# Medicare Fraud Detection using Machine Learning

PRANJAL CHAUDHARI[1], PRATIBHA KOLI[2], HARSHADA MALI[3], SUMIT PAWAR[4], PROF. MANISHA PATIL[5]

[1, 2, 3, 4] *CSE (Data Science), R. C. Patel Institute of Technology, Shirpur, India.*

[5] *Department of CSE (Data Science) & AIML, R. C. Patel Institute of Technology, Shirpur, India.*

***Abstract-*** ***Medicare fraud is a significant issue that poses a threat to the integrity of the medicare system, leading to substantial financial losses and potentially compromising patient care. In response to this challenge, the utilization of machine learning models has emerged as a promising approach for detecting and preventing fraudulent activities within Medicare. This research paper proposes a machine learning approach for detecting fraud among healthcare providers. The approach involves utilizing machine learning algorithms to analyze diverse datasets containing information on billing patterns, patient demographics, service types, and geographic locations. By training the model on labelled data indicating instances of fraud, it learns to identify patterns and anomalies indicative of fraudulent behavior. Key findings from this study include the successful development of a machine learning model capable of accurately detecting healthcare provider fraud. The model demonstrates high precision, recall, and accuracy rates when tested on both training and unseen data, indicating its robustness and effectiveness.***

***Indexed Terms-*** ***Medicare fraud, Machine Learning, Fraud detection, Support Vector Machine, Logistic Regression, LightGBM, Naïve Bayes.***

## I. INTRODUCTION

Medicare fraud encompasses a range of illegal activities designed to exploit the medicare program for financial benefit this type of fraud significantly undermines the integrity of healthcare systems globally acting as a form of white-collar crime where fraudulent healthcare claims are submitted for monetary gain the consequences of medicare fraud extend beyond mere financial losses as they also strain healthcare resources and affect individual beneficiaries.

One of the most pressing issues in this domain is provider fraud, where healthcare providers submit false claims for services not rendered or for more expensive services than those provided. The government reports that Medicare fraud contributes substantially to the overall rise in healthcare costs. This type of fraud often involves complex schemes where networks of healthcare professionals and beneficiaries collude to defraud the system. Analyzing Medicare data has uncovered numerous instances of fraudulent activity among physicians who manipulate billing codes to claim higher payments for expensive treatments and medications.

Such fraudulent activities expose insurance companies and institutions to significant vulnerabilities, leading to increased insurance premiums and, consequently, higher healthcare costs for consumers. Medicare fraud and abuse manifest in various forms, with some of the most common being:

1) Submitting claims for services not actually provided.
2) Filing multiple claims for the same service.
3) Misrepresenting the nature of the service provided.
4) Upcoding, or billing for a more complex service than what was performed.
5) Billing for covered services when the actual service provided is not covered.

This research paper provides a comprehensive analysis of healthcare provider fraud detection and analysis using machine learning techniques. The study's objective is to construct and evaluate supervised machine learning models to effectively identify fraudulent healthcare providers. The datasets employed for this project, sourced from Kaggle, include various types of healthcare data distributed across four datasets. Each dataset undergoes extensive preprocessing, and new features are engineered to enhance their value, thereby improving

the accuracy of the machine learning models in addressing the research objectives.

The aim is to extract valuable insights from these datasets, offering a proactive and automated approach to fraud detection that surpasses the limitations of traditional methods. The healthcare industry is increasingly leveraging advanced technologies, particularly machine learning, to enhance its fraud detection capabilities. Machine learning models significantly improve the accuracy and efficiency of fraud detection in healthcare by identifying patterns and anomalies that are often difficult for human reviewers to detect. This project underscores the transformative potential of machine learning in fortifying healthcare fraud detection systems.

## II. RELATED WORKS

In the field of Medicare fraud, numerous studies have explored the application of machine learning to Medicare fraud detection, showcasing a variety of approaches and their effectiveness.

Bauder and Khoshgoftaar (2017) evaluated several machine learning methods for Medicare fraud detection, emphasizing the significant potential of these techniques to identify fraudulent claims effectively. Their work highlighted the applicability of models like decision trees and ensemble methods in detecting anomalies within Medicare data [1].

Johnson and Khoshgoftaar (2019) extended this research by employing neural networks to detect Medicare fraud, demonstrating the neural network's capability to handle large, complex datasets. Their study revealed the advantages of deep learning in uncovering intricate patterns indicative of fraudulent behavior, thus improving detection accuracy [2].

Detecting Healthcare Fraud and Abuse with Machine Learning and Rule-based Systems" by Zhang et al. (2019). This paper presents a hybrid approach for healthcare fraud detection, combining machine learning models and rule-based systems. The authors evaluate the performance of their system on a dataset of Medicare claims and demonstrate that their method achieves high detection accuracy with a low false positive rate.

Zhang, Xiao, and Wu (2020) developed a machine learning-based system for detecting medical fraud and abuse. They illustrated how integrating advanced algorithms can streamline the detection process, making it more efficient and effective in real-world applications. Their system significantly reduced false positives while maintaining high detection rates [3].

Lavanya et al. (2021) conducted a comparative analysis of various machine learning approaches for healthcare fraud detection. Their study compared models like logistic regression, random forests, and support vector machines, providing insights into their performance, strengths, and weaknesses in different fraud detection scenarios [4].

Jenita Mary and Angelin Claret (2022) carried out an analytical study on fraud detection in healthcare insurance claims using machine learning classifiers. Their research supported the efficacy of various classifiers, such as decision trees and ensemble methods, in effectively processing and analyzing large-scale insurance data to detect fraud [5].

Hole and Joshi (2023) performed a comprehensive analysis of provider fraud detection using machine learning. They underscored the importance of sophisticated models and feature engineering in reducing false positives and improving overall model reliability [6].

Lekkala (2023) discussed the critical role of machine learning models in healthcare fraud detection, advocating for their broader adoption. Lekkala's study emphasized the models' capability to process vast amounts of data quickly and accurately, identifying fraudulent patterns those traditional methods might miss [7].

## III. METHODOLOGY

The effectiveness of machine learning-based Medicare fraud detection significantly depends on the availability and preprocessing of relevant data. Additionally, the selection of appropriate algorithms plays a crucial role in the detection framework (Chalapathy & Chawla, 2019). Choosing the right algorithm simplifies the evaluation of model performance.

*A. Data Collection*

The initial phase of this project involved acquiring data. Publicly available datasets from the Kaggle repository were used to train and evaluate the machine learning models.

The data comprised four distinct CSV files:

Training File: This file contained the primary target variable, with each entry representing a provider identified by a unique ID and a binary label indicating potential fraud (1 for fraud providers and 0 for non-fraud providers).

Beneficiary File: This file included comprehensive information about all beneficiaries in the dataset, covering both medical and personal details.

Inpatient File: This file detailed inpatient claims and the corresponding treatments provided to hospitalized patients.

Outpatient File: This file documented outpatient claims and the treatments administered during patient visits.

These four files formed the basis for our analysis. After data acquisition, we processed and merged these files to create a comprehensive master dataset tailored for research into fraudulent provider detection.

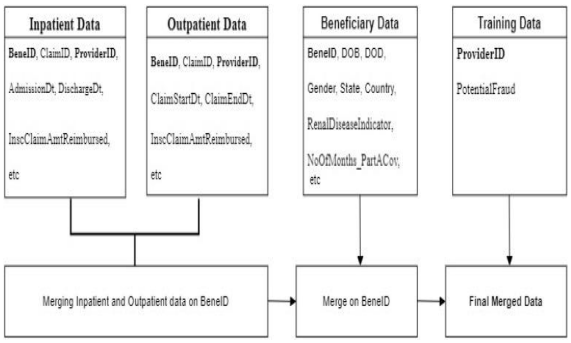


Figure 1: Dataset merging process

*B. Exploratory Data Analysis (EDA) and Data Preprocessing*

In the context of Medicare fraud detection, Exploratory Data Analysis (EDA) is crucial for examining datasets to identify patterns and anomalies that may suggest fraudulent activities. EDA involves summarizing the key characteristics of the data using descriptive statistics, visualizing feature distributions with histograms and box plots, and exploring correlations between variables through heatmaps. This process helps in understanding the data structure, detecting outliers, and uncovering insights that inform the next steps in the analysis.

Data preprocessing is essential to prepare raw data for machine learning applications. This step includes cleaning the data by addressing missing values and outliers, and transforming the data through feature engineering, encoding categorical variables, and scaling numerical features. The dataset is then split into training and testing sets to ensure robust model evaluation. Addressing class imbalance through techniques such as oversampling or undersampling is crucial for enhancing the model's capability to detect fraudulent cases. These preprocessing steps ensure the data is in an optimal format for developing effective predictive models.

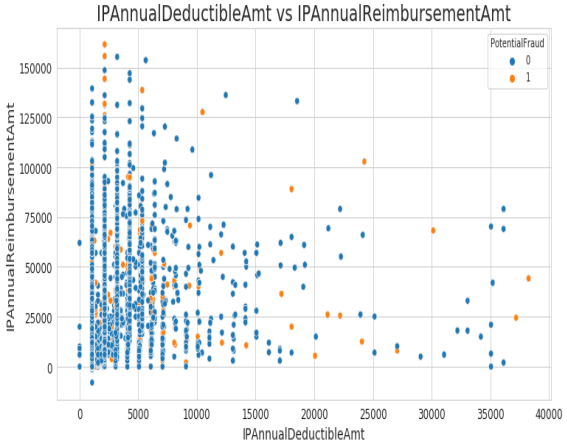


Figure 2: Inpatient Annual Deductible amount vs Reimbursement Amounts

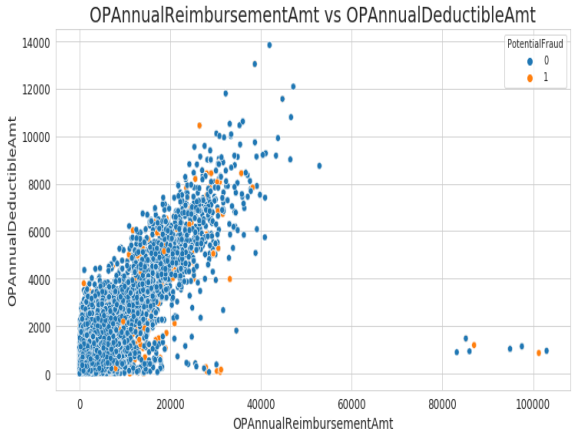


Figure 3: Outpatient Annual Deductible amount vs Reimbursement Amounts

*C. Machine Learning Algorithms*

Logistic Regression: Logistic regression is a foundational algorithm for binary classification tasks, such as distinguishing between fraudulent and non-fraudulent healthcare providers. It models the probability of a binary outcome using one or more predictor variables and is known for its simplicity and interpretability.

Naive Bayes: Naive Bayes is a probabilistic classifier based on Bayes' Theorem, assuming that predictors are independent. Despite its simplicity, it is powerful for fraud detection, particularly with categorical data. The algorithm calculates the posterior probability of a provider being fraudulent given their features, making it efficient for handling large datasets with minimal computational cost.

Decision Tree: A Decision Tree is a non-parametric, tree-structured classifier that splits the dataset into subsets based on the most significant feature at each node. Each internal node represents a decision based on a feature, while each leaf node represents an outcome.

Random Forest: Random Forest is an ensemble learning method that constructs multiple decision trees during training and aggregates their results to make a final prediction. This approach enhances predictive accuracy and controls overfitting by leveraging the collective decision-making of multiple trees.

LightGBM: LightGBM (Light Gradient Boosting Machine) is an advanced gradient boosting framework designed for high efficiency and scalability. It uses tree-based learning algorithms and offers faster training speeds and lower memory usage compared to other boosting methods. LightGBM's features, like histogram-based algorithms and leaf-wise growth, make it particularly effective for large datasets and complex classification tasks such as Medicare fraud detection.

In our study, we implemented supervised learning models using Logistic Regression, Naive Bayes, Decision Tree, Random Forest, and LightGBM to classify Medicare fraud. These algorithms were selected for their proven effectiveness in financial fraud detection. We compared their performances using metrics such as accuracy, precision, recall, and F1-score to determine the most effective model for our dataset.

*C. Evaluation Metrics*

To assess the effectiveness of our machine learning models in detecting Medicare fraud, we relied on several standard evaluation metrics:

Receiver Operating Characteristic (ROC) Curve: The ROC curve visually represents the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) across various threshold settings. A model with a higher ROC curve, closer to the upper left corner, demonstrates superior discriminatory ability between fraudulent and non-fraudulent providers.

Confusion Matrix: The confusion matrix offers a detailed breakdown of the model's performance by presenting the counts of true positives, true negatives, false positives, and false negatives. It serves as a valuable tool for assessing the model's ability to accurately classify instances of fraud and non-fraud.

Test AUC (Area Under the Curve): The Test AUC quantifies the model's discriminative ability on the test data. A higher AUC value signifies better performance in distinguishing between fraud and non-fraud cases. Proper interpretation of the AUC score is crucial for understanding the model's overall effectiveness.

Test F1-Score: The Test F1-Score balances the trade-off between precision and recall on the test data. It provides a single metric that encapsulates the model's ability to correctly identify fraudulent cases while minimizing false alarms.

By analyzing these evaluation metrics, we gain insights into the strengths and weaknesses of our machine learning models in Medicare fraud detection. This information guides us in making informed decisions about model selection and deployment in real-world scenarios.

## IV. RESULTS AND DISCUSSION

### *A. Model Performance*

The performance of five machine learning models - Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, and LightGBM - was evaluated for Medicare fraud detection. The models were assessed using Test AUC and Test F1-score to understand their accuracy and ability to balance between precision and recall. Below the comparative table showcasing the Test AUC and Test F1-scores, we observe notable differences in model effectiveness. LightGBM outperformed all other models with an AUC of 0.84105 and a F1-score of 0.71214, indicating excellent predictive accuracy and balance between precision and recall.

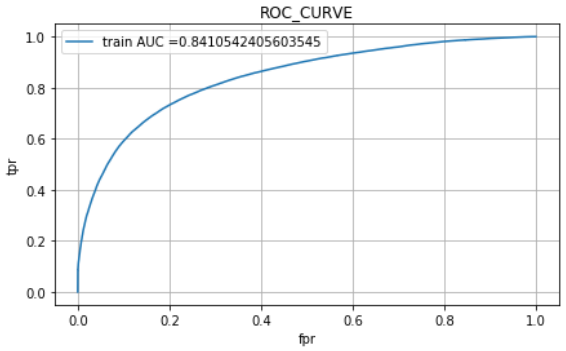


Figure 4: ROC Curve

Table 1: Comparison Between Different Models on AUC and F1 Score

| Model | AUC | F1 |
|---|---|---|
| Logistic Regression | 0.73514 | 0.62004 |
| Naïve Bayes | 0.51854 | 0.55965 |
| Decision Tree | 0.80175 | 0.68137 |
| Random Forest | 0.73544 | 0.60794 |
| LightGBM | 0.84105 | 0.71214 |

### *B. Discussion*

The LightGBM model demonstrated the highest performance, indicating its strong capability in identifying fraudulent activities with a good balance between precision and recall. Decision Tree also performed well, offering both accuracy and interpretability. Logistic Regression and Random Forest provided moderate results, suitable for baseline comparisons and further optimization. Naïve Bayes, however, showed limited effectiveness, suggesting that its assumptions do not align well with the complexities of the Medicare fraud dataset. Implementing the best-performing models, particularly LightGBM, could significantly enhance fraud detection efforts, leading to substantial cost savings and improved healthcare system integrity. Future work should focus on integrating additional data sources and advanced techniques to further refine model performance and adaptability.

## CONCLUSION

In this study, we developed and evaluated various machine learning models to detect Medicare fraud, including Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, and LightGBM. Among these models, LightGBM demonstrated the highest performance, showcasing its robust capability in accurately identifying fraudulent providers while maintaining a balanced approach between precision and recall. Decision Tree also exhibited commendable performance, offering both accuracy and interpretability.

Our findings underscore the significance of leveraging advanced machine learning techniques for effective fraud detection, which can substantially mitigate financial losses and enhance the integrity of Medicare systems. Future research endeavors should explore alternative advanced models, delve into the application of deep learning techniques, and experiment with different architectural configurations such as varying numbers of layers, dropout rates, filter sizes, and max-pooling layers to further optimize model performance.

Implementing the best-performing models, particularly LightGBM, holds the potential to yield significant cost savings and streamline healthcare services by minimizing false positives and accurately identifying fraudulent activities. This could lead to enhanced efficiency and integrity within the Medicare system, ultimately benefiting both providers and beneficiaries.

REFERENCES

[1] Bauder, R. A. and Khoshgoftaar, T. M., Medicare fraud detection using machine learning methods, 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 858–865, 2017.
[2] Johnson, J. M., & Khoshgoftaar, T. M., Medicare Fraud Detection Using Neural Networks. Journal of Big Data, 6, Article No. 63, 2019.
[3] Conghai Zhang, Xinyao Xiao and Chao Wu, Medical Fraud and Abuse Detection System Based on Machine Learning. Int. J. Environ. Res. Public Health 2020.
[4] S. Lavanya1, S. Manoj Kumar, P. Mohan Kumar. Machine Learning Based Approaches for Healthcare Fraud Detection: A Comparative Analysis. Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021.
[5] A. Jenita Mary, S. P. Angelin Claret. Analytical study on fraud detection in healthcare insurance claim data using machine learning classifiers, AIP Conf. Proc. 2516, 240006, 2022.
[6] Hole Prajakta Parshuram, Prof. S. G. Joshi. A Comprehensive Analysis of Provider Fraud Detection through Machine Learning. International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), Volume 3, Issue 2, 2023.
[7] Lekkala, L. R., Importance of Machine Learning Models in Healthcare Fraud Detection. Voice of the Publisher, 9, 207-215, 2023.
[8] https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis/data.