# MAR ATHANASIUS COLLEGE OF ENGINEERING, KOTHAMANGALAM

## Initial Project Report

## HEALTHCARE FRAUD DETECTION USING MACHINE LEARNING


Done by

**ARCHANA P**

Reg No: MAC23MCA-2016


Under the guidance of

Prof. Manu John

# ABSTRACT

The Healthcare Fraud Detection Using Machine Learning project is focused on utilizing artificial intelligence techniques to detect healthcare fraud by analyzing claim data. Healthcare fraud is a complex issue that includes several fraudulent tactics used to get unjustified advantages from healthcare institutions. This project offers a more accurate and scalable solution for the problem. The relevance of this project lies in its potential to enhance the accuracy and efficiency of fraud detection, thereby reducing financial losses and improving the integrity of healthcare systems.

A comprehensive literature survey reveals that traditional fraud detection methods are often time-consuming and ineffective. Studies show that AI models are superior in detecting complex fraud patterns and improving detection rates, making them ideal for healthcare fraud detection.

The proposed system employs a methodical approach to training, validating, and testing machine learning models. The study employs machine learning algorithms such as Random Forest and Light GBM, trained on a comprehensive medicare claims dataset sourced from Kaggle, which includes inpatient, outpatient, and beneficiary information. The dataset contains several thousand records, making it suitable for training robust machine learning models. The preprocessing steps include data cleaning to handle missing values and data transformation to encode categorical variables into numerical values. The models are trained, validated, and tested using an 80-20 split of the data. The performance of these algorithms is then compared to see which algorithm gives the best prediction.

The dataset utilized in this study is sourced from the Kaggle platform and consists of Medicare claims data, encompassing information on inpatient and outpatient visits, as well as beneficiary details. The dataset includes various features, such as admission and discharge dates, diagnosis codes, and procedures performed, which are essential for predicting fraud.The project aims to predict whether a claim is fraudulent or not. By using machine learning techniques, it helps to reduce the number of frauds in the health sector. The expected outcome is more accurate fraud detection and improved healthcare efficiency.


**References:**


1. Singh K. HEALTHCARE FRAUDULENCE: LEVERAGING ADVANCED ARTIFICIAL INTELLIGENCE TECHNIQUES FOR DETECTION International Research Journal of Modernization in Engineering Technology and Science, 6 (2), 966-976 https://www.doi.org/10.56726.IRJMETS49394.2024

2. Pranjal Chaudhari ; Pratibha Koli ; Harshada Mali ; Sumit Pawar ; Prof. Manisha Patil . "Medicare Fraud Detection using Machine Learning" Iconic Research And Engineering Journals Volume 7 Issue 11 2024 Page 650-655.

3. Yoo Y, Shin J, Kyeong S. Medicare fraud detection using graph analysis: a comparative study of machine learning and graph neural networks. IEEE Access. 2023 Aug 17. https://www.doi.org/10.1109/ACCESS.2023.3305962



Submitted By:
Archana P
Reg No: MAC23MCA-2016
2023 –25 Batch MCA Department, MACE

Faculty Guide:
Prof. Manu John
Associate Professor S3 MCA

# INTRODUCTION

Healthcare fraud is a significant issue, encompassing various deceptive practices like billing for unprovided services, upcoding, and unbundling. Such fraudulent activities not only strain financial resources but also degrade the quality of patient care by diverting funds from legitimate medical needs. The National Healthcare Anti-Fraud Association (NHCAA) estimates that healthcare fraud costs the industry over $68 billion annually in the United States, highlighting the critical need for effective detection methods. Traditional fraud detection methods, including manual reviews and rule-based systems, are often labor-intensive and fail to identify sophisticated fraud schemes.

The proposed system follows a structured approach to develop and evaluate machine learning models. Machine learning algorithms like Random Forest and Light GBM are employed to predict whether a claim is fraudulent. The goal of the project is to utilize these techniques to minimize fraud in the healthcare sector. By analyzing the performance of these algorithms, the study aims to identify which model provides the most accurate predictions.

The proposed system follows a structured approach to develop and evaluate machine learning models. The process begins with training and validating models using a comprehensive dataset from Kaggle, which includes Medicare claims data. The data is cleaned to handle missing values and transformed to convert categorical variables into numerical ones. The models are then trained and tested using an 80-20 split of the data, allowing for effective performance comparison.

The dataset used in this study, titled "Healthcare Provider Fraud Detection," is a comprehensive collection of data related to Medicare claims. It was sourced from Kaggle, a popular platform for data science competitions and datasets. InpatientData.csv contains information about claims filed for patients who were admitted to hospitals. It includes details such as patient admission and discharge dates, diagnosis codes, and the procedures performed during the hospital stay. OutpatientData.csv includes information about claims filed for patients who visited hospitals but were not admitted. It captures data on the patient's date of service, diagnosis codes, and the procedures performed during their outpatient visits. BeneficiaryData.csv provides beneficiary Know Your Customer (KYC) details, including health conditions and regional information.

The study's results, visualized through various data visualization techniques, demonstrate the models effectiveness in detecting fraudulent claims. Evaluation metrics like accuracy and F1-score show that AI models significantly outperform traditional methods. However, the research also acknowledges the potential for false positives and negatives, emphasizing the need for continuous monitoring and adaptation of these models in dynamic healthcare fraud scenarios. The expected outcome is improved fraud detection and increased efficiency within healthcare systems.

In conclusion the use of this dataset in training machine learning models highlights the potential of AI in revolutionizing healthcare fraud detection. The dataset's comprehensive nature and detailed preprocessing steps facilitated the development of robust models capable of accurately identifying fraudulent claims, significantly outperforming traditional fraud detection methods.

# LITERATURE REVIEW

**Paper 1: Healthcare Fraudulence: Leveraging Advanced Artificial Intelligence Techniques For Detection**

Traditional fraud detection methods like manual reviews and rule-based systems are slow and often miss evolving fraud schemes. AI techniques offer a more dynamic and adaptive solution. AI models such as Logistic Regression, Decision Tree, and Random Forest are better at detecting complex fraud patterns. This makes AI crucial for improving healthcare fraud detection.

| | |
|---|---|
| **Title of the paper** | Singh K. HEALTHCARE FRAUDULENCE: LEVERAGING ADVANCED ARTIFICIAL INTELLIGENCE TECHNIQUES FOR DETECTION International Research Journal of Modernization in Engineering Technology and Science, 6 (2), 966-976, 2024 |
| **Area of work** | The project focuses on the application of artificial intelligence techniques to detect fraud in healthcare claims. |
| **Dataset** | BeneficiaryData.csv contains 25 features<br>InpatientData.csv contains 30 features<br>OutpatientData.csv contains 27 features |
| **Methodology/Strategy** | The "Healthcare Fraud Detection using Machine Learning" project uses AI techniques like Logistic Regression, Decision Tree, and Random Forest to detect healthcare fraud. The model is trained on a large Medicare claims dataset from Kaggle, the models undergo data cleaning, transformation, and feature selection. Logistic Regression models fraud probability, Decision Trees maximize information gain, and Random Forests combine predictions for accuracy. Evaluated using accuracy and F1-score, Random Forests perform best. The input includes claims data with details like admission dates and diagnosis codes. The goal is to accurately identify fraudulent claims through careful preprocessing and fine-tuning. |
| **Algorithm** | Logistic Regression , Decision Tree , Random Forest. |
| **Result/Accuracy** | **Random Forest: 95.17%**<br>Logistic Regression: 92.69%<br>Decision Tree:92.59% |
| **Advantages** | Enhance accuracy and efficiency by balancing the identification of fraudulent claims and maintain the integrity and trust of the healthcare system |
| **Future Proposal** | Broader applications in various healthcare-related fraudulent activities.<br>Improving Model Accuracy |

**Paper 2: Medicare Fraud Detection using Machine Learning**

The paper "Medicare Fraud Detection using Machine Learning" addresses Medicare fraud, causing financial losses and harming healthcare integrity. It uses machine learning to analyze data and detect fraud, finding models like Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, and LightGBM effective, with LightGBM being the most accurate. This research shows machine learning can improve fraud detection.

| | |
|---|---|
| **Title of the paper** | Pranjal Chaudhari ; Pratibha Koli ; Harshada Mali ; Sumit Pawar ; Prof. Manisha Patil . "Medicare Fraud Detection using Machine Learning" Iconic Research And Engineering Journals Volume 7 Issue 11 2024 Page 650-655 |
| **Area of work** | The researchers analyze data like billing patterns, patient demographics, and service types to identify fraudulent activities. By training the model with known instances of fraud, they develop an effective tool for accurately detecting fraud in healthcare. |
| **Dataset** | Publicly available datasets from the Kaggle repository were used to train and evaluate the machine learning models |
| **Methodology/Strategy** | The “Medicare Fraud Detection using Machine Learning” project uses AI techniques like Logistic Regression, Naive Bayes , Light GBM , Decision Tree, and Random Forest to detect healthcare fraud. The models undergo data cleaning 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 |
| **Algorithm** | Logistic Regression , Decision Tree , Random Forest , Naive Bayes , Light GBM |
| **Result/Accuracy** | **Light GBM: 84.10%**<br>Random Forest: 73.54%<br>Decision Tree: 80.17%<br>Naive Bayes: 51.85%<br>Logistic Regression: 73.51% |
| **Advantages** | The study highlighted that advanced machine learning models like LightGBM significantly enhance fraud detection efforts by maintaining a balanced approach between precision and recall, leading to substantial cost savings and improved healthcare system integrity. |
| **Limitations** | The primary limitation noted was the moderate performance of some models, such as Logistic Regression and Random Forest, and the limited effectiveness of Naive Bayes, suggesting its assumptions do not align well with the complexities of the Medicare fraud dataset |
| **Future Proposal** | Integrating additional data sources. Exploring advanced techniques and models, including deep learning. |

**Paper 3: Medicare Fraud Detection Using Graph Analysis: A comparative study of Machine Learning and Graph Neural Networks**

The paper titled "Medicare Fraud Detection Using Graph Analysis: A Comparative Study of Machine Learning and Graph Neural Networks" focuses on the detection of Medicare fraud, which is a significant issue that causes substantial financial losses annually. The study emphasizes the use of graph analysis that extracts graph information from a network among providers, physicians, and beneficiaries and uses this information as a feature of conventional machine learning algorithms, such as logistic regression, XGBoost, and multi-layer perceptron

| | |
|---|---|
| **Title of the paper** | Yoo Y, Shin J, Kyeong S. Medicare fraud detection using graph analysis: a comparative study of machine learning and graph neural networks. IEEE Access. 2023 Aug 17. |
| **Area of Work** | The area of work in this project focuses on developing and implementing an advanced machine learning technique to detect fraud in healthcare insurance claims. |
| **Dataset** | The study uses open-source tabular datasets that include beneficiary information, inpatient and outpatient claims, and indications of potentially fraudulent providers. |
| **Methodology/Strategy** | To develop a Medicare fraud detection model, we used two approaches: GNN algorithms and traditional machine learning with graph features. For the GNN approach, we converted the tabular data into graph-structured data and developed GNN models for node classification to detect fraudulent providers. For the machine learning approach, we extracted relationships between Providers-Physician and Providers-Beneficiary, created bipartite graphs, and used graph centrality information as input features for conventional machine learning algorithms. |
| **Algorithm** | Logistic Regression, Random Forest , Light GBM, XGBoost, GNN Models |
| **Result/Accuracy** | **Light GBM: 83%**<br>Random Forest: 74%<br>Logistic Regression: 76%<br>XGBoost: 82%<br>GNN Model: 73.99% |
| **Advantages** | Improved Detection Performance: The use of graph centrality measures in machine learning models significantly improves fraud detection accuracy.<br>Efficiency: The traditional machine learning models with graph features require much less computational power and time compared to GNNs. |
| **Limitations** | Computational Burden of GNNs and the Complexity of Graph Construction is a major problem. |
| **Future Proposal** | Focusing on improving the detection system further and exploring other machine learning techniques. This could help in making fraud detection even better and faster, enhancing the overall effectiveness of the system |

## LITERATURE SUMMARY

| | TITLE | DATASET | ALGORITHM | ACCURACY |
|---|---|---|---|---|
| **PAPER 1** | . Singh K.HEALTHCARE FRAUDULENCE:LEVERAGING ADVANCED ARTIFICIAL TECHNIQUES FOR DETECTION International Research Journal Of Modernization in Engineering Technology And Science, 6(2),966-976 | BeneficiaryData.csv contains 25 features<br>InpatientData.csv contains 30 features<br>OutpatientData.csv contains 27 features | Random Forest<br>Logistic Regression<br>Decision Tree | **95.17%**<br>92.67%<br>92.59% |
| **PAPER 2** | Pranjal Chaudhari ; Pratibha Koli ; Harshada Mali ; Sumit Pawar ; Prof. Manisha Patil . "Medicare Fraud Detection using Machine Learning" Iconic Research And Engineering Journals Volume 7 Issue 11 2024 Page 650-655 | Publicly available datasets from the Kaggle repository were used to train and evaluate the machine learning models. | Light GBM<br>Random Forest<br>Naive Bayes<br>Decision Tree<br>Logistic Regression | **84.10%**<br>73.54%<br>51.85%<br>80.17%<br>73.51% |
| **PAPER 3** | Yoo Y, Shin J, Kyeong S. Medicare fraud detection using graph analysis: a comparative study of machine learning and graph neural networks. IEEE Access. 2023 Aug 17. | The study uses open-source tabular datasets that include beneficiary information, inpatient and outpatient claims, and indications of potentially fraudulent providers. | Light GBM<br>Random Forest<br>GNN Model<br>XGBoost<br>Logistic Regression | **83%**<br>74%<br>73.99%<br>82%<br>76% |

# PROPOSED MODEL

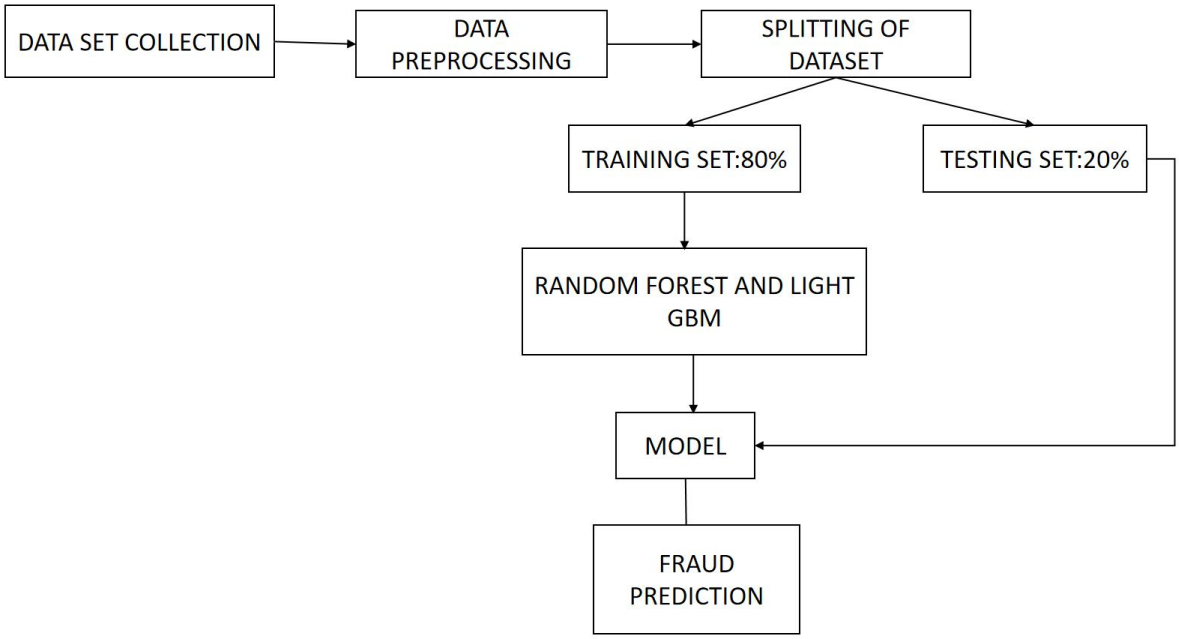
**Healthcare Fraud Detection Using Machine Learning**

From the above three papers, we get to know that different approaches are used for Healthcare Fraud Detection. First paper is the detection of fraud claims using selective features, the second paper uses more machine learning algorithms to detect fraud claims and the third paper aims at a hybrid approach that combines data processing with a specialized version of the support vector machine algorithm .

The proposed system employs both Random Forest and LightGBM algorithms to detect healthcare fraud. Their performance is compared using key metrics such as accuracy, precision, recall, and F1-score. This comparison aims to identify which algorithm performs better and provides the most accurate predictions. By evaluating both models, we aim to ensure the system's robustness and reliability in detecting fraudulent claims.

Using feature engineering and a correlation matrix helped make the model more accurate and less likely to overfit. AI models are better than traditional methods for detecting fraud because they can adapt and find complex patterns. However, they can still make mistakes, so they need constant monitoring and adjustment. This research shows that advanced AI can greatly improve healthcare fraud detection, but it needs regular updates to stay effective.

**PIPELINE DIAGRAM:**



# DATASET DESCRIPTION

**Dataset Overview:**

The dataset for this project is sourced from Kaggle's "Healthcare Provider Fraud Detection" dataset, the data is divided into three main parts:

InpatientData.csv contains information about claims filed for patients who were admitted to hospitals. This data includes the patient's admission and discharge dates, the diagnosis code, and the procedures performed.

OutpatientData.csv contains information about claims filed for patients who visited hospitals but were not admitted. This data includes the patient's date of service, the diagnosis code, and the procedures performed.

BeneficiaryData.csv contains beneficiary details like health conditions, region they belong to, etc.

**Source:**

The "Healthcare Provider Fraud Detection" dataset is publicly available on Kaggle.

Dataset Link : https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis

| DOB | DOD | Gender | Race | RenalDisea: | State | County |
|---|---|---|---|---|---|---|
| 01-01-1943 | NA | 1 | 1 | 0 | 39 | 230 |
| 01-09-1936 | NA | 2 | 1 | 0 | 39 | 280 |
| 01-08-1936 | NA | 1 | 1 | 0 | 52 | 590 |
| 01-07-1922 | NA | 1 | 1 | 0 | 39 | 270 |
| 01-09-1935 | NA | 1 | 1 | 0 | 24 | 680 |

**Features:**

Beneficiary data.csv contains columns such as ['BeneID', 'DOB', 'DOD', 'Gender', 'Race', 'RenalDiseaseIndicator', 'State', 'County', 'NoOfMonths_PartACov', 'NoOfMonths_PartBCov', 'ChronicCond_Alzheimer', 'ChronicCond_Heartfailure', 'ChronicCond_KidneyDisease', 'ChronicCond_Cancer', 'ChronicCond_ObstrPulmonary', 'ChronicCond_Depression', 'ChronicCond_Diabetes', 'ChronicCond_IschemicHeart', 'ChronicCond_Osteoporasis', 'ChronicCond_rheumatoidarthritis', 'ChronicCond_stroke', 'IPAnnualReimbursementAmt', 'IPAnnualDeductibleAmt']

InpatientData.csv contains columns such as ['Provider', 'InscClaimAmtReimbursed', 'AttendingPhysician', 'OperatingPhysician', 'OtherPhysician', 'ClmDiagnosisCode_1', 'ClmDiagnosisCode_2', 'ClmDiagnosisCode_3', 'ClmDiagnosisCode_4', 'ClmDiagnosisCode_5', 'ClmDiagnosisCode_6', 'ClmDiagnosisCode_7', 'ClmDiagnosisCode_8', 'ClmDiagnosisCode_9', 'ClmDiagnosisCode_10', 'ClmProcedureCode_1',

'ClmProcedureCode_2', 'ClmProcedureCode_3', 'ClmProcedureCode_4', 'ClmProcedureCode_5', 'ClmProcedureCode_6', 'DeductibleAmtPaid', 'ClmAdmitDiagnosisCode']

OutpatientData.csv contains columns such as ['Provider', 'InscClaimAmtReimbursed', 'AttendingPhysician', 'OperatingPhysician', 'OtherPhysician', 'ClmDiagnosisCode_1', 'ClmDiagnosisCode_2', 'ClmDiagnosisCode_3', 'ClmDiagnosisCode_4', 'ClmDiagnosisCode_5', 'ClmDiagnosisCode_6', 'ClmDiagnosisCode_7', 'ClmDiagnosisCode_8', 'ClmDiagnosisCode_9', 'ClmDiagnosisCode_10', 'ClmProcedureCode_1', 'ClmProcedureCode_2', 'ClmProcedureCode_3', 'ClmProcedureCode_4', 'ClmProcedureCode_5', 'ClmProcedureCode_6', 'DeductibleAmtPaid', 'ClmAdmitDiagnosisCode']

# CONCLUSION

In the end, this study supports the idea that AI has the potential to revolutionize fraud detection in the healthcare industry and beyond.These models may strike a delicate balance between recognizing fraudulent claims and limiting false positives, according to a thorough review of performance indicators, including accuracy, precision, recall, and F1-score.The foundation for more reliable fraud detection systems has been created by choosing pertinent characteristics and improving the dataset.