Analysis of ML models & Data Balancing Techniques for Medicare Fraud Using PySpark

Konduru Praveen Karthik[a,1], Taduvai Satvik Gupta[a,2], Doradla Kaushik[a,3] and T. K. Ramesh[a,4*]
[a]*Department of Electronics and Communication Engineering*,
Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham, India
praveenkarthik2290@gmail.com[1], satviktaduvai@gmail.com[2], kaushikdoradla@gmail.com[3], tk_ramesh@blr.amrita.edu[4*]

***Abstract*—Medicare fraud is a persistent and costly issue, contributing to billions of dollars in annual losses in healthcare systems worldwide. With the increasing complexity of healthcare data and fraudulent schemes, traditional detection methods struggle to keep up, making Big Data analytics and machine learning essential for effective fraud detection. This paper proposes a robust methodology for detecting Medicare fraud using a PySpark framework, leveraging datasets from the CMS, LEIE, FDA. The approach integrates these datasets to create a comprehensive profile for each claim and applies feature selection to identify indicators of fraudulent behavior. The methodology employs data balancing techniques, such as RUS and SMOTE to address class imbalances in fraud data. Various machine learning models, including Gradient Boosted Trees, Random Forest, and Logistic Regression, are trained and evaluated on their accuracy, precision, recall, and specificity scores. The experimental results show that the proposed models effectively identify fraudulent claims, with the model trained on data balanced at a 1:5 ratio of fraud to non-fraud cases achieving the highest specificity of 73%. This work not only improves Medicare fraud detection but also offers a scalable approach that can be applied to large-scale healthcare schemes globally, such as India's Ayushman Bharat program. The results underscore the potential of Big Data solutions in safeguarding healthcare resources by mitigating fraud.**

***Keywords*—PySpark, RUS, SMOTE, Random Forest, GBT, Logistic Regression, Accuracy, Specificity.**

I. Introduction

Medicare fraud represents a significant challenge in healthcare, resulting in an estimated loss of up to $64 billion annually due to fraudulent claims [1]. This problem is exacerbated by the vast and rapidly increasing volume of healthcare data, expected to reach 163 zettabytes by 2025 [2]. In India, large-scale healthcare schemes like the Ayushman Bharat – Pradhan Mantri Jan Arogya Yojana (AB-PMJAY) provide health insurance coverage to over 500 million economically vulnerable individuals. The scale of these programs makes them attractive targets for fraud, leading to substantial financial losses and impacting the delivery of healthcare to genuine beneficiaries.Reports indicate that fraud in the healthcare sector accounts for an estimated Rs. 6,000 crore ($800 million USD) annually across various schemes, including Ayushman Bharat. [3] Traditional methods for fraud detection, which rely on manual audits and basic rule-based algorithms, are often inadequate in handling the complexity and scale of this data, leading to delays and inaccuracies in detecting fraudulent activities.

By utilizing sophisticated machine learning algorithms and the integration of various healthcare data sources, big data analytics has become a potent remedy for these issues. The problem of class imbalance, in which fraudulent claims are far less common than valid claims, is one of the main challenges in Medicare fraud detection. Machine learning models are frequently biased toward the majority class as a result of this mismatch, which lowers their capacity to detect fraud [4].

To address these issues, the goal of this research is to use machine learning and Big Data analytics in the PySpark framework to create a thorough Medicare fraud detection system. To improve model performance, the proposed approach employs several data balancing strategies and incorporates datasets from the Food and Drug Administration (FDA), the List of Excluded Individuals/Entities (LEIE), and the Centers for Medicare & Medicaid Services (CMS). The system uses methods like feature selection, data sampling, and model assessment using measures like accuracy, precision, recall, and specificity in an effort to increase the accuracy of fraud detection and decrease false positives. In addition to improving Medicare fraud detection efficiency, this research intends to show the scalability and usefulness of Big Data solutions in the healthcare sector. This work's primary contributions include:

- Use the latest version of dataset for the analysis i.e. 2024 versions.
- Data integration is being performed on 3 different datasets and model has been trained and tested using the Integrated dataset.
- The model has been trained and tested using different ratios of fraud and non-fraud data.
- Model has been trained using different ML models and evaluated using the evaluation metrics like Specificity, Accuracy, Precision and Recall.

The research is structured as follows: Section 2 reviews recent studies on Medicare fraud detection, Section 3 discusses the architectures used in this work, Section 4 highlights key findings, and Section 5 provides conclusions and suggests directions for future research.

II. Literature Survey

This section unfolds the recent works in the medicare fraud detection and are outlined as follows:

Researchers in [5] tackle class imbalance in Medicare fraud detection, where only 0.062% of the dataset is fraudulent. Using sampling methods and models like Logistic Regression and Random Forest, they find random under sampling improves detection performance, emphasizing the need to address class imbalance.

The study in [6] evaluates CatBoost for Medicare fraud detection using claims data with many categorical features. The authors highlight CatBoost's reduced data pre-processing needs and superior performance over XGBoost in terms of AUC. By incorporating features like provider state, the study shows CatBoost's potential to improve fraud detection, addressing $52 billion in improper payments in 2017.

The authors in [7] explore improving Medicare fraud detection by combining graph analysis and machine learning, using GNNs and traditional models with graph centrality features. They highlight the billions in annual losses and show that graph-based features enhance detection accuracy, benefiting government and insurers.

The researchers in [8] provides a comprehensive review of various methodologies and techniques for detecting healthcare fraud, particularly focusing on upcoming and provider fraud. It highlights the challenges posed by high-dimensional big data and the inefficiencies of current static detection systems, emphasizing the need for innovative approaches such as machine learning, Bayesian modeling, and data mining to uncover unknown fraudulent patterns.

The research in [9] utilizes multiple datasets from the U.S. Centers for Medicare and Medicaid Services (CMS), implementing a neural network architecture with different activation functions,comparing them in terms of accuracy, training time, and computational cost, using maxout's unique property of selecting the maximum value across feature maps to optimize fraud detection tasks.On average, SeLU outperformed maxout in both classification accuracy and training speed, suggesting it as a preferred activation function for big data medical fraud detection tasks.

Study in [10] involves creating a medical behavior model based on frequent pattern mining and association rules to detect specific patterns in large healthcare datasets. The system is built on the Hadoop platform, leveraging its distributed nature to manage massive datasets and uncover fraud patterns in healthcare data. By using a distributed anomaly detection algorithm (DCMMAB), the model identifies "medical aggregation behavior" — patterns where multiple medical insurance cards are used too frequently at the same hospital within a short period, signaling potential fraud.The proposed system demonstrates a performance increase of over 20% in accuracy compared to existing methods.

Class imbalance, where minority classes have significantly fewer samples than majority classes has been discussed in [11]. Techniques to handle class imbalance include resampling methods like RUS and SMOTE (Synthetic Minority Oversampling Technique). RUS, in particular, is effective in reducing dataset size, thus speeding up training without adding synthetic data.RUS specifically with XGBoost, noting improvements in fraud detection for Medicare datasets at a 1:1 class ratio.

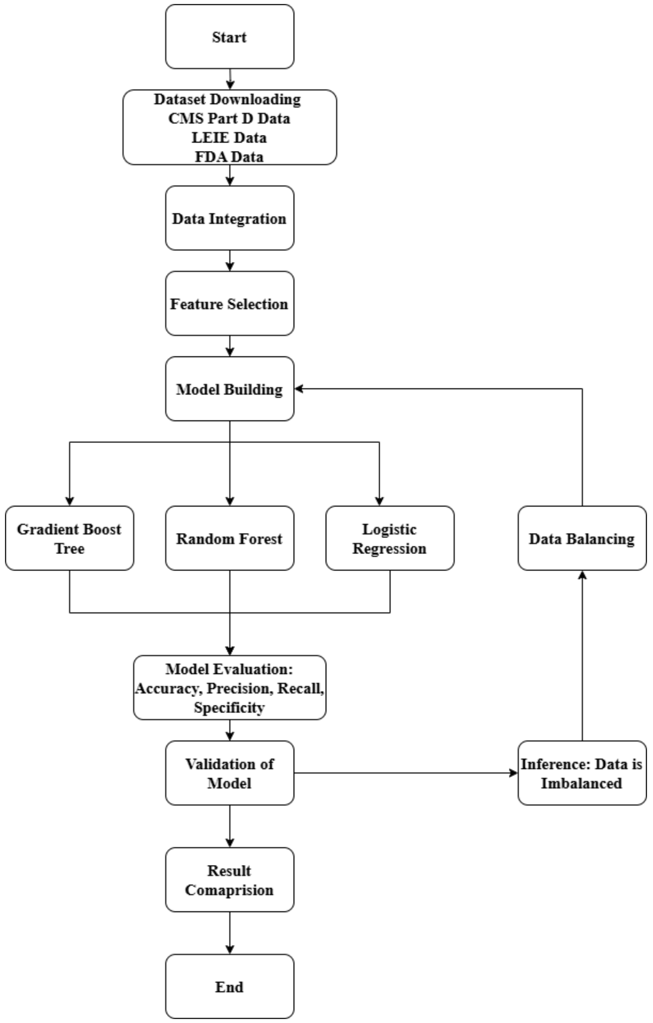


Fig. 1: Methodology of the Proposed Model

The reviewed studies highlight the challenges of class imbalance and high-dimensional data in Medicare fraud detection. Techniques such as sampling methods (RUS, SMOTE), specialized models like CatBoost and neural networks, and innovative approaches using graph analysis and pattern mining have shown improvements in detection performance. These methods emphasize the importance of targeted model choices and data handling strategies to enhance fraud detection accuracy and efficiency. Research in this paper uses low computation machine learning to increase the detection of fraudulent cases in the medicare fraud detection.

III. Proposed Methodology

The methodology for this research involves several key steps as shown in Fig. 1. Each step is designed to leverage large-scale datasets and advanced machine learning techniques to accurately identify potential fraud cases. The following subsections outline the methodology:

A. Data Collection

This research utilizes three primary datasets:

- CMS (Centers for Medicare & Medicaid Services) Dataset: Contains records of healthcare providers, patient services, and billing information for Medicare claims. [12]
- LEIE (List of Excluded Individuals/Entities): People and organizations that have been denied access to Medicare, Medicaid, and other federally sponsored health care programs are listed in the LEIE dataset. [13]
- FDA (Food and Drug Administration) Dataset: Provides data on approved drugs, medical devices, and adverse event reports to verify compliance with regulations. [14]

B. Data Integration

By connecting common identifiers like provider IDs, patient IDs, or claims numbers, three datasets are combined into a single one. Integrating the datasets involves combining data from CMS, LEIE, and FDA datasets using unique attributes such as First_name, Last_name, City and State. Integration is crucial for creating a unified dataset with all necessary information for fraud detection.

C. Feature Selection

Feature selection is a critical step in improving model performance by reducing noise and focusing on the most predictive attributes. Given the complexity of healthcare fraud detection, selecting appropriate features helps in building a model that accurately distinguishes fraudulent and non-fraudulent claims. Categorical attributes like First_name, Last_name, etc., were removed while keeping numerical attributes.

D. Data Balancing

Imbalanced datasets are a common challenge in fraud detection, where fraudulent claims (the minority class) make up a small fraction of total claims. Without addressing this imbalance, models tend to be biased toward the majority class, resulting in low recall for the minority class (fraudulent cases). An initial analysis was conducted to determine the degree of class imbalance. The imbalance ratio was found to be significant i.e., fraudulent claims might constitute less than 5% of total claims.

Synthetic Minority Over-sampling Technique (SMOTE) was applied to synthetically increase the number of fraudulent cases by generating synthetic examples [15]. By interpolating across minority class samples that already exist, SMOTE generates synthetic samples that aid the model in better identifying fraudulent patterns. The choice of SMOTE over random oversampling was made to prevent over fitting to exact duplicate samples.

In addition to oversampling, Random Under-Sampling (RUS) was also tested to reduce the size of the majority class (non-fraudulent claims), and is effective in reducing dataset size, thus speeding up training without adding synthetic data. [16]

Combination of SMOTE and Under-sampling: A hybrid approach was ultimately used, combining SMOTE and random under-sampling. This approach balanced the dataset while retaining a representative sample of non-fraudulent claims, which improved the model's ability to generalize on unseen data.

E. Model Building

After data pre-processing and balancing, the next step was to train and evaluate using the 5-fold cross validation on several machine learning models viz. Logistic Regression(LR), Gradient Boost Trees(GBT) and Random Forest(RF). Each of the following models were selected based on its suitability for handling imbalanced data and its potential for interpretability and accuracy.

- Gradient Boosting(GBT): Using decision trees as base learners, Gradient Boosting iteratively reduces error by focusing on the hardest-to-predict cases. Hyperparameters such as learning rate, maximum depth, and number of estimators were tuned optimize the model's performance [17].
- Random Forest(RF): Random Forest was chosen for its high interpretability and resilience to over-fitting. Hyperparameters, including the number of trees and maximum features, were optimized through cross-validation [18].
- Logistic Regression(LR): Logistic Regression was used as a baseline model due to its simplicity and interpretability. L2 regularization (Ridge) was applied to prevent overfitting, and the model was tested with various regularization strengths [19].

F. Model Evaluation

The evaluation of model performance was tailored to balance accuracy, precision, recall and specificity, as high specificity is critical in fraud detection to avoid classifying legitimate claims as fraudulent. The following metrics were used:

- Accuracy: Accuracy, while commonly used, was not the primary metric due to the imbalanced nature of the dataset. However, it provided an overview of model performance. The formula for accuracy is given in Eqn. 1.

$$Accuracy = \frac{No.of Correct Predictions}{Total No.of Predicitions} \tag{1}$$

- Precision and Recall: Precision measured the proportion of correctly identified fraud cases out of all predicted fraud cases, while recall measured the proportion of actual fraud cases correctly identified by the model. While recall is important to capture as many fraud cases as possible, precision is equally critical to prevent high false-positive rates [20]. The formula for precision and recall are given in Eqns. 2 & 3

$$Precision = \frac{True Positive}{True Positive + False Positive} \tag{2}$$

$$Recall = \frac{True Positive}{True Positive + False Negative} \tag{3}$$

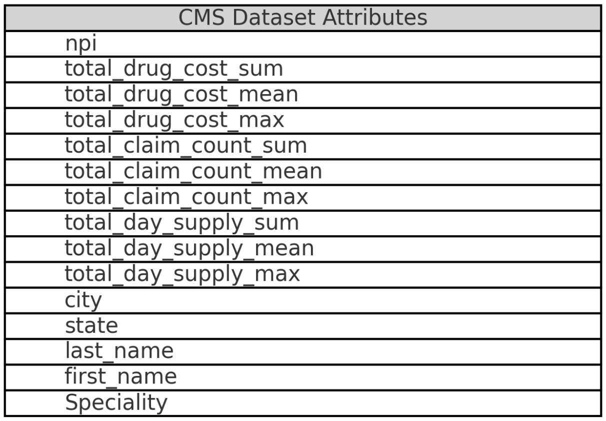
- Specificity: Specificity, or the True Negative Rate, measures the proportion of non-fraudulent cases that were correctly identified as such. This metric is particularly important in fraud detection, as high specificity ensures that legitimate claims are less likely to be incorrectly flagged as fraud. Maximizing specificity is critical for avoiding disruptions for legitimate claimants and minimizing unnecessary investigations. The formula for specificity is given in Eqn. 4

$$Specificity = \frac{TrueNegative}{TrueNegative + FalsePositive} \tag{4}$$

IV. Implementation and Results

PySpark is used to implement the proposed methodology. It is the Python API for Apache Spark, that enables the use of Python to write scalable, distributed data processing applications. It enables users to take advantage of Spark's large data processing, machine learning, and analytics capabilities in Python.

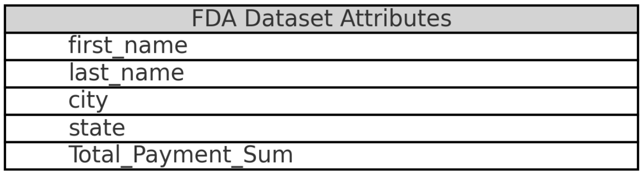
A PySpark session is created initially for building the model. Then the all three data sets are cleaned to remove the missing data and redundancy from all the three datasets. After data cleaning the Total drug costs and claims of each user is grouped and combined to have only one row for each user. The attributes of the CMS dataset after the Pre-processing is shown in the Fig. 2. Then FDA data is pre-processed and duplicates, null values in this dataset are removed and then common attributes with CMS data are kept and remaining columns are dropped. The Attributes of this dataset after pre-processing are shown in the Fig. 3.



CMS Dataset Attributes
npi
total_drug_cost_sum
total_drug_cost_mean
total_drug_cost_max
total_claim_count_sum
total_claim_count_mean
total_claim_count_max
total_day_supply_sum
total_day_supply_mean
total_day_supply_max
city
state
last_name
first_name
Speciality

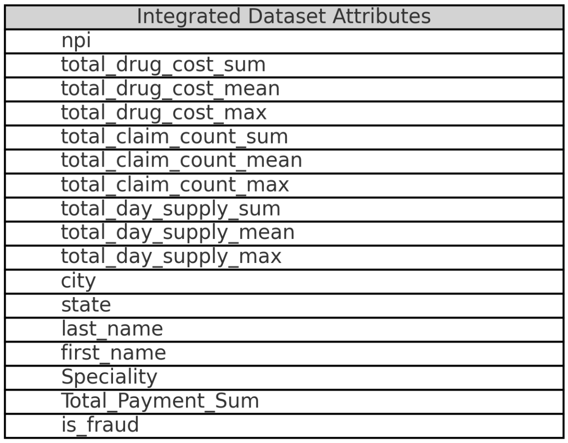
Fig. 2: Attributes of CMS dataset after Pre-processing

LEIE dataset is then preprocessed this contains the npi_id and details of the users who are excluded from the medicare. Hence, all the other columns except for npi_id are dropped and a new column is_fraud is created and marked as 1 for all in this dataset. Finally all the three dataset are integrated using the common attributes, that is npi_id in the CMS and LEIE dataset and first_name, last_name, city, state in CMS and FDA. Using these attributes the three datasets are joined to form a single dataset consisting the attributes shown in the Fig. 4.



FDA Dataset Attributes
first_name
last_name
city
state
Total_Payment_Sum

Fig. 3: Attributes of FDA dataset after Pre-processing



Integrated Dataset Attributes
npi
total_drug_cost_sum
total_drug_cost_mean
total_drug_cost_max
total_claim_count_sum
total_claim_count_mean
total_claim_count_max
total_day_supply_sum
total_day_supply_mean
total_day_supply_max
city
state
last_name
first_name
Speciality
Total_Payment_Sum
is_fraud

Fig. 4: Attributes of the integrated dataset after Pre-processing

After the data integration the features are selected for the training. Unwanted attributes like city, state, last_name, first_name, Specialty are dropped. The final attribute used for the training are npi_id, total_drug_cost, total_claims, total_day_supply, total_payment_sum, is_fraud. The sample of the dataset used for the training is shown in the Table. I

TABLE I: Sample of the Training Dataset

npi_id	total_drug_cost	total_claims	total_day_supply	total_payment_sum	is_fraud
1003838830	1409.26	88	6720	50412	0
1003868399	748.47	20	674	1123.42	0
1023191889	172.22	22	459	15079.68	0
1023498268	481.26	54	427	9988.95	0
1033120993	89.02	23	198	45693	1

Using npi_id, total_drug_cost, total_claims, total_day_supply, total_payment_sum as the feature vectors and is_fraud as the target variable, models like LR, RF and GBT are trained on the integrated dataset. The evaluation of this model is done using the 5-fold cross validation and the metrics used are Accuracy, Precision, Recall and Specificity. The results of which are tabulated in the Table. II.

TABLE II: Performance Comparison of Models

S.No	Model	Accuracy	Precision	Recall	Specificity
1	LR	99	99	99	10
2	GBT	98	99	98	10
3	RF	97	99	96	10

From the Table. II it is observed that even if accuracy of the model is high, the specificity of the model is very low indicating that the model is unable to predict the fraudulent classes properly. After analyzing the dataset it is observed that

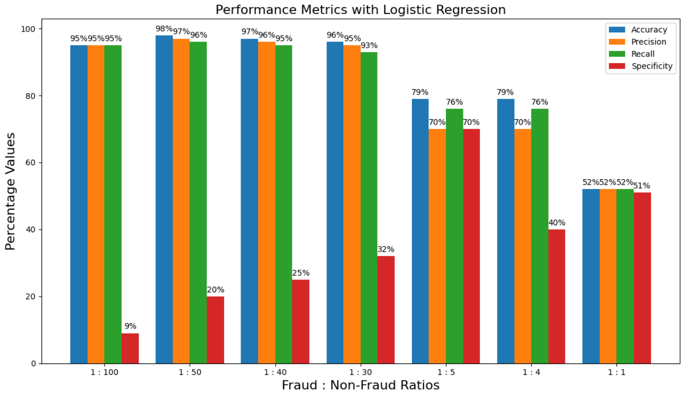


Fig. 5: Results of Data balancing for LR model

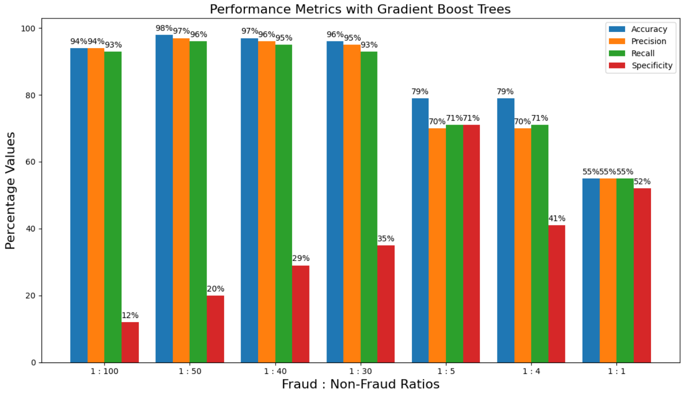


Fig. 6: Results of Data balancing for GBT model

the ratio of no.of fraud to no.of non-fraud samples is observed to be 1:3900 i.e. the data is heavily imbalanced hence the cause for low specificity. Several data balancing strategies, such as RUS and SMOTE, are employed to mitigate this, and models are retrained using the balanced data. The results of these are shown in the Fig. 5, 6 and 7

From the Fig. 5, 6 and 7 it is observed that at a ratio of 1:5 for fraud cases to non-fraud cases the models are performing better with accuracy of 80%, precision of 82%, recall of 74% and a specificity of 73%. This model can be used to detect all the fraud and non fraud cases accurately compared to the other models where the specificity is low even if it is having a higher accuracy than this model and the model with a 1:5 ratio can be stated as an optimal model for the prediction of the fraud and non-fraud cases in the medicare.

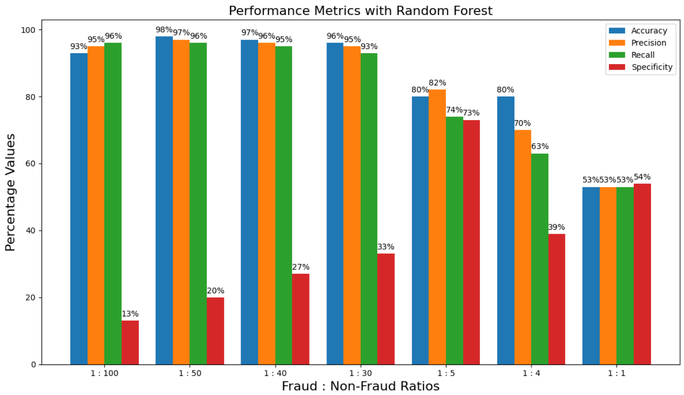


Fig. 7: Results of Data balancing for RF model

V. Conclusion and Future Scope

Medicare fraud detection models were evaluated based on metrics such as specificity, precision, recall, and accuracy, with the results showing a promising high performance after applying the data balancing techniques. Each model—Gradient Boosting Trees (GBT), Random Forest (RF), and Logistic Regression (LR)—achieved 99% accuracy, reflecting their strong ability to classify instances correctly overall. However, specificity was initially very low (around 10%) across all models, indicating that legitimate claims were frequently misclassified as fraudulent. This low specificity highlighted the severe class imbalance issue, with a fraud-to-non-fraud ratio of approximately 1:3900, severely impacting the models' ability to correctly identify legitimate claims. To address this, the study applied data balancing techniques, such as Random Under-Sampling (RUS), Synthetic Minority Over-Sampling Technique (SMOTE), and a hybrid approach combining both. This hybrid technique proved effective, leading to improved recall and precision for identifying fraud cases. Specificity also increased after balancing, indicating that the models became better at distinguishing between fraudulent and non-fraudulent claims.

Future advancements for this Medicare fraud detection system could include incorporating deep learning models for better fraud pattern recognition and developing real-time processing to detect fraud as it occurs. Additional data sources, like electronic health records and demographic information, could enhance decision-making, while advanced feature selection techniques and adaptive sampling methods could improve specificity. Integrating explainable AI would aid transparency, and optimizing for cloud scalability would make this solution more cost-effective and accessible for widespread healthcare use.

References

[1] National Health Care Anti-Fraud Association (NHCAA): "The Challenge of Health Care Fraud" (NHCAA). At: https://www.nhcaa.org/

[2] Reinsel, D., Gantz, J., & Rydning, J. (2018). "The Digitization of the World: From Edge to Core." IDC White Paper.At: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

[3] Economic Times article on healthcare fraud costs in India: "India's Healthcare Fraud May Cost Rs 6,000 Crore Annually." At: https://health.economictimes.indiatimes.com/news/industry/indias-healthcare-fraud-may-cost-rs-6000-crore-annually/80769002

[4] 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.

[5] R. A. Bauder, T. M. Khoshgoftaar and T. Hasanin, "Data Sampling Approaches with Severely Imbalanced Big Data for Medicare Fraud Detection," 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI), Volos, Greece, 2018, pp. 137-142.

[6] J. Hancock and T. M. Khoshgoftaar, ”Medicare Fraud Detection using CatBoost,” 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI), Las Vegas, NV, USA, 2020, pp. 97-103.

[7] Y. Yoo, J. Shin and S. Kyeong, ”Medicare Fraud Detection Using Graph Analysis: A Comparative Study of Machine Learning and Graph Neural Networks,” in IEEE Access, vol. 11, pp. 88278-88294, 2023.

[8] Mary, A. Jenita and S. P. Angelin Claret. “Design and development of big data-based model for detecting fraud in healthcare insurance industry.” Soft Computing 27 (2023): 8357-8369.

[9] G. Castaneda, P. Morris and T. M. Khoshgoftaar, ”Maxout Neural Network for Big Data Medical Fraud Detection,” 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), Newark, CA, USA, 2019, pp. 357-362.

[10] S. Zhou, J. He, H. Yang, D. Chen and R. Zhang, ”Big Data-Driven Abnormal Behavior Detection in Healthcare Based on Association Rules,” in IEEE Access, vol. 8, pp. 129002-129011, 2020J.

[11] Hancock, T. M. Khoshgoftaar and J. M. Johnson, ”The Effects of Random Undersampling for Big Data Medicare Fraud Detection,” 2022 IEEE International Conference on Service-Oriented System Engineering (SOSE), Newark, CA, USA, 2022, pp. 141-146.

[12] “Centers for Medicare & Medicaid Services Data,” Cms.gov, 2024. https://data.cms.gov/provider-summary-by-type-of-service/medicare-part-d-prescribers.

[13] “LEIE Downloadable Databases — Office of Inspector General — U.S. Department of Health and Human Services,” oig.hhs.gov. https://oig.hhs.gov/exclusions/exclusions_list.asp

[14] C. for D. E. and Research, “Drugs@FDA Data Files,” FDA, Dec. 2021, Available: https://www.fda.gov/drugs/drug-approvals-and-databases/drugsfda-data-files.

[15] T. T. S. et. al, “A Comparative Study of Various Oversampling Techniques for dealing With Tiny Sense Classes in Preposition Sense Disambiguation”, IJAST, vol. 29, no. 04, pp. 602 - 607, Feb. 2020.

[16] M. P. Paing, C. Pintavirooj, S. Tungjitkusolmun, S. Choomchuay and K. HAMAMOTO, ”Comparison of Sampling Methods for Imbalanced Data Classification in Random Forest,” 2018 11th Biomedical Engineering International Conference (BMEiCON), Chiang Mai, Thailand, 2018, pp. 1-5.

[17] R. Thomas and E. R. Vimina, “Enhancing the Classification Accuracy of Credit Default Using Extreme Gradient Boosting with Recursive Feature Selection,” Lecture notes in electrical engineering, pp. 585–591, Nov. 2021.

[18] V. K. Daliya, T. K. Ramesh and S. -B. Ko, ”An Optimised Multivariable Regression Model for Predictive Analysis of Diabetic Disease Progression,” in IEEE Access, vol. 9, pp. 99768-99780, 2021.

[19] M. G. Deepika and P. Sarika, “A Comparative Analysis of MFIs in India Using ANOVA and Logistic Regression Model,” Advances in intelligent systems and computing, pp. 503–515.

[20] D. V. K, T. K. Ramesh and S. A, ”A Machine Learning based Ensemble Approach for Predictive Analysis of Healthcare Data,” 2020 2nd PhD Colloquium on Ethically Driven Innovation and Technology for Society (PhD EDITS), Bangalore, India, 2020, pp. 1-2.