ENHANCING MEDICARE FRAUD DETECTION TROUGH MACHINE LEARNING

Azra Ismath Mohammed[1], Samreen Sultana[2]

[1]*PG Scholar, Department of CSE, Shadan Women's College of Engineering and Technology, Hyderabad, azraismath@gmail.com*

[2]*Asst. Professor, Department of CSE, Shadan Women's College of Engineering and Technology*

samreencme@gmail.com

ABSTRACT

The identification of healthcare fraud is a crucial problem that is complicated by unbalanced datasets, which frequently lead to less-than-ideal model performance. Traditional machine learning (ML) techniques have been the mainstay of previous research, but they suffer from problems like overfitting brought on by Random Oversampling (ROS), noise introduced by the Synthetic Minority Oversampling Technique (SMOTE), and the loss of important information due to Random Undersampling (RUS). In this work, we focus on the Medicare Part B dataset and provide a unique solution to the unbalanced data issue in healthcare fraud detection. We start by carefully extracting the category characteristic "Provider Type," which enables the creation of new, synthetic instances by copying pre-existing types to increase diversity in the minority class.

We use SMOTE-ENN, a hybrid resampling strategy that combines the Synthetic Minority Oversampling strategy (SMOTE) and Edited Nearest Neighbors (ENN) to create synthetic data points while eliminating noisy, unnecessary instances, in order to better balance the dataset. In addition to balancing the dataset, this combination strategy aids in reducing the possible negative consequences of unbalanced data. With the use of standard assessment measures including accuracy, F1 score, recall, precision, and the AUC-ROC curve, we assess the logistic regression model's performance on the resampled dataset. Furthermore, we stress the significance of the Area Under the Precision-Recall Curve (AUPRC) as a crucial statistic for assessing model performance in situations when there is an imbalance. With an astounding 98% accuracy rate, the experimental results demonstrate that logistic regression performs better than other methods and proves the efficacy of our proposed approach for detecting healthcare fraud in unbalanced datasets.

INTRODUCTION

Detecting healthcare fraud is essential to protecting healthcare systems from financial exploitation, which can result in needless expenses, subpar patient treatment, and a decline in public confidence in healthcare organizations. Healthcare costs are greatly increased by fraudulent claims, such as invoicing for treatments that were never provided or performing needless activities to boost compensation. But because of the enormous amount healthcare data, difficult to identify these fraudulent operations. Fraud-detection is made more difficult by the uneven distribution created by fraudulent claims often make up a small percentage of the millions of claims in huge databases. Healthcare fraud detection has been approached using conventional machine learning methods like Random Forests, Decision Trees, and Logistic Regression; however, they are not without their limitations. The unbalanced dataset problem, in which (non-fraudulent claims) outnumbers (fraudulent claims), causes biassed predictions in many of these models. The algorithms' tendency might result in an imbalance that makes it difficult for models since they generate a lot, or fraudulent cases that are incorrectly labeled as non-fraudulent. A number of overcome the issues unbalanced data, such as Under sampling (RUS). By replicating existing minority class instances, ROS expands the number of minority class instances; yet, by adding redundant information, it runs the danger of overfitting. Although SMOTE can generate noise or outliers, it produces synthetic examples. RUS decreases, which might result in important data. While these strategies assist balance the dataset, set of constraints that might impair the overall efficacy of the fraud detection algorithms. Using a hybrid resampling technique dubbed SMOTE-ENN (Synthetic Minority Oversampling Technique-Edited Nearest Neighbors), we provide a unique approach to healthcare fraud detection in this work that directly solves the issues of dataset imbalance. SMOTE-ENN eliminates noisy and redundant examples from the dataset by combining the advantages of Edited Nearest Neighbors (ENN) and SMOTE for creating synthetic minority class instances.

OBJECTIVE

With an emphasis on the Medicare Part B dataset, unique strategy for resolving the inherent difficulties presented by unbalanced datasets in healthcare fraud detection. Fraudulent class and removing noisy, unnecessary instances from the majority (non-fraudulent) class, the hybrid resampling technique SMOTE-ENN (Synthetic Minority Oversampling Technique-Edited Nearest Neighbors) attempts. The model may overcome the bias frequently found in conventional by better differentiating between fraudulent and non-

fraudulent claims by enhancing the balance of the dataset. In situations when fraud instances valid ones, this method is intended to greatly improve systems. This study is logistic regression as a classifier on the resampled dataset, resolving the imbalance in the dataset. The objective is to show how hybrid resampling methods, like SMOTE-ENN, may enhance a number of performance indicators and classification accuracy. With an emphasis on evaluating false claims in comparison to conventional techniques, these metrics include the F1 score, recall, precision, and AUC-ROC. Through the use of these performance criteria, the study aims to offer assessment efficacy in the setting of unbalanced data, when conventional accuracy might not be enough for precise fraud detection. This experiment also demonstrates employing comprehensive evaluation metrics, particularly the AUPRC, as a critical performance metric for unbalanced datasets. Because fraudulent claims are frequently rare, traditional metrics like model's ability to identify fraudulent activity. AUPRC offers a more accurate imbalanced classes, because it accounts for the minority class's precision and recall, ensuring accurately evaluated. Lastly, by verifying the potential use operational healthcare fraud detection systems, demonstrate its practicality. The techniques employed can be expanded, extended, and successfully implemented to manage the intricacies of unbalanced and big healthcare datasets, providing a more dependable means of identifying false claims in actual healthcare environments.

PROBLEM STATEMENT

Healthcare fraud detection is a critical task, but it is severely challenged by the presence of highly unbalanced datasets, such as Medicare Part B records, where fraudulent cases are much rarer than legitimate ones. Traditional machine learning approaches struggle in this context, as common resampling techniques like Random Oversampling (ROS), SMOTE, and Random Under sampling (RUS) can introduce overfitting, noise, or loss of important information, leading to suboptimal model performance. There is a need for a robust method that can effectively handle class imbalance, improve the diversity of minority class instances, and reduce noise, while ensuring reliable detection of fraudulent activities.

EXISTING SYSTEM

Traditional ML methods categorize instances of fraudulent activity in healthcare datasets, including the Medicare Part B dataset, and are a major component of healthcare fraud detection systems. These systems frequently employ Decision Trees, Random Forests, SVMs, and Logistic Regression as algorithms. Based on characteristics extracted from medical claims data, these models are trained to differentiate between fraudulent (minority class) and non-fraudulent (majority class) actions. The capacity of the models to correctly identify fraudulent cases while keeping valid transactions is a major factor in how well they function. However, the persistence of class inequality remains a significant barrier. Since fraudulent cases are rare non-fraudulent claims, (fraudulent cases) is underrepresented in most healthcare statistics. Traditional machine-learning algorithms are more likely to correctly predict the majority class (non-fraudulent claims) while failing to detect fraudulent instances, this discrepancy could lead to biased model training. This is especially troublesome when it comes to healthcare fraud detection, because detecting false claims is essential to minimizing losses and maintaining system integrity.

Disadvantage of Existing System

- Because ROS duplicates existing instances, it might result in overfitting by oversampling the minority class.
- The synthetic instances may overlap with the majority class as a result.
- RUS may lead to underfitting and an underrepresented and uninformative majority class.

PROPOSED SYSTEM

We suggest a unique method to enhance model performance, specifically for the Medicare Part B dataset, difficulties presented by unbalanced datasets in healthcare fraud detection. Overfitting (caused by Random Oversampling), noise introduction (caused by SMOTE), and information loss (caused by Random Undersampling) have all been problems for the current systems. By concentrating on enhancing the dataset's balance and making sure more varied, pertinent data, our suggested solution seeks to address these shortcomings. The approach begins with a meticulous feature engineering stage in which the Medicare Part B dataset's categorical feature "Provider Type" is extracted. Representativeness generating synthetic instances that add more variety by copying the current "Provider Type" values. Because logistic regression works effectively for identifying fraud in unbalanced datasets, the system uses it as its classification method. Accuracy, F1 score, recall, precision, AUC-ROC, Precision-Recall Curve (AUPRC), which is important in situations involving unbalanced datasets, are some of the measures we use to assess the model's performance. To guarantee excellent detection performance, the suggested solution efficiently balances the dataset, improves model training, and offers a thorough assessment.

Advantages of Proposed System

- Because it forecasts probabilities, logistic regression is helpful for confidence estimates and ranking.
- Particularly for small to medium-sized datasets, it is straightforward and computationally efficient.
- Compared to more intricate models like ensemble techniques or neural networks, it uses fewer computer resources.
- In situations involving binary classification, it serves as a trustworthy baseline model for comparison.

RELATED WORKS

Healthcare fraud detection has been widely studied using traditional machine learning (ML) techniques, such as logistic regression, decision trees, and support vector machines. Many studies focus on feature engineering and statistical analysis to identify patterns indicative of fraud in large healthcare datasets like Medicare claims. While these methods have shown promise in detecting fraudulent claims, their performance often degrades when dealing with unbalanced datasets. Techniques such as Random Oversampling (ROS) and Random Undersampling (RUS) have been applied to address class imbalance, but ROS can lead to overfitting, and RUS may discard critical information, reducing model generalization. Consequently, the challenge of effectively detecting healthcare fraud persists in real-world scenarios.

Recent research has explored the use of advanced oversampling strategies like the Synthetic Minority Oversampling Technique (SMOTE) to generate synthetic minority instances and improve classifier performance. SMOTE has been combined with ensemble learning methods such as Random Forests and XGBoost to enhance detection accuracy. However, SMOTE can introduce noisy samples and outliers, which negatively affect model performance. Hybrid approaches, such as SMOTE with Edited Nearest Neighbors (SMOTE-ENN), have been proposed to mitigate these drawbacks by simultaneously oversampling the minority class and removing potentially noisy points. These hybrid resampling strategies have shown improved performance in highly imbalanced datasets across various domains, including healthcare fraud detection.

Feature selection and categorical variable handling have also been emphasized in recent works. Studies highlight that domain-specific features, like provider types, procedure codes, and claim amounts, are critical in improving model discrimination between fraudulent and legitimate claims. Researchers have experimented with feature transformation and embedding techniques to enhance the representation of categorical data. Additionally, evaluation metrics beyond accuracy, such as F1-score, AUC-ROC, and Area Under Precision-Recall Curve (AUPRC), have been increasingly recommended for assessing model performance on imbalanced datasets. These advancements provide a foundation for combining effective resampling, feature engineering, and robust evaluation strategies to detect healthcare fraud reliably.

METHODOLOGY OF PROJECT

In this study, we address the challenge of healthcare fraud detection in highly imbalanced datasets using a hybrid resampling and machine learning approach. We begin by preprocessing the Medicare Part B dataset and carefully extracting the categorical feature “Provider Type” to enhance minority class representation. To tackle class imbalance, we employ the SMOTE-ENN technique, which generates synthetic instances for the minority class while removing noisy or ambiguous samples, ensuring a cleaner and more balanced dataset. Logistic regression is then trained on the resampled data, and model performance is evaluated using multiple metrics, including accuracy, precision, recall, F1-score, AUC-ROC, and AUPRC, with special emphasis on precision-recall evaluations.

MODULE DESCRIPTION:

Data Collection:

Obtaining the pertinent data, in this case the Medicare Part B information, is the first step. Information on healthcare claims, including patient and provider identities, billing amounts, and whether or not the claim is false, is included in this dataset. The machine learning models are trained and assessed using this data as the basis.

Data Preparation:

Preparing the data for analysis comes next once it has been collected. This entails resolving missing values, fixing discrepancies, and converting categorical variables into numerical representations in order to clean the data. A crucial component of this process is feature engineering, which involves extracting important characteristics like "Provider Type" and perhaps creating more synthetic features to enhance the model's functionality.

Data Splitting:

Following preparation, the data is divided into training and testing datasets. The model is typically trained using 70–80% of the data, with the remaining 20–30% set aside for testing and validation. This division makes it possible to test the model's capacity for generalization using data that hasn't been seen yet.

Model Training:
A machine learning model—in this example, logistic regression—is trained on the training dataset during the training phase. In order to ensure that the model learns to properly identify both the majority (non-fraudulent) and minority (fraudulent) classes, the resampling approach SMOTE-ENN is used to solve the issue of class imbalance in the dataset. To maximize the model's performance, hyperparameter adjustment may also be done at this stage.

Model Evaluation:
The model is assessed using a range of performance indicators following training in order to determine its efficacy. The model's capacity to identify false claims is evaluated using key measures including accuracy, precision, recall, F1 score, AUC-ROC, and AUPRC, particularly when considering the extremely unbalanced dataset. The model's performance on both the minority and majority classes may be determined with the use of these indicators.

Model Prediction:
Lastly, fresh, unknown data is predicted using the trained model. This makes it possible to spot possibly false medical claims in real-time situations. The predicted accuracy and dependability of the model in identifying fraud are then evaluated by contrasting its predictions with the actual results.

ALGORITHM USED IN PROJECT

In this project, Logistic Regression is employed as the primary classification algorithm for detecting healthcare fraud in the Medicare Part B dataset. Logistic Regression is a widely used statistical method for binary classification, as it estimates the probability of a particular outcome based on input features. It is especially suitable for datasets with categorical variables, like "Provider Type," and performs well in cases of moderate imbalance when combined with appropriate resampling techniques. Its simplicity, interpretability, and efficiency make it an ideal choice for large healthcare datasets, where understanding feature contributions is as important as achieving high predictive performance.

To address the severe class imbalance, Logistic Regression is trained on a dataset balanced using SMOTE-ENN, a hybrid resampling technique. SMOTE generates synthetic minority class instances to enhance diversity, while ENN removes noisy or misclassified points to reduce the risk of overfitting. This combination ensures that Logistic Regression can learn meaningful patterns from both majority and minority classes, improving detection accuracy, recall, precision, and overall reliability in identifying fraudulent healthcare claims.

DATA FLOW DIAGRAM

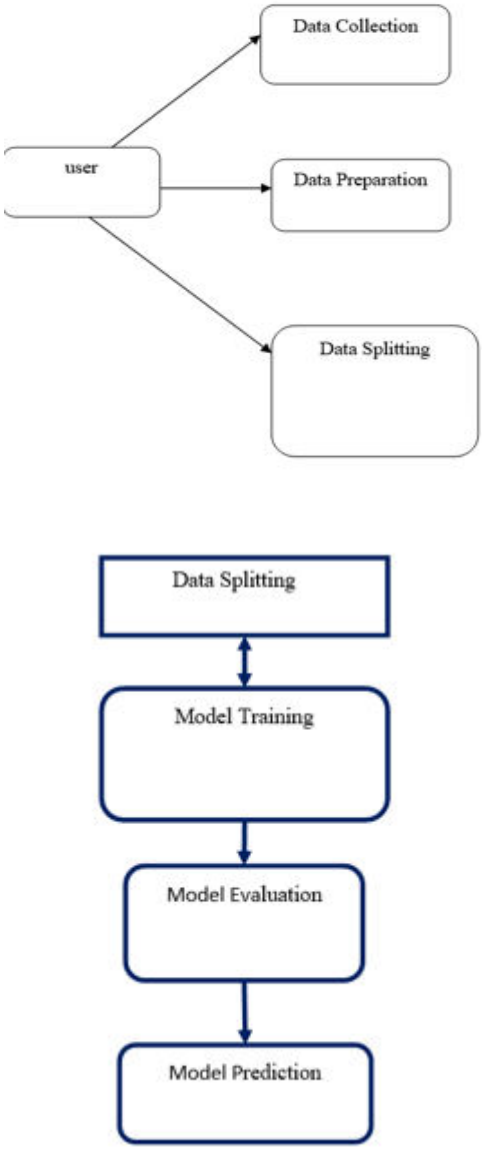

Fig: 1 Flow Diagram

SYSTEM ARCHITECTURE

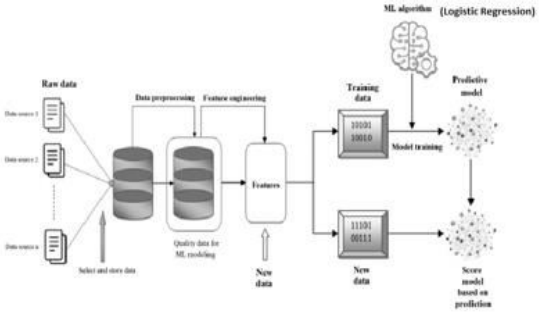

Fig: 2 SYSTEM ARCHITECTURE OF PROJECT

RESULTS

Fig 3: Home Screen

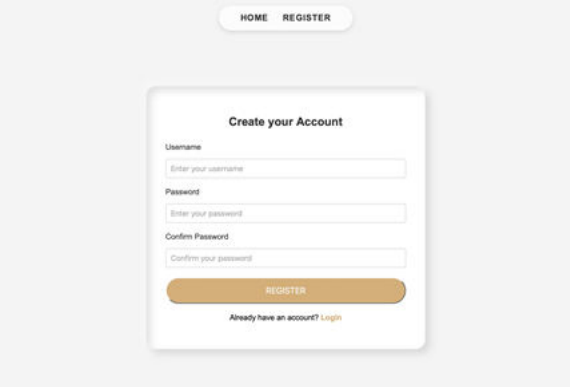

Fig 4: User Registration

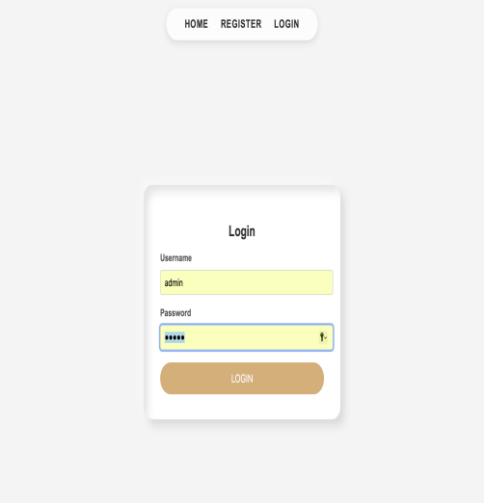

Fig 5: User Login

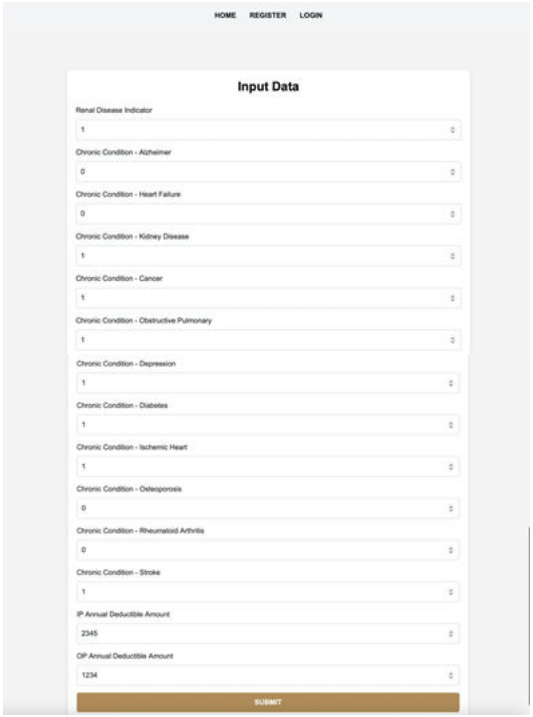

Fig 6: Input Screen

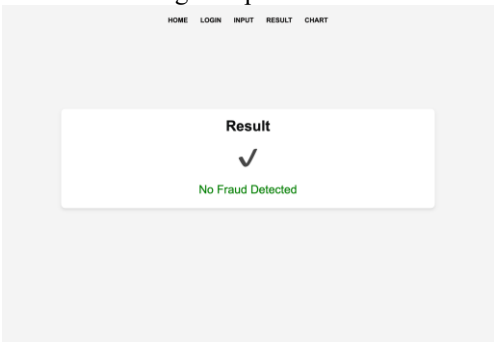

Fig 7: Result

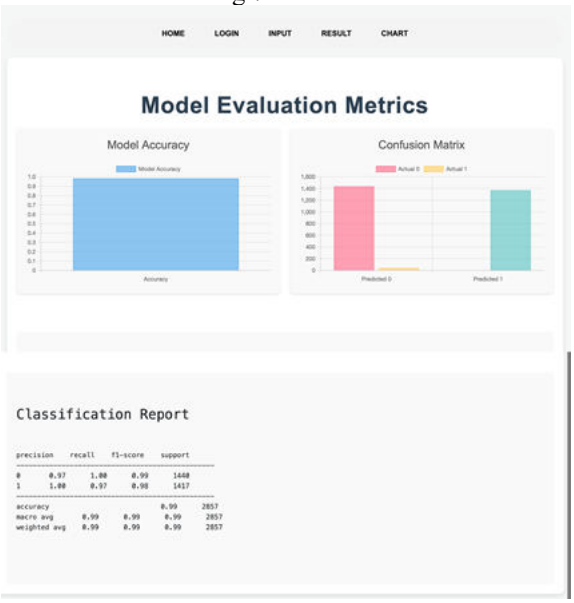

Fig 8: Model Evaluation Metrics

FUTURE ENHANCEMENT

Future research can investigate cutting-edge AI approaches, such as DL architectures like

Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to expand on the study. Identifying intricate, non-linear patterns in medical data that conventional models could overlook. Additionally, by integrating the advantages of several models, ensemble learning techniques like stacking, bagging, and boosting can improve prediction performance. This strategy would result in more reliable systems that may increase accuracy and better manage a variety of fraud detection scenarios.

Additional crucial topics for future researches include scalability and generality. To make sure the SMOTE-ENN framework is reliable and flexible across a range of healthcare fraud scenarios, it should be tested on bigger and more varied datasets. Furthermore, to make sure the system can react enhancing operational efficiency, validating the model's effectiveness in scenarios will be crucial. In dynamic healthcare settings, this real-time capacity is especially crucial for the prompt identification and avoidance of fraudulent claims.

CONCLUSION

This study emphasizes how urgently the ongoing issues with unbalanced datasets in healthcare fraud detection must be addressed. Effective because fraudulent actions in healthcare systems place a heavy financial and operational strain on these systems. By combining the creation of synthetic instances with the elimination of unnecessary and noisy data, the suggested framework—which is based on the SMOTE-ENN hybrid resampling method—has effectively addressed the problem of data imbalance. In addition to improving dataset quality, this dual strategy raises machine learning models' overall performance and dependability.

This framework's use of logistic regression demonstrated how well it can handle binary classification issues while still being interpretable and computationally economical. Furthermore, a full evaluation of the model's capabilities was assured by the inclusion of comprehensive evaluation measures including Accuracy, F1 Score, Precision, Recall, AUC, and AUPRC. AUPRC was particularly useful in assessing performance in unbalanced situations, reflecting the harmony between recall and precision in identifying the minority class (fraudulent instances). The 98% accuracy attained confirms the suggested framework's strength and effectiveness and shows that it has the potential to be a useful tool for real-world healthcare fraud detection applications. Along with its technological advantages, this study offers and advancement in this field. Researchers and practitioners may use the methods and results to help develop are scalable and flexible enough to accommodate a variety of healthcare datasets and changing fraud trends.

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.