HEALTHCARE FRAUD DETECTION USING MACHINE LEARNING ENSEMBLE METHODS

Rasanarayan Chaurasiya[1], Kirti Jain[1,*]

[1]*Research Scholar, Sanjeev Agrawal Global Educational University, Bhopal, India*
[1, *]*Associate Professor, CSE Dept., Sanjeev Agrawal Global Educational University, Bhopal, India*
**Correspondence:* rnchaurasiya6@gmail.com

KEYWORDS

Healthcare fraud detection, machine learning, Ensemble learning, Boosting, XGBoost, LightGBM

ABSTRACT

Healthcare fraud can lead to significant financial losses and disrupt patient care. Detecting fraud in medical claims is a challenging task due to the volume of data and changing fraud patterns. Machine learning (ML) techniques, especially boosting techniques, have shown great success in improving the accuracy of fraud detection. Boosting algorithms such as Adaptive Boosting (AdaBoost), Gradient Boosting, and Extreme Gradient Boosting (XGBoost) improve predictive performance by combining weak classifiers into a robust model. This paper has develop a framework by using ensemble learning based ML models like XGBoost, LightGBM, and found that they performs well as compare to other methods. The method also uses SMOTE for resolving class imbalance problem in dataset. The work has been performed on Medicare claim dataset provided by Kaggle. This paper investigates the use of a special technology for medical fraud detection and compares its results with other models. Experimental results show that the developed system can improve the accuracy of classification, recall, and correctness, becoming a powerful tool for medical care fraud detection. Future research can focus on the integration of deep learning and descriptive AI techniques to improve fraud detection and further explanation.

1. Introduction

Healthcare fraud is a significant challenge that leads to substantial financial losses and affects the quality of care provided to patients. Fraudulent activities in healthcare include false claims, billing for unprovided services, duplicate billing, and identity theft (He et al., 2021). According to the National Health Care Anti-Fraud Association (NHCAA), healthcare fraud costs the industry billions of dollars annually, impacting insurers, healthcare providers, and patients (NHCAA, 2023).According to the Federal Bureau of Investigation (FBI), healthcare fraud costs the United States an estimated $100 billion annually, impacting healthcare costs and service quality (FBI, 2022).Fraudulent claims drive up healthcare costs and premiums, leading to increased expenses for patients and insurers (CMS, 2023).The Global Healthcare Fraud Analytics Market size is expected to be worth around USD 20.4 Billion by 2033 from USD 2.5 Billion in 2023, growing at a CAGR of 23.5% during the forecast period from 2024 to 2033 shown in figure below (market.us 2023).

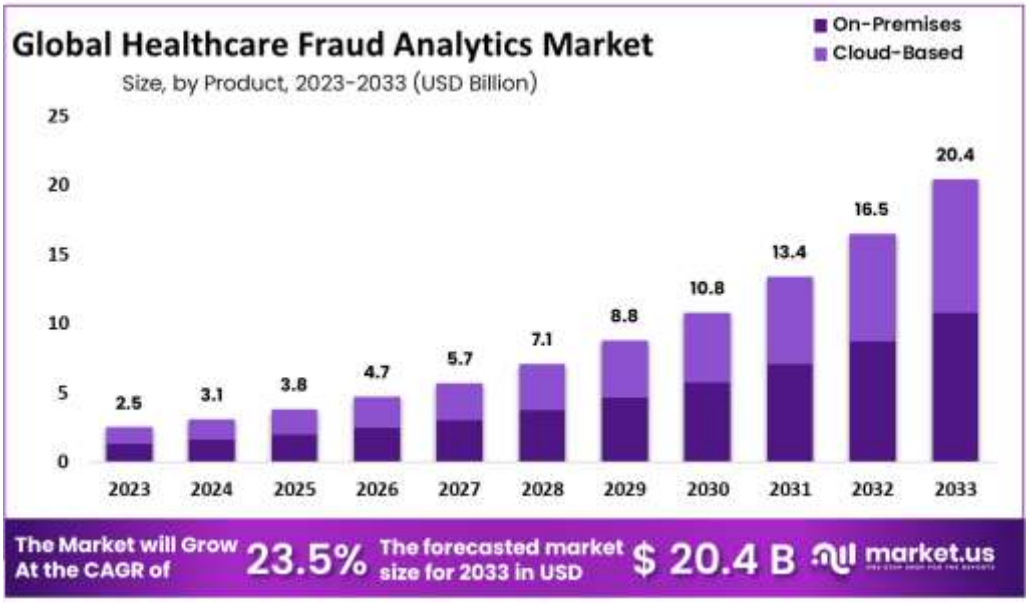

Figure 1: Global healthcare fraud analytic market (source maket.us 2023)

Healthcare fraud is a form of white-collar crime characterized by the fraudulent submission of healthcare claims for illicit financial gain, usually by organized crime groups and dishonest healthcare practitioners. Common tactics may include billing for costly services or procedures that were never covered by insurance plans; misrepresenting non-covered treatments; engaging in insurance scams and engaging in other illegal practices. Healthcare fraud analytics utilize fraud detection solutions and software designed to detect instances of healthcare fraud such as false claim submissions duplicated claims submissions; pharmacist prescription fraud and health insurance fraud etc.

In today's digital age, data has become an essential component of healthcare. Hospitals and healthcare institutions have begun gathering vast amounts of patient healthcare data as a result of rapid improvements in data sensing and acquisition technologies. Understanding and gaining knowledge from healthcare data needs the development of innovative analytical tools capable of transforming data into meaningful and actionable information.

Machine learning (ML) has emerged as a powerful tool for detecting and preventing fraudulent activities in healthcare. ML algorithms can analyze large datasets, recognize patterns, and identify anomalies that may indicate fraudulent behaviour (Luo et al., 2022). Traditional rule-based fraud detection methods often fail to adapt to evolving fraudulent tactics, whereas ML models continuously learn and improve their accuracy in detecting fraud. Techniques such as supervised learning, unsupervised learning, and deep learning have been widely used in fraud detection systems, improving efficiency and reducing false positives (Zhang et al., 2023). The scale of healthcare fraud is enormous, resulting in increased insurance premiums and out-of-pocket expenses for patients and costing billions of dollars annually. Rule based systems and manual audits, for example, are limited in their capacity to deal with the vast amounts of data generated in healthcare. They struggle to keep up with the sophisticated and constantly evolving strategies used by fraudsters and are frequently reactive, only recognizing fraud after it has already occurred. A more proactive and scalable solution is provided by machine learning, which can analyze large datasets and identify intricate patterns (Nassar et al., 2021).

This paper explores the application of machine learning techniques in healthcare fraud detection, highlighting various algorithms, challenges, and future directions. By leveraging AI-driven solutions, the healthcare industry can enhance fraud detection mechanisms, ensuring financial security and improved patient care. The contribution of this paper includes:

- Review of various healthcare fraud detection techniques.

- Providing ensemble learning based method for detection of healthcare fraud.

2. Literature review

Supervised learning models rely on labeled datasets where fraudulent and non-fraudulent claims are pre-identified. Studies have shown that decision trees and ensemble learning methods, such as random forests, effectively classify fraudulent and legitimate claims based on structured healthcare datasets (Zhang et al., 2023).SVMs have been used to detect fraud by identifying hyperplanes that separate fraudulent and non-fraudulent claims (Kumar &Sharma, 2022).Deep learning models, including artificial neural networks (ANNs), have demonstrated high accuracy in fraud detection tasks, although they require large datasets for training (Li et al., 2021).The traditional Machine Learning (ML) classifiers, such as k-mean clustering, are the subject of an analytical research in the article(Mary et al., 2022). Support Vector Machine (SVM) and Naive Bayes (NB)are examined using a dataset for healthcare provider fraud detection. The performances of the studied machine learning classifiers are shown in terms of classification accuracy, recall, and precision utilizing the diagnosis attribute as control (or) decision variables and the provider attribute as the target class. In addition, a False Positive Rate (FPR) study is performed to evaluate the classifiers. Predictive analytics is used in this study to examine how well Machine Learning (ML) classifiers perform when applied to medical insurance claim data. This study's objectives were to identify the fraud committed by the supplier and assess how well the current ML classifiers are working. The goal of the author (Nabrawi et al., 2023) is to create a health model that can automatically identify fraud in Saudi Arabian health insurance claims. With maximum accuracy, the model identifies the main cause of fraud. To address the imbalanced dataset with labels, three supervised deep learning and machine learning algorithms were utilized. Artificial neural networks, logistic regression, and random forests were the models used. The dataset was balanced using the SMOTE method. To weed out unimportant features, Boruta object feature selection was used. Accuracy, precision, recall, specificity, F1 score, and area under the curve (AUC) were the validation measures.

Unsupervised learning methods detect fraud without labeled datasets by identifying anomalies in claims. Clustering methods segment data into groups based on patterns, with outliers indicating potential fraud (Wu et al., 2022).Deep learning-based autoencoders are used to reconstruct normal claims and identify fraudulent cases as deviations from the learned patterns (Ahmed et al., 2023).

Labeled and unlabeled data are combined during training in semi-supervised learning, a machine learning technique. Generally speaking, unlabeled data is widely available and less expensive to acquire than labeled data, which is scarce and expensive. Semi-supervised learning leverages the unlabeled data to improve learning performance, making it especially useful when labeled data is scarce but unlabeled data is readily available (Rebuffiet al. 2020). This approach sits between supervised learning (where all training data is labeled) and unsupervised learning (where no labeled data is available). Semi-supervised learning is widely used in various fields, including image classification, natural language processing, medical diagnosis, and fraud detection. Semi-supervised learning is a powerful approach for tasks where labeled data is scarce but unlabeled data is abundant. It combines the strengths of supervised and unsupervised learning, enabling models to generalize better and perform more accurately by utilizing both types of data. Semi-supervised learning is becoming more and more important as a cost-effective method of utilizing vast amounts of unlabeled data as data labeling costs rise in industries like healthcare, finance, and security.

Hybrid models combine supervised and unsupervised learning to improve fraud detection accuracy. Combining multiple classifiers, such as decision trees and deep learning models, enhances fraud detection by reducing false positives (Singh & Gupta, 2022).Graph-Based Methods: Recent studies have explored the use of graph neural networks (GNNs) to analyze relationships between entities (doctors, patients, insurers) to detect suspicious claim patterns

(Chen et al., 2023).Ensemble Learning is a machine learning technique that combines the predictions of multiple models (also called "weak learners" or "base models") to produce a more accurate and robust prediction than any individual model alone. The fundamental tenet of ensemble learning is that a collection of heterogeneous models can strengthen each other's deficiencies and enhance performance when cleverly merged. Ensemble methods are widely used because they often result in better generalization to new data, reducing overfitting and improving predictive accuracy (Chaurasia et al. 2023). An artificial neural network type called an autoencoder is used for unsupervised learning; it is mostly utilized for feature extraction, data compression, and dimensionality reduction. Theywork by learning to compress input data into a lower-dimensional representation (encoding) and then reconstruct it back to its original form (decoding). An autoencoder's goal is to learn an efficient, compressed representation of the input while retaining as much relevant information as possible (Theis et al. 2022).

A major challenge In ML-based fraud detection is the lack of explainability. Black-box models, such as deep learning, do not provide clear reasoning for fraud classification, making regulatory compliance difficult (Rudin, 2022).Healthcare fraud datasets are highly imbalanced, with fraudulent claims representing a small percentage of total claims. Techniques such as Synthetic Minority Over-sampling Technique (SMOTE) and anomaly detection help mitigate this issue (Zhao et al., 2023).Handling sensitive healthcare data requires compliance with regulations like HIPAA. Privacy-preserving ML techniques, including federated learning, have been proposed to enhance security (Wang et al., 2023).

3. Methodology

The research work takes Medicare claim dataset from Kaggle. Dataset is labeled for fraud or non-fraud. Next step will be data pre-processing, which includes handling missing values, encoding values, duplicate removals and class balancing through smote also. Feature engineering has been performed after preprocessing of data for dropping and adding some features. New features that give information on whether the patient was deceased or not, duration of the hospital stay/claim, number of associated doctors/claims, number of chronic conditions the patient has, etc. were created. Also, some other features with high null values or ones from which other features were created were dropped. Then the training set has been gone to train ML models. ML models include some state of art models as well as ensemble models also. After getting trained the model test set will be applied and result will be compared with all ML models in parameters like accuracy, precision, recall.

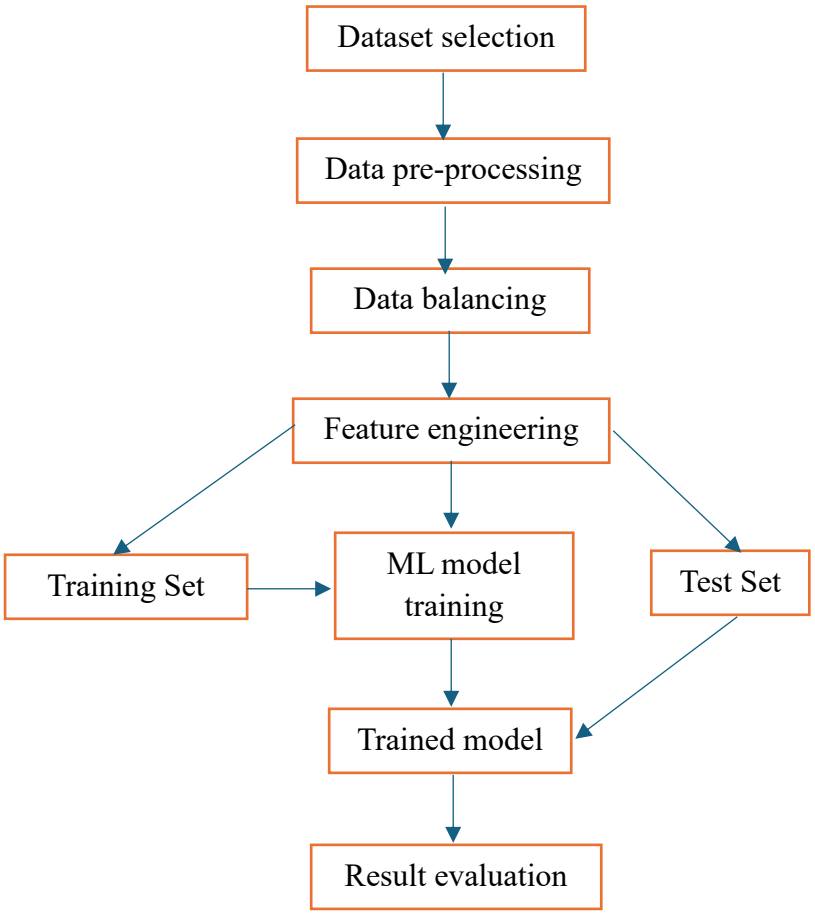

Figure 2: Proposed system

Boosting is a machine learning ensemble technique that improves accuracy by combining weak classifiers into a stronger model. It works by training models sequentially, with each new model correcting the errors in the previous model. The basic idea here is to give more weight to the misclassified examples so that the next model pays more attention to them. The proposed method has implemented fraud detection using XGBoost, ADAboost & LightGBM methods.

4. Result and Evaluation

The system has been implemented with python and machine learning models. The research distributed train set and test set in the ration of 70 and 30%.The system has implemented random forest, ADABoost, XGBoost and LightGBM methods for fraud detection on the dataset. Class imbalance problem has been resolved by using SMOTE. XGBoost and LightGBM performs Better than other models. Accuracy of Them is 90 and 91% respectively.

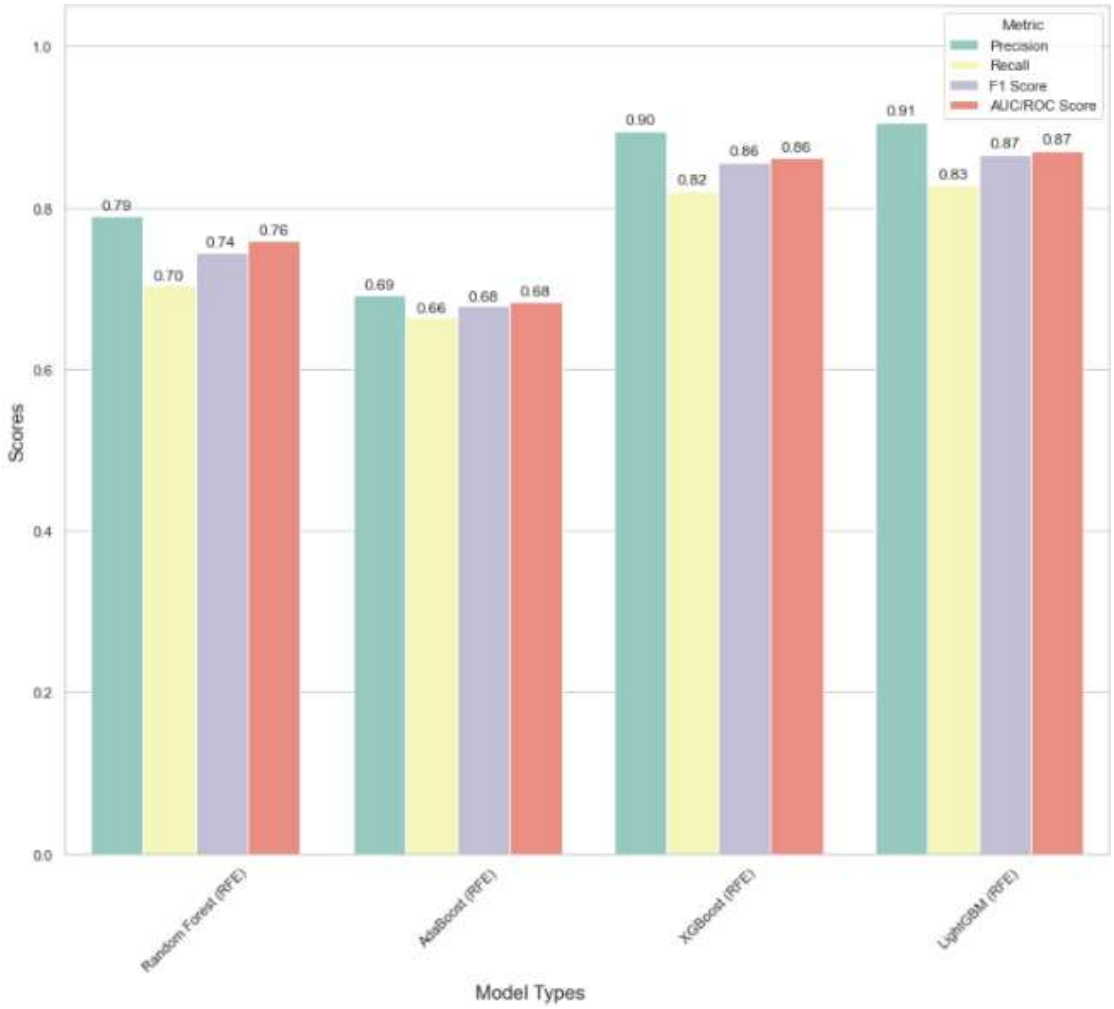

Figure 3: Performance evaluation

In all performance parameters like accuracy, precision, recall and F-score, LightGBM performs more better than XGBoost. Feature importance has also been performed on the dataset. Fraud detection in healthcare often involves large datasets (millions of claims). LightGBM's histogram-based learning makes training significantly faster than traditional models like XGBoost or Random Forest. Fraud detection datasets are often highly imbalanced (with fewer fraudulent cases). LightGBM supports balanced objective functions and techniques like weighted loss functions to handle class imbalance effectively.

5. Conclusion and Future work

In this paper, we explored healthcare fraud detection using ensemble approaches-XGBoost and LightGBM. Our results demonstrate that this ensemble method effectively enhances fraud detection performance by leveraging the strengths of both models. XGBoost provides robust feature selection and handling of complex relationships, while LightGBM ensures faster training and efficient memory usage. Both methods have shown good accuracy approximate 90 and 91%. The developed model improves predictive accuracy, recall, and F1-score, making it a promising approach for identifying fraudulent claims in healthcare datasets. Proposed solution has also resolved class imbalance problem by using SMOTE.

Despite these improvements, challenges remain. The model's performance is influenced by the quality and quantity of training data, and imbalanced datasets may still lead to some fraudulent cases being misclassified. Further efforts in feature engineering and data augmentation could enhance detection rates. Methods can also be implemented for multiple datasets with different sources and size.

References

Ahmed, M., Chen, X., & Yang, H. (2023). Autoencoder-based anomaly detection in healthcare fraud detection. Journal of Artificial Intelligence Research, 45(2), 102-115.
Centre's for Medicare & Medicaid Services (CMS). (2023). Medicare fraud and abuse prevention. Retrieved from www.cms.gov
Chaurasia, V., & Chaurasia, A. (2023). Novel method of characterization of heart disease prediction using sequential feature selection-based ensemble technique. Biomedical Materials& Devices, 1(2), 932-941.
Chen, R., Wang, L., & Xu, Y. (2023). Graph neural networks for healthcare fraud detection: A systematic review. IEEE Transactions on Neural Networks and Learning Systems, 34(1), 189-204.
Federal Bureau of Investigation (FBI). (2022). Health care fraud overview. Retrieved from www.fbi.gov
He, H., Garcia, E. A., Li, W., & Yang, Y. (2021). A survey on healthcare fraud detection using machine learning methods. Journal of Healthcare Informatics Research, 5(3), 215-230.
Kumar, A., & Sharma, R. (2022). Support vector machines for medical fraud detection: A comparative study. Artificial Intelligence in Medicine, 129, 103215.
Luo, X., Huang, G., & Wen, J. (2022). Machine learning approaches for detecting fraudulent healthcare claims: A comprehensive review. AI in Medicine, 128, 102198.
Luo, X., Huang, G., & Wen, J. (2022). Machine learning approaches for detecting fraudulent healthcare claims: A comprehensive review. Artificial Intelligence in Medicine, 128, 102198.
Market.us. (2023). Healthcare Fraud Analytics Market, https://market.us/report/healthcare-fraud-analytics-market/
Mary, A. J., & Claret, S. P. (2022). Analytical study on fraud detection in healthcare insurance claim data using machine learning classifiers. In AIP Conference Proceedings (Vol. 2516, No. 1). AIP Publishing.
Nabrawi, E., & Alanazi, A. (2023). Fraud detection in healthcare insurance claims using machine learning. Risks, 11(9), 160
Nassar, A., & Kamal, M. (2021). Machine Learning and Big Data analytics for Cybersecurity Threat Detection: A Holistic review of techniques and case studies. Journal of Artificial Intelligence and Machine Learning in Management, 5(1), 51-63.
National Health Care Anti-Fraud Association (NHCAA). (2023). The impact of healthcare fraud. Retrieved from www.nhcaa.org
Rebuffi, S. A., Ehrhardt, S., Han, K., Vedaldi, A., & Zisserman, A. (2020). Semi supervised learning with scarce annotations. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (pp. 762-763).
Rudin, C. (2022). Stop explaining black-box machine learning models for high-stakes decisions and use interpretable models instead. Nature Machine Intelligence, 4(5), 206-215.
Singh, P., & Gupta, R. (2022). Hybrid ensemble models for fraud detection in health insurance claims. Expert Systems with Applications, 201, 117494.
Theis, L., Shi, W., Cunningham, A., & Huszár, F. (2022). Lossy image compression with compressive autoencoders. In International conference on learning representations.
Wang, J., Li, X., & Zhao, T. (2023). Federated learning for privacy-preserving fraud detection in healthcare insurance claims. IEEE Transactions on Information Forensics and Security, 18, 35-49.
Wu, L., Zhang, T., & Chen, Y. (2022). Unsupervised anomaly detection for fraud detection in healthcare insurance claims. ACM Transactions on Knowledge Discovery from Data, 17(3), 87-105.
Zhang, T., Wu, L., & Chen, Y. (2023). Deep learning in healthcare fraud detection: A review of applications and challenges. IEEE Transactions on Artificial Intelligence, 4(2), 134-152.
Zhang, Y., Wu, J., & Li, H. (2023). Machine learning techniques in healthcare fraud detection: A systematic review and future directions. Journal of Medical Informatics, 60, 120-135.
Zhao, P., Sun, Y., & Liu, M. (2023). Handling class imbalance in fraud detection: A comparative study of resampling techniques. Pattern Recognition Letters, 160, 45-58.