International Journal of Scientific Research in Engineering and Management (IJSREM)
Special Edition - Volume 09 NCFT -2025 SJIF Rating: 8.586 ISSN: 2582-3930

Leveraging Machine Learning for Improved Detection of Medicare Fraud

Velishala Aarthi
Student,Computer Science and Engineering
Guru Nanak Institutions Technical Campus (Autonomous)
Hyderabad,Telangana,India-501506
aarthivelishala@gmail.com

V.Sri Raghavendra
Student,Computer Science and Engineering
Guru Nanak Institutions Technical Campus (Autonomous)
Hyderabad,Telangana,India-501506
raghavendray967@gmail.com

V.Deekshith Rao
Student,Computer Science and Engineering
Guru Nanak Institutions Technical Campus (Autonomous)
Hyderabad,Telangana,India-501506
raodeekshith12@gmail.com

Mrs.Hyma Birudaraju
Assistant Professor,Computer Science and Engineering
Guru Nanak Institutions Technical Campus (Autonomous)
Hyderabad,Telangana,India-501506
bhyma.gnitc@gniindia.org

***Abstract*— In order to overcome imbalanced datasets in healthcare fraud detection, the effort focuses on the Medicare Part B dataset. The hybrid resampling technique (SMOTE-ENN) is used with categorical feature extraction in this unique approach to balance the dataset. For fraud detection, logistic regression is used, and performance is assessed using a variety of measures. By using this method, problems with conventional resampling techniques like noise, overfitting, and information loss are lessened. The significance of AUPRC in situations with unbalanced data is emphasised by the study. Results demonstrate increased accuracy in detecting fraud, confirming the efficacy of the suggested approach.**

I. INTRODUCTION

Healthcare systems, particularly in the U.S., suffer significant financial losses due to fraud, with Medicare being a primary target. Fraudulent activities account for 3–10% of total healthcare costs, equating to losses between $19 billion and $65 billion annually. This not only affects financial resources but also undermines trust in healthcare services. Traditional fraud detection methods, such as rule-based systems, lack the flexibility to detect evolving fraud patterns. In contrast, machine learning (ML) offers a more adaptable solution by analyzing large-scale Medicare data to identify anomalies and suspicious behavior.

A major challenge in using ML for fraud detection is the imbalance in Medicare datasets, where non-fraudulent cases vastly outnumber fraudulent ones. This skew leads to poor model performance, especially in identifying the minority class (fraud cases), increasing the risk of false negatives. To overcome this, resampling techniques such as Random Oversampling (ROS), SMOTE, ADASYN, and Random Undersampling (RUS) are employed. However, each method has limitations—ROS can cause overfitting, SMOTE may introduce noise, and RUS risks losing valuable data.

This study addresses these challenges by proposing a novel approach combining separate handling of categorical features (e.g., Provider Type) and a SMOTE-ENN hybrid resampling method. This technique not only balances the dataset but also filters out noise. The approach is evaluated using ensemble classifiers and measured by the Area Under the Precision-Recall Curve (AUPRC), offering a more accurate assessment of model performance. The key contributions include a synthetic generation of categorical features and an integrated method to enhance fraud detection in imbalanced healthcare datasets.

II. RELATED WORK

Healthcare fraud detection has become a key area of research, especially with the growing application of artificial intelligence (AI) and machine learning (ML) techniques. Many studies have demonstrated how ML can identify suspicious patterns and fraudulent activities in healthcare systems. In addition, another important research focus is addressing the challenge of imbalanced datasets, common in fraud detection, where non-fraud cases heavily outnumber fraudulent ones. Tackling this issue is crucial for improving the accuracy of fraud detection models.

Several studies have proposed innovative AI-based methods for fraud detection. For instance, researchers have used large Medicare datasets and applied ensemble techniques, decision trees, and Bayesian networks to enhance detection performance. Feature engineering and data preparation have also been shown to play significant roles in improving model effectiveness. Although these approaches yield promising results, many of them are still limited by the heavily skewed data distributions found in real-world Medicare datasets.

To combat data imbalance, various resampling techniques have been introduced. Random Undersampling (RUS) and Random Oversampling (ROS) are commonly applied, each with its own advantages and limitations. RUS simplifies computation by reducing the volume of majority class data, but at the cost of losing potentially valuable information. In contrast, ROS duplicates minority class instances to balance the dataset, which may lead to overfitting and reduced model generalizability.

Other techniques include semantic embeddings and hybrid resampling. Semantic embedding methods convert medical codes into vector representations to enrich the feature space, and are often paired with resampling to improve class balance. Some studies have proposed combining over- and under-sampling approaches, although details on how these hybrid methods are implemented remain unclear. Despite these advancements, a clear understanding of their effectiveness across diverse datasets is still lacking.

Advanced techniques like ADASYN and Class Weighing Schemes have also been explored. ADASYN generates new samples based on the density of the minority class, offering greater variety in synthetic data. However, it can introduce noise, especially in complex healthcare datasets. Similarly, CWS adjusts model sensitivity toward minority classes, though its effectiveness varies depending on the classifier used. These studies suggest potential, but more work is needed to fine-tune such approaches for real-world healthcare data.

In conclusion, while machine learning offers strong potential for healthcare fraud detection, the issue of data imbalance remains a persistent challenge. There is a need for further exploration of methods like SMOTE and more robust hybrid strategies that generate diverse yet clean minority samples. Additionally, combining these data-balancing techniques with ensemble models could significantly enhance performance and accuracy. Future research should focus on developing scalable and interpretable models that are resilient to noise and capable of handling complex Medicare datasets.

III. PROBLEM STATEMENT AND EXISTING SYSTEM

Figure 1 illustrates the stark class imbalance in the Medicare Part B dataset, where fraudulent claims are significantly outnumbered by non-fraudulent ones at a ratio of 1:11,312. The effectiveness of conventional machine learning models is significantly impacted by this high skew, which biases them in favor of the dominant class and impairs their capacity to correctly identify fraudulent activity.

We recommend using the SMOTE-ENN approach to lessen this problem. We define a set of fraud detection models in this study, each trained on a different data partition Dm and represented by the notation $\{fm\}M_{m=1}$. Dm = $\{(x_{mi}, y_{mi})\}N_{mi=1}$, where x_{mi} indicates the input features for the i-th data point and y_{mi} is the associated label, contains pairs of feature vectors and class labels for each subset Dm.

To create a more balanced version of the dataset, D′m, the SMOTE-ENN approach is applied to each subset Dm. By removing noisy or unclear samples and synthesizing new minority class examples, this method improves the dataset. The formula for the transformation is zD′m = SMOTE-ENN(Dm). (1)

By eliminating inaccurate data points, this method seeks to improve the dataset and lessen class imbalance. The goal of combining noise reduction and synthetic oversampling is to improve the training data's overall quality.

Determining if SMOTE-ENN enhances model performance is one of the process's main goals. The effectiveness of each model fm trained on the modified dataset D′m will be compared to that of its counterpart trained on the original, imbalanced data Dm in order to assess this.

IV. PROPOSED SYSTEM AND ARCHITECTURE

To address the significant class imbalance in the Medicare Part B dataset, a hybrid approach is used combining SMOTE (Synthetic Minority Over-sampling Technique) and ENN (Edited Nearest Neighbors). This strategy helps both in augmenting minority class data and reducing noise from the dataset.

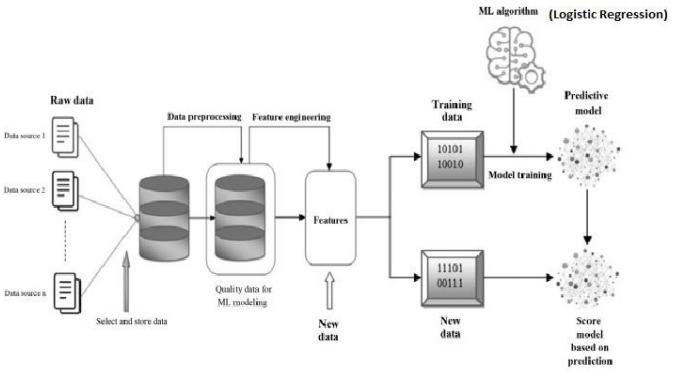

Fig:System Architecture for health care fraud detection based on SMOTE-ENN

Architecture Overview

The system architecture divides data into numerical and categorical types in order to address the class imbalance

imbalance by separating data into numerical and categorical types. Numerical features undergo balancing via SMOTE-ENN, while categorical data is expanded using random sampling without replacement. Ensemble classifiers are then used to detect fraudulent claims.

- **Data Collection**

Two datasets are used:

1. **Medicare Part B (2020)** – contains claims data with 29 features, primarily provider and service-related.
2. **LEIE** – The LEIE is a list of medical professionals who are not eligible for federal programs. It's used to label claims as fraudulent when conditions match specific exclusion criteria.

- **Preprocessing**

Data cleaning follows CMS guidelines, handling missing values and irrelevant features. A “Year” column is added, and categorical values like gender are encoded numerically. The LEIE dataset is merged to label records as fraud or non-fraud based on the NPI and exclusion date logic.

- **Data Splitting**

The dataset is divided into numerical and categorical subsets. The categorical feature, “Provider Type,” includes 102 classes and is augmented fairly using random sampling without replacement to avoid class bias.

- **Train-Test Split**

Data is split in an 80:20 ratio to create training and testing sets, ensuring reliable model evaluation on unseen data.

- **SMOTE-ENN Method**

SMOTE creates synthetic examples of the minority class using interpolation, while ENN removes noisy or misclassified instances. Together, they produce a cleaner, balanced dataset of over 7 million synthetic samples, boosting model generalization.

- **Classification Models**

For classification, six machine learning models are employed:

- **XGBoost**: High performance and efficient gradient boosting.
- **AdaBoost**: Focuses on correcting errors from weak learners.
- **LightGBM**: Fast and memory-efficient gradient boosting.
- **Random Forest (RF)** and **Decision Trees (DT)**: Tree-based models for classification.
- **Logistic Regression (LR)**: Baseline linear model.

V. ALGORITHM DESCRIPTION

VI. SMOTE-ENN Algorithm (in brief)

Purpose: to combine undersampling (ENN) and oversampling (SMOTE) in order to balance an unbalanced dataset.

Steps:

1. **Input:** Training dataset DmD_m with features and labels.
2. **Oversampling (SMOTE):**
- Randomly select a minority class sample.
- Find its *k* nearest minority neighbors.
- By interpolating between a neighbor and the chosen sample, create a synthetic sample.
- Add the synthetic sample to the dataset.
3. **Undersampling (ENN):**
    - Randomly select a sample from the dataset
    - .Find its *k* nearest neighbors.
    - If most neighbors are from the opposite (majority) class, remove the sample to reduce noise.
4. **Output:** A more balanced dataset Dm′D'_m with reduced class imbalance and noise.

VI. ADVANTAGES AND LIMITATIONS

Advantages:

1. **Cost Savings:**
    - Detecting fraud early helps prevent massive financial losses to healthcare systems and insurance providers.
2. **Improved Accuracy with AI/ML:**
    - can uncover hidden patterns and anomalies that humans might miss.
3. **Real-Time Detection:**
    - Advanced systems can flag suspicious claims in real-time, reducing delayed payouts or reimbursements.
4. **Enhanced Regulatory Compliance:**
    - Helps ensure adherence to legal standards, reducing the risk of lawsuits and penalties.
5. **Improved Service Delivery:**
    - Reducing fraud means resources are better allocated to genuine patients and services.
6. **Data-Driven Insights:**
    - Fraud detection systems can also provide insights into common fraud schemes, helping improve policies and prevention.

Limitations:

1. **False Positives/Negatives:**
    - Algorithms might wrongly flag legitimate claims or miss fraudulent ones, affecting trust and efficiency.
2. **Data Quality Issues:**
    - Incomplete, inconsistent, or incorrect data can hinder model accuracy and performance.
3. **Complex Fraud Tactics:**
    - Fraudsters continually adapt; static models may become outdated quickly without retraining.
4. **Privacy Concerns:**
    - Handling sensitive medical data raises ethical and legal challenges, especially regarding patient confidentiality.
5. **High Implementation Costs:**
    - Building and maintaining fraud detection systems, especially AI-based, can be expensive and resource-intensive.
6. **Dependence on Historical Data:**
    - Models trained on past fraud patterns may fail to detect novel or evolving fraudulent schemes.

VII. RESULTS AND DISCUSSIONS

The primary goal of our study is to evaluate various machine learning models and examine their effectiveness in improving fraud detection within the healthcare sector. This section presents a detailed analysis and discussion of the results derived from implementing the proposed methodology on the Medicare Part B dataset. We employ several classification algorithms—including Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), XGBoost, AdaBoost, and Light Gradient Boosting Machine (LGBM)—to perform the classification task. The performance of these models is assessed using key evaluation metrics such as accuracy, F1-score, precision, and recall. The SMOTE-ENN technique, which combines oversampling of minority class instances with the removal of overlapping or noisy samples, has demonstrated varying impacts across different classifiers. It

significantly enhances the performance of tree-based and ensemble models, while showing limited improvement for classifiers like Logistic Regression (LR) and AdaBoost. For example, Decision Trees (DT) achieve near-perfect performance, with an accuracy, F1-score, and recall of 0.99 using the train-test split method. In cross-validation, the model even reaches perfect scores of 1.00 across these metrics. These encouraging findings demonstrate how well the suggested strategy works to address class disparity. The varying outcomes among algorithms stem from LR's inherent linearity and AdaBoost's sensitivity to noisy data, which limit their ability to fully benefit from synthetic samples and noise reduction.

VIII. Conclusion and Future Work

This study highlights the importance of addressing data imbalance in healthcare fraud detection by proposing a novel machine learning framework that leverages the SMOTE-ENN hybrid resampling technique. By generating synthetic minority class samples and removing noisy data points, SMOTE-ENN effectively balances the dataset, leading to improved model accuracy. Additionally, the study incorporates evaluation metrics such as the Area Under the Curve (AUC) and Area Under the Precision-Recall Curve (AUPRC), which provide a more comprehensive assessment of model performance, especially important in highly imbalanced datasets. The AUPRC, in particular, proves crucial for evaluating predictive power in fraud detection scenarios. This framework offers a solid foundation for future research, encouraging SMOTE-ENN's applicability in various healthcare fraud contexts and integrating it with advanced AI technologies, such as deep learning, to further enhance detection capabilities.

References

[1] L. Morris, ''Combating fraud in health care: An essential component of any cost containment strategy,'' Health Affairs, vol. 28, no. 5, pp. 1351–1356,Sep. 2009.

[2] J. T. Hancock, R. A. Bauder, H. Wang, and T. M. Khoshgoftaar, ''Explainable machine learning models for medicare fraud detection,'' J. Big Data, vol. 10, no. 1, p. 154, Oct. 2023.

[3] A. Alanazi, ''Using machine learning for healthcare challenges and opportunities,'' Informat. Med. Unlocked, vol. 30, 2022, Art. no. 100924.

[4] R. A. Bauder and T. M. Khoshgoftaar, ''The detection of medicare fraud using machine learning methods with excluded provider labels,'' in Proc.Thirty-First Int. Flairs Conf., 2018, pp. 1–6.

[5] R. A. Bauder and T. M. Khoshgoftaar, ''Medicare fraud detection using machine learning methods,'' in Proc. 16th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA), Dec. 2017, pp. 858–865.

[6] V. Nalluri, J.-R. Chang, L.-S. Chen, and J.-C. Chen, ''Building prediction models and discovering important factors of health insurance fraud using machine learning methods,'' J. Ambient Intell. Humanized Comput. vol. 14, no. 7, pp. 9607–9619, Jul. 2023.

[7] P. Dua and S. Bais, ''Supervised learning methods for fraud detection in healthcare insurance,'' in Machine Learning in Healthcare Informatics (Intelligent Systems Reference Library), vol. 56, S. Dua, U. Acharya, and P. Dua, Eds. Berlin, Germany: Springer, 2014, doi: 10.1007/978-3-642-40017-9_12.

[8] R. Bauder, R. da Rosa, and T. Khoshgoftaar, ''Identifying medicare provider fraud with unsupervised machine learning,'' in Proc. IEEE Int. Conf. Inf. Reuse Integr. (IRI), Jul. 2018, pp. 285–292.

[9] Centers for Medicare and Medicaid Services. (2017). Research, Statistics, Data, and Systems. [Online]. Available: https://www.cms.gov/researchstatistics-data-and-systems/research-statistics data-and-systems.html

[10] P. Brennan, ''A comprehensive survey of methods for overcoming the class imbalance problem in fraud detection,'' Inst. Technol. Blanchardstown Dublin, Dublin, Ireland, Tech. Rep., 2012.

[11] N. Agrawal and S. Panigrahi, ''A comparative analysis of fraud detection in healthcare using data balancing & machine learning techniques,'' in Proc. Int. Conf. Commun., Circuits, Syst. (IC3S), May 2023, pp. 1–4.

[12] M. Herland, R. A. Bauder, and T. M. Khoshgoftaar, ''The effects of class rarity on the evaluation of supervised healthcare fraud detection models,'' J. Big Data, vol. 6, no. 1, pp. 1–33, Dec. 2019.

[13] J. Hancock, T. M. Khoshgoftaar, and J. M. Johnson, ''The effects of random undersampling for big data medicare fraud detection,'' in Proc. IEEE Int. Conf. Service-Oriented Syst. Eng. (SOSE), Aug. 2022, pp. 141–146.

[14] A. Mehbodniya, I. Alam, S. Pande, R. Neware, K. P. Rane, M. Shabaz, and M. V. Madhavan, ''Financial fraud detection in healthcare using machine learning and deep learning techniques,'' Secur. Commun. Netw., vol. 2021, pp. 1–8, Sep. 2021.

[15] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, ''Learning from class-imbalanced data: Review of methods and applications,'' Expert Syst. Appl., vol. 73, pp. 220 239, May 2017.

[16] J. Hancock and T. M. Khoshgoftaar, ''Optimizing ensemble trees for big data healthcare fraud detection,'' in Proc. IEEE 23rd Int. Conf. Inf. Reuse Integr. Data Sci. (IRI), Aug. 2022, pp. 243 249.

[17] N. Kumaraswamy, M. K. Markey, J. C. Barner, and K. Rascati, ''Feature engineering to detect fraud using healthcare claims data,'' Expert Syst. Appl., vol. 210, Dec. 2022, Art. no. 118433.

[18] N. Kumaraswamy, T. Ekin, C. Park, M. K. Markey, J. C. Barner, and K. Rascati, ''Using a Bayesian belief network to detect healthcare fraud,'' Expert Syst. Appl., vol. 238, Mar. 2024, Art. no. 122241.

[19] J. M. Johnson and T. M. Khoshgoftaar, ''Data-centric AI for healthcare fraud detection,'' Social Netw. Comput. Sci., vol. 4, no. 4, p. 389, May 2023.

[20] R. A. Bauder and T. M. Khoshgoftaar, ''The effects of varying class distribution on learner behavior for medicare fraud detection with imbalanced big data,'' Health Inf. Sci. Syst., vol. 6, no. 1, pp. 1–14,Dec. 2018.

[21] R. A. Bauder, T. M. Khoshgoftaar, and T. Hasanin, ''Data sampling approaches with severely imbalanced big data for medicare fraud detection,'' in Proc. IEEE 30th Int.

Conf. Tools Artif. Intell. (ICTAI), Nov. 2018, pp. 137–142.
[22] J. M. Johnson and T. M. Khoshgoftaar, ‘‘Hcpcs2Vec: Healthcare procedure embeddings for medicare fraud prediction,’’ in Proc. IEEE 6th Int. Conf. Collaboration Internet Comput. (CIC), Dec. 2020, pp. 145–152.
[23] J. M. Johnson and T. M. Khoshgoftaar, ‘‘Medical provider embeddings for healthcare fraud detection,’’ Social Netw. Comput. Sci., vol. 2, no. 4, p. 276, Jul. 2021. [Online]. Available: https://link.springer.com/10.1007/s42979-021-00656-y
[24] M. Suesserman, S. Gorny, D. Lasaga, J. Helms, D. Olson, E. Bowen,and S. Bhattacharya, ‘‘Procedure code overutilization detection from healthcare claims using unsupervised deep learning methods,’’ BMC Med. Informat. Decis. Making, vol. 23, no. 1, p. 196, Sep. 2023.
.