International Journal of Multidisciplinary Engineering in Current Research - IJMEC
Volume 11, Issue 4s, April-2026, http://ijmec.com/, ISSN: 2456-4265

Full Length Article

Enhancing Medicare Fraud Detection Through Machine Learning

Rafath Tabrez[1], Loqman Uzair[2], Ms.Imreena Ali[3]
[1,2]B.E students; Department Of Computer Science Engineering, ISLEC, Hyderabad India.
[3]Assistant Professor Department Of Computer Science Engineering, ISLEC, Hyderabad India.
Mail id: tabrezz1437@gmail.com
Accepted 24-04-2026

Abstract
Medicare fraud is one of the major challenges faced by the healthcare industry, leading to significant financial losses and reduced trust in healthcare systems. Traditional fraud detection methods mainly rely on manual auditing and rule-based systems, which are often inefficient in detecting complex and evolving fraudulent activities. This project proposes an intelligent Medicare fraud detection system using machine learning techniques to improve the accuracy and efficiency of identifying fraudulent healthcare claims. The system analyzes historical medical claim data and applies various machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine to classify claims as fraudulent or legitimate. Data preprocessing, feature selection, and class balancing techniques are used to enhance model performance. The proposed model helps in detecting suspicious billing patterns, abnormal claim behavior, and provider fraud with reduced human intervention. Experimental results demonstrate that machine learning-based approaches provide higher accuracy, faster detection, and better scalability compared to traditional methods. The developed system can assist healthcare organizations and insurance providers in minimizing financial losses and improving the reliability of Medicare services.

Keywords: *Medicare Fraud Detection, Machine Learning, Healthcare Analytics, Fraud Prevention, Random Forest, Decision Tree, Healthcare Claims, Artificial Intelligence, Data Mining, Predictive Analytics.*

Introduction:
Healthcare fraud detection plays a vital role in safeguarding healthcare systems against financial exploitation, which can lead to unnecessary costs, compromised patient care, and erode trust in medical institutions. Fraudulent claims, such as billing for services that were never rendered or providing unnecessary services to increase reimbursement, contribute significantly to healthcare expenditures. However, detecting these fraudulent activities is increasingly challenging due to the vast volume and complexity of healthcare data. With large datasets containing millions of claims, the fraudulent claims are often a small fraction compared to legitimate claims, creating an imbalanced distribution that complicates the identification of fraud. Traditional machine learning techniques such as Random Forests, Decision Trees, and Logistic Regression have been applied to tackle healthcare fraud detection, but they are not without their limitations.

Many of these models struggle with the imbalanced dataset issue, where the majority class (non-fraudulent claims) outnumbers the minority class(fraudulent claims), leading to biased predictions.

This imbalance can result in models that fail to accurately detect fraud, as the algorithms tend to favor the majority class, leading to a high number of false negatives (fraudulent cases incorrectly identified as non-fraudulent).To mitigate the challenges posed by imbalanced data, various techniques like Random Oversampling (ROS), Synthetic Minority Oversampling Technique (SMOTE), and Random Undersampling (RUS) have been proposed. ROS increases the number of minority class instances by duplicating existing ones, but it risks overfitting by introducing duplicate information. SMOTE creates synthetic instances of the minority class, but it can introduce noise or outliers. RUS reduces the majority class, potentially losing valuable information. While these methods help balance the dataset, they come with their own set of limitations that can undermine the overall performance of the fraud detection models. In this study, we present a novel approach to healthcare fraud detection that directly addresses the challenges of dataset imbalance by employing a hybrid resampling technique called SMOTE-ENN (Synthetic Minority Oversampling Technique - Edited Nearest Neighbors). SMOTE-ENN combines the strengths of SMOTE for generating synthetic minority class instances with Edited Nearest Neighbors (ENN) to remove noisy and redundant instances from the dataset.

https://doi.org/10.63665/IJMEC.1104s.49
ISSN: 2456-4265
IJMEC 2026

Literature Review

Healthcare fraud has become a major challenge in modern medical systems, especially in Medicare insurance programs where false claims and illegal billing practices cause significant financial losses. Traditional fraud detection methods mainly depend on manual auditing and rule-based systems, which are time-consuming and less effective for identifying complex fraud patterns. Researchers have therefore focused on applying Machine Learning (ML) techniques to improve the accuracy and efficiency of fraud detection.

Several studies have used supervised learning algorithms such as Decision Trees, Random Forest, Logistic Regression, and Support Vector Machines (SVM) for detecting fraudulent medical claims. These algorithms analyze historical claim data and classify transactions as genuine or fraudulent based on patterns and behaviors. Among these methods, Random Forest and Decision Tree algorithms have shown high accuracy because they can handle large datasets and complex relationships between variables.

Deep learning and Artificial Neural Networks (ANN) have also been explored for Medicare fraud detection. These methods can identify hidden patterns in large healthcare datasets and improve prediction performance. However, deep learning models require high computational power and large amounts of training data.

Researchers have further applied anomaly detection and unsupervised learning techniques to identify unusual claim activities without labeled fraud data.
Clustering algorithms such as K-Means help group similar claim patterns and detect abnormal provider behavior. These methods are useful in real-world healthcare systems where fraud cases are rare and continuously changing.

Recent studies emphasize the importance of feature engineering and data preprocessing in improving fraud detection accuracy.
Features such as billing amount, frequency of claims, patient history, and provider behavior are commonly used for model training. Data balancing techniques like SMOTE are also applied to overcome class imbalance problems because fraudulent claims are much fewer than legitimate claims.

Overall, the literature shows that Machine Learning techniques significantly improve Medicare fraud detection by increasing detection accuracy, reducing manual effort, and identifying hidden fraud patterns. The combination of predictive analytics, big data processing, and intelligent algorithms provides an effective solution for modern healthcare fraud prevention systems.

Methodology

The proposed system for enhancing Medicare fraud detection uses Machine Learning techniques to identify fraudulent healthcare claims accurately and efficiently. The methodology consists of several stages including data collection, preprocessing, feature extraction, model training, prediction, and result evaluation.

Initially, healthcare claim datasets are collected from Medicare records and related healthcare sources. The dataset contains information such as patient details, billing amount, diagnosis codes, treatment procedures, provider information, and claim history. Since raw medical data may contain missing values, duplicate records, and inconsistent formats, data preprocessing is performed to clean and normalize the dataset.

After preprocessing, feature extraction and selection techniques are applied to identify important attributes related to fraudulent activities. Features such as claim amount, frequency of claims, unusual billing patterns, and provider behavior are used for analysis. Data balancing methods like SMOTE may also be used to handle class imbalance between fraudulent and non-fraudulent claims.

The processed data is then divided into training and testing datasets. Machine Learning algorithms such as Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM) are trained using the training data. These algorithms learn patterns from historical claim records and classify claims as legitimate or fraudulent.

During the prediction phase, the trained model analyzes new healthcare claims and detects suspicious activities based on learned patterns. The system then generates fraud alerts for high-risk claims, helping healthcare authorities take preventive actions quickly.

Finally, the performance of the model is evaluated using parameters such as accuracy, precision, recall, and F1-score. The algorithm with the best performance is selected for the final fraud detection system. This methodology helps improve fraud detection efficiency, reduce financial losses, and support secure healthcare management systems.

https://doi.org/10.63665/IJMEC.1104s.49
ISSN: 2456-4265
IJMEC 2026

Implementation

The implementation of the Medicare fraud detection system is carried out using Machine Learning algorithms and healthcare claim datasets. The system is developed in a structured manner to identify fraudulent medical claims efficiently.

First, the healthcare dataset is imported into the system using Python programming and data analysis libraries such as Pandas and NumPy. Data preprocessing techniques are applied to remove missing values, duplicate records, and unwanted data. The cleaned dataset is then transformed into a suitable format for Machine Learning model training.

After preprocessing, important features such as claim amount, provider details, billing frequency, diagnosis codes, and patient history are selected. The dataset is divided into training and testing sets to evaluate the model performance accurately.

Machine Learning algorithms such as Decision Tree, Random Forest, Logistic Regression, and Support Vector Machine (SVM) are implemented using Scikit-learn libraries. The models are trained using historical claim data to learn fraudulent and non-fraudulent patterns.

Once training is completed, the trained model predicts whether a new healthcare claim is legitimate or fraudulent. The prediction results are analyzed using evaluation metrics like accuracy, precision, recall, and F1-score. The algorithm with the highest accuracy and better fraud detection capability is selected as the final model.

The implementation helps automate fraud detection, reduce manual auditing work, improve detection speed, and minimize financial losses in Medicare healthcare systems.

Algorithm and Methodology

The proposed Medicare fraud detection system begins by importing the healthcare claim dataset, which contains information related to patient claims, billing details, and provider activities. Data preprocessing is then performed to clean and prepare the dataset by handling missing values, removing inconsistencies, and transforming the data into a suitable format for analysis. After preprocessing, important features are selected to identify the most relevant attributes contributing to fraud detection.

The dataset is then divided into training and testing sets to enable model development and evaluation. Various Machine Learning algorithms are trained using the training dataset, and the trained models are subsequently tested on unseen data. Based on the learned patterns, the system predicts whether a healthcare claim is fraudulent or non-fraudulent. The performance of the models is evaluated using standard metrics such as accuracy, precision, recall, and F1-score. Finally, the fraud detection results are displayed, and the process concludes.

Flowchart Description

The workflow of the proposed system follows a sequential process. It starts with importing the healthcare claim dataset, followed by data

https://doi.org/10.63665/IJMEC.1104s.49
ISSN: 2456-4265
IJMEC 2026

preprocessing and feature selection. The processed data is then split into training and testing datasets. Machine Learning models are trained and tested using the prepared data. The system then performs fraud prediction, evaluates model performance, displays the detection results, and finally stops.

Results and Performance Analysis

The proposed Medicare fraud detection system was tested using healthcare claim datasets to analyze the effectiveness of different Machine Learning algorithms in detecting fraudulent claims. The system successfully classified both fraudulent and non-fraudulent claims with high accuracy, thereby reducing the need for manual verification.

Among the implemented algorithms, the Random Forest model achieved the best overall performance because of its ability to handle large datasets and capture complex fraud patterns effectively. The system analyzed factors such as billing behavior, claim frequency, and provider activities to identify suspicious transactions.

The comparative analysis showed that Logistic Regression achieved an accuracy of 89%, precision of 87%, recall of 85%, and an F1-score of 86%. The Decision Tree algorithm improved the performance with 91% accuracy, 89% precision, 88% recall, and 88% F1-score. Support Vector Machine (SVM) further enhanced the results by obtaining 93% accuracy, 91% precision, 90% recall, and 90% F1-score. However, Random Forest demonstrated the highest efficiency among all models, producing the most reliable fraud detection results.

The evaluation metrics, including accuracy, precision, recall, and F1-score, showed significant improvement compared to traditional rule-based fraud detection approaches. The proposed model effectively minimized false predictions and improved the overall reliability and robustness of healthcare fraud identification.

The final results demonstrate that Machine Learning techniques can effectively improve Medicare fraud detection by increasing accuracy, reducing financial losses, and supporting faster healthcare claim verification processes.

Conclusion

The project "Enhancing Medicare Fraud Detection through Machine Learning" successfully demonstrates the use of Machine Learning techniques for identifying fraudulent healthcare claims. Traditional fraud detection methods are often slow and less effective in detecting complex fraud patterns, whereas Machine Learning models provide faster and more accurate fraud identification.

In this project, various algorithms such as Logistic Regression, Decision Tree, Support Vector Machine (SVM), and Random Forest were implemented and evaluated using healthcare claim datasets. Among these algorithms, the Random Forest model achieved the highest accuracy and showed better performance in detecting suspicious claim activities.

The proposed system improves fraud detection efficiency by analyzing large volumes of healthcare data, reducing manual effort, minimizing false claims, and helping healthcare organizations prevent financial losses. The project also highlights the importance of data preprocessing, feature selection, and model evaluation in building an effective fraud detection system.

Overall, the implementation of Machine Learning in Medicare fraud detection provides a reliable, intelligent, and scalable solution for modern healthcare systems.

Future improvements may include the use of deep learning techniques, real-time fraud monitoring, and integration with big data technologies for enhanced performance and security.

References

1) Machine Learning, McGraw-Hill Education, 1997.
2) Pattern Recognition and Machine Learning, Springer Publication, 2006.
3) Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers, 2016.
4) Introduction to Machine Learning with Python, O'Reilly Media, 2016.
5) Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, O'Reilly Media, 2019.
6) Data Mining Concepts and Techniques, Elsevier Publication, 2011.
7) Artificial Intelligence: A Modern Approach, Pearson Education, 2021.
8) Python Machine Learning, Packt Publishing, 2019.
9) Deep Learning, MIT Press, 2016.
10) Applied Predictive Modeling, Springer Publication, 2013.
11) Healthcare Fraud Detection Using Machine Learning, International Journal of Advanced Computer Science and Applications (IJACSA), 2020.

https://doi.org/10.63665/IJMEC.1104s.49
ISSN: 2456-4265
IJMEC 2026

12) Medical Insurance Fraud Detection Using Data Mining Techniques, IEEE International Conference on Computing and Communication Technologies, 2019.
13) Fraud Detection in Healthcare Insurance Using Random Forest, International Journal of Computer Applications, 2021.
14) Machine Learning Approaches for Healthcare Fraud Detection, Springer Journal of Big Data Analytics, 2020.
15) Anomaly Detection Techniques for Medicare Fraud Analysis, IEEE Access, 2021.
16) Predictive Analytics in Healthcare Fraud Detection, Journal of Healthcare Informatics Research, 2019.
17) Healthcare Claim Fraud Detection Using Support Vector Machine, International Journal of Innovative Technology and Exploring Engineering, 2020.
18) Artificial Intelligence Techniques in Medical Fraud Detection, Elsevier Procedia Computer Science, 2021.
19) Big Data Analytics for Healthcare Fraud Prevention, Journal of Information Security and Applications, 2020.
20) Centers for Medicare & Medicaid Services – Medicare Fraud Prevention Guidelines.
21) World Health Organization – Digital Health and Healthcare Data Reports.
22) Scikit-learn Official Documentation
23) TensorFlow Official Documentation
24) Python Official Documentation
25) Pandas Official Documentation

https://doi.org/10.63665/IJMEC.1104s.49
ISSN: 2456-4265
IJMEC 2026