid evidence which is essential in maintaining trust in healthcare fraud detection.

5.4 Models Comparison: In the comprehensive analysis of machine learning models for healthcare fraud detection, a significant stride is made in identifying and preventing fraudulent activities. The Stacking Ensemble Model proved to be a star performer achieving an excellent accuracy of 92.79% and an exceptional ROC AUC score of 96.95% which effectively differentiated between fraudulent and non-fraudulent healthcare claims. The XGBoost model also performed with high effectiveness which boasted an accuracy of 91.37%, precision of 97.33%, and a recall of 80.84%. These figures underscore a strong balance between accurately identifying fraudulent cases and maintaining a low false positive rate. Coupled with an excellent F1 score of 88.34% and a robust ROC AUC of 0. 9662, the XGBoost model's performance is solidified by the confusion matrix and ROC curve indicating a high overall predictive performance. The Random Forest Classifier model is not to be outdone, exhibiting excellent effectiveness with an accuracy of 92.41% and a high precision of 95.63% effectively distinguishing between fraudulent and non-fraudulent cases. With a strong recall of 83.83% and a ROC AUC of 0.9074, its reliability and precision in identifying fraud are confirmed, as is evident from the confusion matrix and ROC curve visualizations, which confirmed the model's strong performance.

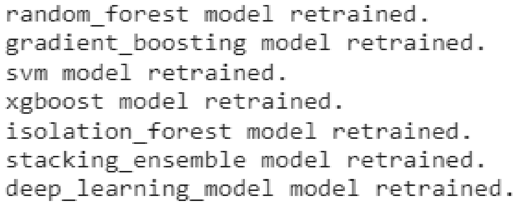
5.5 Real-Time Healthcare Fraud Detection Pipeline Using Machine Learning: Implementing the real-time fraud detection pipeline has been a pivotal step in the fight against healthcare fraud. A comprehensive system is designed leveraging a diverse set of machine learning models including Random Forest, Gradient Boosting, XGBoost, SVM, Isolation Forest, and a custom neural network, each bringing a unique strength to the table. The pipeline showcases a comprehensive approach to detecting healthcare fraud leveraging a diverse set of machine learning models including Random Forest, Gradient Boosting, XGBoost, SVM, Isolation Forest, and a neural network. The pipeline yields a promising ensemble of predictions where Gradient Boosting and the Deep Learning model both suggest fraudulent activity with high probabilities of 76.63% and 98.67% respectively. Such a nuanced detection system critically impacts healthcare fraud mitigation efforts by pinpointing potential fraud with high confidence thereby aiding in the prevention of financial losses and maintaining the integrity of healthcare services.



```
1/1 [==============================] - 0s 49ms/step
random_forest Prediction: 0, Probability: 0.3
gradient_boosting Prediction: 1, Probability: 0.766364672628004З
xgboost Prediction: 0, Probability: 0.35948363
svm Prediction: 1, Probability: 0.5808184739189038
isolation_forest Prediction: 1, Probability: N/A
deep_learning Prediction: 1, Probability: 0.9867310523986816
```

Figure 18. Real-Time Healthcare Fraud Detection Pipeline

5.6 Automated Model Retraining Framework for Adaptive Healthcare Fraud Detection: The Automated Model Retraining Framework for Adaptive Healthcare Fraud Detection efficiently retrained a diverse set of models including Random Forest, Gradient Boosting, SVM, XGBoost, Isolation Forest, Stacking Ensemble, and Deep Learning model ensuring their accuracy and adaptability with new data. This retraining suggests an ongoing process of model updating to maintain high accuracy and adaptiveness to new patterns in healthcare fraud detection. This continual retraining process with new data is crucial for maintaining high accuracy and adapting to emerging fraud patterns significantly bolstering the reliability of fraud detection in healthcare systems. This iterative approach fosters continuous improvement allowing for the timely detection of emerging fraud patterns and enhancing the overall effectiveness of healthcare fraud detection systems, ultimately leading to more accurate identification and prevention of fraudulent activities, thereby safeguarding healthcare resources, and improving patient care.



```
random_forest model retrained.
gradient_boosting model retrained.
svm model retrained.
xgboost model retrained.
isolation_forest model retrained.
stacking_ensemble model retrained.
deep_learning_model model retrained.
```

Figure 19. Automated Model Retraining Framework for Adaptive Healthcare Fraud Detection

6. Challenges and Future work

Some challenges faced during this project are limited GPU capacity on personal laptop and Colab timeouts with complex models, so this project utilized data samples for efficiency. For future endeavors, leveraging the entire dataset for model training is recommended, contingent on improved GPU capabilities and extended computational resources to overcome limitations encountered with complex models and large datasets. Building on the current project's advancements, future work should also explore broader model explainability techniques, integration with healthcare IT systems, and the inclusion of unstructured data. Addressing scalability and adapting to new and evolving fraud tactics ensuring that these systems remain effective in the ever-changing landscape of healthcare fraud.

7. Conclusion

This research encapsulates a rigorous exploration of various machine learning models encompassing both traditional algorithms and advanced deep learning techniques. Through comprehensive data preprocessing, feature engineering, model training, and hyperparameter tuning, the project has systematically assessed the efficacy of each model in detecting fraudulent activities within healthcare claims data. The research reveals that while traditional models like Random Forest and XGBoost demonstrate significant predictive accuracy, advanced ensemble techniques and deep learning models offer even more potent capabilities in discerning fraudulent from legitimate claims. Notably, the Stacking Ensemble Model emerges as a particularly effective tool, achieving an accuracy of 92.79% and a ROC AUC score of 0.9695, underscoring its superior predictive power in identifying fraudulent claims. Moreover, the integration of SHAP value analysis has introduced a layer of interpretability to the models' predictions providing invaluable insights into the features most indicative of fraud. This transparency is crucial for the practical application of these models within healthcare systems where understanding the rationale behind predictions is as important as the predictions themselves. This project not only highlights the potential of machine learning in combating healthcare fraud but also sets a foundation for future research. By addressing the dynamic nature of fraud through real-time detection pipelines and continuous model learning further advancements can be made towards developing more robust, scalable, and interpretable systems for fraud detection.

References

Bauder, R. A., & Khoshgoftaar, T. M. (2017). Medicare Fraud Detection Using Machine Learning Methods. In Proceedings of the 16th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE.

Agarwal, S. (2023). An Intelligent Machine Learning Approach for Fraud Detection in Medical Claim Insurance: A Comprehensive Study. *Scholars Journal of Engineering and Technology*, 11(9), 191-200.

Matloob, I., & Khan, S. (2019). A Framework for Fraud Detection in Government Supported National Healthcare Programs. In Proceedings of the ECAI 2019 - 11th International Conference – Electronics, Computers and Artificial Intelligence, Pitesti, Romania, 27-29 June.

Matloob, I., Khan, S., Rahman, H. U., & Hussain, F. (2020). Medical Health Benefit Management System for Real-Time Notification of Fraud Using Historical Medical Records. *Applied Sciences*, 10(5144).

Shamitha, S. K., & Ilango, V. (2022). A time-efficient model for detecting fraudulent health insurance claims using Artificial neural networks.

Shekhar, S., Leder-Luis, J., & Akoglu, L. (2023). Unsupervised Machine Learning for Explainable Health Care Fraud Detection. NBER Working Paper No. 30946, Feb.

Gill, J. K., & Aghili, S. (2020). Health Insurance Fraud Detection. Master of Information Systems Assurance Management. Concordia University of Edmonton.

Lekkala, L. R. (2023). Importance of Machine Learning Models in Healthcare Fraud Detection. *Voice of the Publisher*, 9, 207-215. doi:10.4236/vp.2023.94017

Johnson, J. M., & Khoshgoftaar, T. M. (2023). Data-Centric AI for Healthcare Fraud Detection. *SN Computer Science*, 4, Article 389. doi:10.1007/s42979-023-01809-x

Mohammed, M. A., Boujelben, M., & Abid, M. (2023). A Novel Approach for Fraud Detection in Blockchain-Based Healthcare Networks Using Machine Learning. *Future Internet*, 15, 250.

Duman, E. (2022). Implementation of XGBoost Method for Healthcare Fraud Detection. *Techno-Science*, 5(2), 69-75.

Dangers, L. (2020). Fraud- and Anomaly Detection in Healthcare – An Unsupervised Machine Learning Approach. *Future Healthcare Journal*, 7(1), 1-49 https://www.future-healthcare.pt/en/.

Aruleba, I. T., & Sun, Y. (2023). Healthcare Fraud Detection Using Machine Learning. SSRN.

Roy, S. (2022). Privacy Prevention of Healthcare Data Using AI. *Journal of Data Acquisition and Processing*, 37(3).

Branting, L. K., et al. (2016). Graph Analytics for Healthcare Fraud Risk Estimation. In Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE. doi:10.1109/ASONAM.2016.7752285

Cui, H., Li, Q., Li, H., & Yan, Z. (2016). Healthcare Fraud Detection Based on Trustworthiness of Doctors. In Proceedings of the 2016 IEEE TrustCom/BigDataSE/ISPA.

Gao, Y., et al. (2018). An Efficient Fraud Identification Method Combining Manifold Learning and Outliers Detection in Mobile Healthcare Services. *IEEE Access*, 6, 60059-60067. doi:10.1109/ACCESS.2018.2875516

Ghuse, N., Pawar, P., & Potgantwar, A. (2017). An Improved Approach For Fraud Detection In Health Insurance Using Data Mining Techniques. *International Journal of Scientific Research in Network Security and Communication*, 5(5), June 2017.

Ho, C. W. L., et al. (2020). Artificial Intelligence and Big Data Analytics in Health Insurance. *Bulletin of the World Health Organization*, 98, 263–269.

Househ, M. (2023). Artificial Intelligence Solutions to Detect Fraud in Healthcare Settings.

Jambukar, A. (2021). Fraudulent Healthcare Providers Detection Using Machine Learning Algorithms. MSc Research Project. National College of Ireland.

Khan Khayru, R. (2022). Transforming Healthcare: The Power of Artificial Intelligence. *Bulletin of Science, Technology and Society*, 1(3), 15-19.

Liu, Q., & Vasarhelyi, M. (2013). Healthcare Fraud Detection: A Survey and a Clustering Model Incorporating Geo-location Information. Proceedings of the 29th World Continuous Auditing and Reporting Symposium (29WCARS), Brisbane, Australia.

Nabrawi, E., & Alanazi, A. (2023). Fraud Detection in Healthcare Insurance Claims Using Machine Learning. *Risks*, 11(160).

Price II, W. N. (2017). Artificial Intelligence in Health Care: Applications and Legal Issues. *The SciTech Lawyer*, 14(1).

Rangineni, S., & Marupaka, D. (2023). Analysis of Data Engineering for Fraud Detection Using Machine Learning and Artificial Intelligence Technologies. *International Research Journal of Modernization in Engineering Technology and Science*, 05(07), July.

Yeng, P. K., et al. (2021). Artificial Intelligence–Based Framework for Analyzing Health Care Staff Security Practice: Mapping Review and Simulation Study. *JMIR Medical Informatics*, 2021.

Zhang, C., Xiao, X., & Wu, C. (2020). Medical Fraud and Abuse Detection System Based on Machine Learning. *International Journal of Environmental Research and Public Health*, 17(7265). doi:10.3390/ijerph17197265.

Appendix

The complete source code and methodologies employed in this research are accessible at "Healthcare Fraud Detection Using ML".