International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

AI-Powered Fraud Detection in Medicare Claims: Techniques and Analysis

Fnu Anupama

Submitted:15/09/2024 **Revised:** 18/10/2024 **Accepted:** 25/10/2024

Abstract: The implementation of AI-driven LLMs in the healthcare industry has had a profound effect and will continue to shape the healthcare and AI analytics sector. According to Straits research the healthcare analytics field is "valued at **USD 17.61 billion in 2024** and is projected to reach USD 21.78 billion in 2025". The rapidly shaping industry is just starting to grow. These implementations help increase cost effectiveness, implement fraud preventive measures, and risk reduction. The AI-driven implementation process of predictive analytics and finding patterns helps not only the healthcare industry but the beneficiaries indirectly. This study analyzes the use of supervised machine learning algorithms to detect fraudulent claims. This paper explains AI powered fraud Medicaid claims detection framework using machine learning algorithms applying to Medicare synthetic claims data set. Through Supervised Learning, focusing on random classification, along with explainable AI methods, this paper highlights how Medicare fraudulent claims are effectively found with high precision. In addition, this paper demonstrates important prerequisites such as Data preparation, model training, and evaluating performance. Our approach and results highlight the efficiency of AI in automating claims fraud detection, reducing manual laborers' work and improving overall claims authentication processes. In addition, the paper also highlights statistical analysis and graphical representations that evaluate the efficacy of the generated model, contributing to real time issues with Medicare claims fraud.

Keywords: *Medicare Fraud, Machine Learning, Fraud Detection, Artificial Intelligence, Healthcare Claims*

1. Introduction

Health care fraud is a major issue in the United States and worldwide, these issues are resulting in huge financial losses and reducing trust in the healthcare system. According to Thomson and Reuters "The U.S. Department of Justice (DOJ) reported civil settlements and judgments under the False Claims Act related to healthcare fraud that exceeded $1.8 billion in the fiscal year ending Sept. 30, 2023." [10]

Healthcare fraud is a significant issue in the United States and worldwide. Medicare, a health insurance program for people aged 65 or older and younger people with disabilities, has been targeted by fraudulent activities. These include overbilling, phantom billing, upcoding, and unnecessary medical procedures targeted to exploit the Medicare system. They do not evolve enough to deter fraud. The current manual mechanisms in place are labor intensive and not equipped to manage a mass volume of claims. Improving fraud detection is a major priority to increase the integrity of the healthcare system. Taxpayer money should not be wasted.

Historically, Medicare fraud detection has relied on a combination of manual reviews, statistical methods, and rule-based systems.

• **Manual Reviews:** These involve analyzers, and investigators who scrutinize claims for issues such as inconsistencies or errors that could point towards fraudulent claims. Though Medicare manual reviews can also be highly effective and resourceful in identifying fraudulent claims, they take a lot of time, are exhaustive, and are not sustainable because Medicare receives several million claims every year.

• **Statistical Methods:** A statistical test is employed to detect abnormal trends that are anomalous in one way or another. For instance, a provider exhibiting essentially different billing patterns than the average will attract an outcry. However, statistical models fail to capture the details of a normal financial transaction, and therefore, more sophisticated forms of fraud may not be identified easily with statistical methods; the statistical model may not easily be modified to reflect changes in the fraud type, which may also gradually develop over a period. Manual statistical tests without continuous monitoring and not updating the model with new evidence.

• **Rule-Based Systems:** These systems employ predefined parameters to alert the user of unconstitutional claims. For instance, a rule could mark any claim that is above a specified dollar sum or includes a higher number of services. This technique

is extremely useful for detecting organized fraud. However, it is not nearly as useful when used in a more complex environment because it can miss frauds that do not fit into the system's library of behaviors. One major disadvantage of conventional fraud detection approaches is that the fraudulent alarms generated tend to be numerous, and only a small proportion of these represent authentic frauds; the rest are excellent examples of false alarms. This results in time wastage, more paperwork for the providers, and the removal of resources from proper fraud identification. In addition, since fraud schemes are evolving and involve more complications, traditional measures cannot adequately respond to them and, at times, cannot evolve fast enough to counter new types of fraudulent schemes [4].

The approach and methodology explained in this research paper is incremental evidence focusing on manipulating the data set to include fraudulent claims and running the ML algorithms to detect fraudulent claims; This research investigates the application of AI- driven fraud detection models in Medicare claims processing. The data set used for this contains a large Medicare claims data set, specifically synthetic Public Use Files (SynPUFs). These claims data are updated with fraudulent claims to prepare a complete data set that includes both accurate and fraudulent claims which allows for the implementation of supervised learning approaches to classify which claims are fraudulent [5]. This study focuses on data preparation, augmentation and selection of a model, applying a random Forest Classifier to attain high outcomes In addition to this, AI techniques that are explainable have been incorporated to improve transparency in fraud detection decision-making. Statistical analysis and graphical representations, effectiveness of the AI powered fraud detection is demonstrated and derives its impact on healthcare fraud prevention.

On the contrary, Machine Learning driven programs offer advanced solutions by identifying unusual patterns in claims data sets.

1.1. Contributions to this Study

This study's foundation is the usage of random forest for fraud detection; data engineering and visualization approaches are employed to show how well the ML model performs. This model can be readily expanded to real-time claims since it is implemented using synthetic claims.

2. Fraud Detection Methodology in Medicare Claims Using Machine Learning

2.1. Data Preparation and Data Augmentation

There are 66,773 Medicare claims in this data set, total 81 attributes, consisting of the claim's details, patient demographics and services taken. To effectively train the model, the data was prepared by modifying or updating the data set manually.

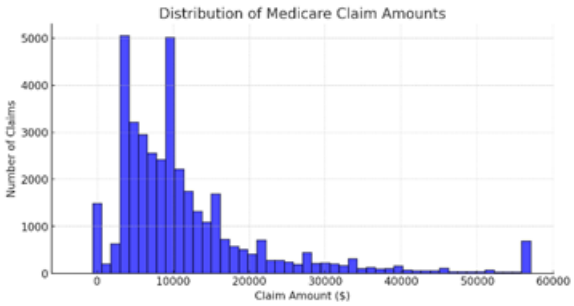

Fig (1): Histogram of Medicare Claim Amounts

Histogram of Medicare Claim Amounts

• This chart shows the **distribution of claim amounts** in the dataset.

• Most claims fall within a lower range, but there are some high-value claims (fraud indicators).

• A **right-skewed distribution** suggests a few claims are exceptionally high.

2.2. Handling Missing Data

The large portion of the dataset had missing values, more likely numerical columns, median imputation was applied to enter missing values, maintaining data consistency, without having a biased data set.

2.3. Feature Selection

Removed attributes which are not required. ID-based fields and categorical data with high cardinality were excluded to prevent data leakage. The final dataset contained **62 numerical attributes** that were used for training the model.

2.4. Synthetic Fraudulent Claim Generation

Since Synpuff claims did not have fraudulent claims, two approaches were taken to update the claims data set.

Modification of Existing Claims: 100 randomly selected claims were altered by inflating cost-related attributes (e.g., **procedure costs increased 1.5x to 3x**) and changing service codes.

• **Addition of New Anomalous Claims:** 100 new claims were added by duplicating real claims and introducing fraudulent indicators, such as extreme values in cost fields and suspicious service patterns.

A new **binary label** column (Fraudulent) was created, where **1** represented fraudulent claims, and **0** indicated legitimate claims.

3. Handling Class Imbalance with SMOTE

Use either SI (MKS) or CGS as primary units. (SI units are strongly encouraged.) English units may be used as secondary units (in parentheses). This applies to papers in data storage**.** For example, write "15 Gb/cm^2 (100 Gb/in^2)." An exception is when English units are used as identifiers in trade, such as "3½-in disk drive."

Fraudulent claims typically constitute a small fraction of total claims, making the dataset highly imbalanced. To mitigate this issue, **Synthetic Minority Over-sampling Technique (SMOTE)** was applied. SMOTE synthetically generates new fraudulent samples by interpolating between existing fraud cases, balancing the dataset. This ensured that the classifier did not develop a bias toward non-fraudulent claims.

4. Machine Learning Model Selection and Training

4.1. Model Choice: Random Forest Classifier

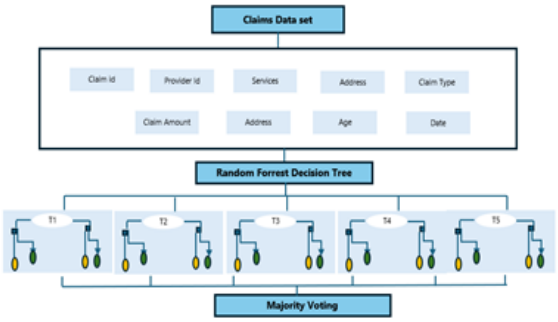

Fig (2): Random Forest Classifier Tree Structure

Random Forest Classifier:

Random forests are supervised by machine learning, this will have multiple tree predictors, each tree will have a value that is defined with random vector sampled independently, all the trees will have the same distribution within the forest. The error generalizing becomes evident when the number of trees in the forest is large. The strength of individual trees in the forest contributes to error generalization and correlation between trees. Likely comparison can be done by using a random selection feature to divide node yields error favourably to Ad but are more robust with respect to noise [3]. Internal estimates monitor error, strength, and correlation and these are used to show the response to increasing the number of features used in the splitting. Internal estimates are also used to measure variable importance. These ideas are also applicable to regression.

For fraud detection, a Random Forest Classifier (RFC) was selected due to its ability to handle high-dimensional data and detect complex patterns. RFC is an ensemble-based learning method that constructs multiple decision trees during training and aggregates their predictions for robust classification [1].

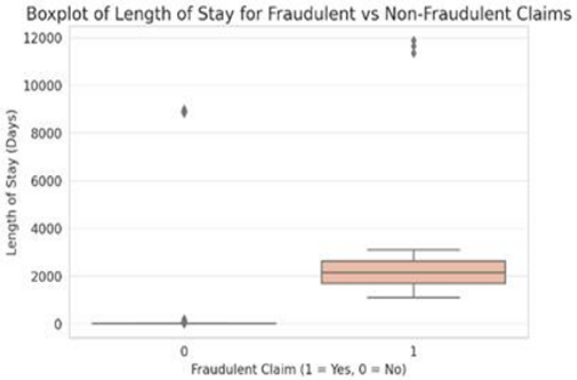

Fig (3) Boxplot

This boxplot compares the length of stay (in days) for fraudulent vs non-fraudulent claims.

The above diagram illustrates how fraudulent or legitimate claims are classified depending on the properties of characteristic of Medicare claims:

X-Axis (Fraudulent Claim - 0 or 1): Important Points to Note

0 Denotes assertions that are not fraudulent.

1 Stands for false statements.

Y-Axis (Stay Duration in Days):

Shows how many days a patient was covered under the claim.
For Claims That Are Not Fraudulent (0):

The period of stay for most claims is short.

There are a few suspicious but uncommon extreme outliers with numbers over 8000 days.

Regarding False Allegations (1):

Compared to claims that are not false, the median length of stay is significantly longer.

Longer hospital stays are generally associated with false claims, according to the interquartile range (IQR).

Outliers with exceptionally lengthy stays (up to 12,000

days) are blatant signs of deception.

What This Signifies for the Identification of Fraud:

Hospital stays for fraudulent claims are typically longer than those for legitimate claims.

Why Random Forest Classifier

According to Markose, "Supervised Learning: Here, an algorithm fed a set of clearly distinguishable examples of fraud and genuine claims. The model acquires the characteristics of fraud and uses them to classify new data that has not been classified as fraud beforehand. Few of the algorithms that can be used in fraud detection are decision trees, random forests, and Support Vector Machines (SVM). These models can be highly effective and precise with the help of an exhaustive and accurate labelled data set for tuning."

4.2. Hyperparameter Optimization

To enhance performance, the model was fine-tuned using **grid search** and **randomized search** methods, optimizing the following hyperparameters:

- **Number of trees (n_estimators):** 200
- **Maximum tree depth (max_depth):** twenty
- **Minimum samples per split (min_samples_split):** five
- **Minimum samples per leaf (min_samples_leaf):** two

These optimizations helped balance model complexity and generalization, reducing overfitting while improving fraud detection accuracy.

4.3. Model Training and Evaluation

The dataset was divided into **80% training** and **20% testing** subsets. **The optimized Random Forest model** trained on the resampled data.

Upon evaluation, the model achieved:

- **Accuracy:** 99.91%
- **Precision for Fraudulent Claims:** 95%
- **Recall for Fraudulent Claims:** 66%
- **F1-score for Fraudulent Claims:** 78%

Despite high accuracy, the recall value indicated potential false negatives (i.e., some fraudulent claims were still undetected). Future work aims to improve recall through additional ensemble techniques and alternative fraud detection algorithms.

Table (1): Key features in fraud detection

Rank	*Feature Name*	*Importance Score*
1	Service Costs	0.245
2	Number of Services Claimed	0.198
3	Patient Visit Frequency	0.153
4	Procedure Code Manipulation	0.13
5	Billing Anomalies	0.102
6	Provider ID Frequency	0.089
7	Length of Stay	0.078
8	Patient Demographics (Age, Gender, Region)	0.065
9	High Reimbursement Amounts	0.058
10	Unusual Claim Submissions	0.047

5. Fraud Detection Insights

To interpret the model's decision-making, **feature importance analysis** was conducted. The **top contributing factors** to fraud detection included:

- Service Costs
- Number of Services Claimed
- Patient Visit Frequency
- Unusual Billing Codes

These insights help policymakers and healthcare administrators identify fraudulent claim patterns and implement targeted fraud prevention strategies.

6. Conclusion and Future Work

The application of **Random Forest with SMOTE** significantly improved fraud detection in Medicare claims. The results demonstrated the potential of machine learning in reducing financial fraud in healthcare systems. Future enhancements will explore:

- **Neural Networks** and **XGBoost** for improved recall.
- **Deep Learning Architectures** for detecting more complex fraud patterns.
- **Integration with Real-Time Monitoring Systems**

for proactive fraud detection.

This study provides a **scalable fraud detection framework,** offering a foundation for future advancements in healthcare fraud analytics . This research is with the limitation of using

synthetic claims, using synthetic data effectively requires making sure the data accurately represents the characteristics and patterns of real-world data and is suggestive of that data. Carefully considered data synthesis techniques and validation against real-world data are necessary to ensure that the synthetic data is of high quality and useful for model construction.

The ability to simulate different scenarios and assess model performance in different contexts is one benefit of using synthetic data. For example, by mimicking fraud methods such as upcoding or charging for services that were never rendered, it allows a business to assess how well AI and ML models detect and classify fraud.

My study adds to the growing literature on efficient ways to detect fraud detecting fraud, but future directions would be to use state-of-the-art learning techniques like deep learning to increase accuracy even more (and/or ability to use unstructured data in the prediction?)

The future of fraud claims detection is deep learning approaches, which should be explored in detail. These are future fraud detection approaches that can be combined with detecting inconsistencies along with neural networks to enhance fraud detection mechanisms.

Appendix

Appendix A: Feature Importance in Fraud Detection

The table below presents the **top 10 most important features** contributing to fraud detection based on the **Random Forest model's feature importance scores.** These features play a critical role in distinguishing **fraudulent vs. non-fraudulent** claims.

7. References and Footnotes

Acknowledgements

ChatGPT model 4.0 is used to execute the model due to restricted access to proprietary LLM.

References

[1] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, Doi: https://doi.org/10.1023/a:1010933404324.

[2] – NHCAA," National Health Care Anti-Fraud Association, 2023. https://www.nhcaa.org/tools-insights/about-health-care-fraud/the-challenge-of-health-care-fraud/

[3] S. J. Rigatti, "Random Forest," Journal of Insurance Medicine, vol. 47, no. 1, pp. 31–39, Jan. 2017, Available: https://meridian.allenpress.com/jim/article/47/1/31/131479/Random-Forest

[4] P GEETHA, J C JENCY, and BALAKUMAARAN R K, "A Study to Assess the Effectiveness of Medication Safety Education on Knowledge among Undergraduate Nursing Students in a Selected College of Nursing, Chennai," International Journal For Multidisciplinary Research, vol. 6, no. 5, Sep. 2024, doi: https://doi.org/10.36948/ijfmr.2024.v06i05.27680.

[5] Medicare, "CMS 2008-2010 Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF) | CMS," Cms.gov, 2024. https://www.cms.gov/data-research/statistics-trends-and-reports/medicare-claims-synthetic-public-use-files/cms-2008-2010-data-entrepreneurs-synthetic-public-use-file-de-synpuf (accessed Feb. 23, 2025).

[6] M. Herland, T. M. Khoshgoftaar, and R. A. Bauder, "Big Data fraud detection using multiple Medicare data sources," *Journal of Big Data*, vol. 5, no. 1, Sep. 2018, Doi: https://doi.org/10.1186/s40537-018-0138-3.

[7] Z. Hamid, F. Khalique, S. Mahmood, A. Daud, A. Bukhari, and Bader Alshemaimri, "Healthcare insurance fraud detection using data mining," BMC medical informatics and decision making, vol. 24, no. 1, Apr. 2024, Doi: https://doi.org/10.1186/s12911-024-02512-4.

[8] Association of Certified Fraud Examiners, "Blog Detail," Acfe.com, 2024. https://www.acfe.com/acfe-insights-blog/blog-detail?s=future-of-healthcare-fraud-artificial-intelligence

[9] J. Hancock, R. A. Bauder, H. Wang, and T. M. Khoshgoftaar, "Explainable machine learning models for Medicare fraud detection," *Journal of Big Data*, vol. 10, no. 1, Oct. 2023, Doi: https://doi.org/10.1186/s40537-023-00821-5.

[10] Melissa D. Berry, "Medicare and Medicaid fraudsters continue to steal taxpayer money - Thomson Reuters Institute," Thomson Reuters Institute, May 13, 2024. https://www.thomsonreuters.com/en-us/posts/investigation-fraud-and-risk/medicare-medicaid-fraud-2