Automated Identification of Unusual Medicare Claim Activities Using Machine Learning

K.Manjunath, department of Computer Science and Engineering, GNITC, 22-5F6, 22wj1a05f6@gniindia.org
K.Anand Kumar, department of Computer Science and Engineering, GNITC, 22-5G1, 22wj1a05g1@gniindia.org
L.Dhanush, department of Computer Science and Engineering, GNITC, 23-519, 23wj5a0519@gniindia.org
Ms.Rajashree Sutrawe, Associate Professor, department of Computer Science and Engineering, GNITC

---***---

Abstract - Healthcare fraud detection is an important challenge in modern healthcare systems due to the large volume of medical claims and the highly imbalanced nature of fraud datasets. Fraudulent claims represent only a small portion of the total claims, which makes it difficult for traditional machine learning models to accurately identify fraudulent activities. Conventional techniques such as Random Oversampling (ROS), Synthetic Minority Oversampling Technique (SMOTE), and Random Undersampling (RUS) are commonly used to address class imbalance, but they often lead to issues such as overfitting, noise generation, or loss of important information. In this study, a machine learning-based approach is proposed to improve Medicare fraud detection using the Medicare Part B dataset. The proposed framework applies a hybrid resampling technique called SMOTE-ENN, which combines SMOTE for generating synthetic minority samples with Edited Nearest Neighbors (ENN) to remove noisy and irrelevant data instances. Logistic Regression is used as the classification algorithm to detect fraudulent healthcare claims. The model performance is evaluated using multiple metrics, including accuracy, precision, recall, F1-score, AUC-ROC, and Area Under the Precision-Recall Curve (AUPRC). Experimental results demonstrate that the proposed approach achieves an accuracy of 98%, indicating its effectiveness in handling imbalanced datasets and improving fraud detection in healthcare systems.

Key Words: Healthcare Fraud Detection, Machine Learning, SMOTE-ENN, Logistic Regression, Imbalanced Data, Medicare Claims.

1. INTRODUCTION

Healthcare systems worldwide handle a massive volume of medical claims and financial transactions every day. While digital healthcare services have improved accessibility and efficiency, they have also increased the risk of fraudulent activities. Healthcare fraud occurs when individuals or organizations intentionally submit false or misleading claims to obtain unauthorized financial benefits. Examples include billing for services that were not provided, performing unnecessary medical procedures, or inflating the cost of treatments. Such fraudulent practices lead to significant financial losses for healthcare providers, insurance companies, and government healthcare programs such as Medicare. Detecting fraudulent claims is a challenging task because healthcare datasets are typically very large and highly imbalanced. In most cases, fraudulent claims represent only a very small percentage of the total claims, while the majority are legitimate. This imbalance makes it difficult for traditional machine learning algorithms to accurately detect fraud, as these models tend to favor the majority class during training. As a result, many fraudulent cases may go undetected, which reduces the effectiveness of existing fraud detection systems.Several machine learning techniques have been applied to healthcare fraud detection, including Decision Trees, Random Forests, Support Vector Machines, and Logistic Regression. However, these models often struggle when dealing with imbalanced datasets. Common data balancing techniques such as Random Oversampling (ROS), Synthetic Minority Oversampling Technique (SMOTE), and Random Undersampling (RUS) have been used to address this issue. Despite their effectiveness in certain scenarios, these methods may introduce problems such as overfitting, noise generation, or loss of important information.

To address these challenges, this research proposes a machine learning-based approach for Medicare fraud detection using a hybrid resampling technique called SMOTE-ENN. This method combines the advantages of SMOTE and Edited Nearest Neighbors (ENN) to generate synthetic minority samples while removing noisy data instances. A Logistic Regression model is then trained on the balanced dataset to identify fraudulent healthcare claims. The proposed system aims to improve fraud detection accuracy and provide a reliable solution for handling imbalanced healthcare datasets.

2. LITERATURE REVIEW

Healthcare fraud detection has become an important research topic due to the increasing number of fraudulent medical claims and the financial losses they cause in healthcare systems. Researchers have applied various machine learning techniques to identify fraudulent healthcare activities and improve detection accuracy.Hancock et al. (2023) proposed explainable machine learning models for Medicare fraud detection using an ensemble supervised feature selection technique. Their study focused on improving model interpretability and reducing dataset dimensionality while maintaining high performance. The authors demonstrated that the proposed feature selection method could reduce dataset dimensionality by approximately 87.5% without affecting model performance, which improves efficiency and reduces the risk of overfitting.

Nalluri et al. (2023) developed prediction models for healthcare insurance fraud using multiple machine learning algorithms, including Support Vector Machines (SVM), Decision Trees (DT), Random Forest (RF), and Multilayer Perceptron (MLP). Their study aimed to identify important factors contributing to fraudulent medical claims. The results showed that Decision Tree models could effectively identify key fraud-related features such as healthcare service providers, insurance claim amounts, and procedure codes.

Agrawal and Panigrahi (2023) conducted a comparative analysis of healthcare fraud detection using different machine learning algorithms combined with data balancing techniques. Their study highlighted that fraud detection systems can significantly benefit from machine learning models that are capable of identifying complex fraud patterns in healthcare data.

Suesserman et al. (2023) investigated the use of unsupervised deep learning methods for detecting procedure code overutilization in healthcare claims. Their research applied deep autoencoder models to identify anomalous patterns in healthcare claims data. The experimental results showed that deep learning approaches can effectively detect unusual healthcare procedures that may indicate fraudulent activities.

Although previous studies have made significant progress in healthcare fraud detection, many existing approaches still face challenges related to highly imbalanced datasets. Fraudulent claims represent only a small portion of the data, which makes accurate classification difficult. Therefore, this study proposes the use of a hybrid resampling technique, SMOTE-ENN, combined with logistic regression to improve fraud detection performance and effectively handle imbalanced healthcare datasets.

3. RELATED WORK

Healthcare fraud detection has been widely studied using different machine learning and data mining techniques to identify fraudulent insurance claims. Due to the increasing volume of healthcare data and the presence of highly imbalanced datasets, researchers have explored several approaches to improve fraud detection accuracy.Hancock et al. (2023) proposed explainable machine learning models for Medicare fraud detection by applying an ensemble supervised feature selection technique. Their approach focused on reducing dataset dimensionality and improving model interpretability while maintaining classification performance. The study demonstrated that feature selection could significantly reduce the number of features while still preserving the predictive capability of fraud detection models.Nalluri et al. (2023) investigated healthcare insurance fraud detection using several machine learning algorithms such as Support Vector Machines (SVM), Decision Trees (DT), Random Forest (RF), and Multilayer Perceptron (MLP). Their research aimed to identify important factors that contribute to fraudulent claims. The results indicated that machine learning models can effectively analyze complex healthcare datasets and detect fraudulent activities.Agrawal and Panigrahi (2023) conducted a comparative analysis of fraud detection techniques in healthcare using different machine learning algorithms and data balancing approaches. Their study highlighted that data imbalance significantly affects the performance of fraud detection systems and emphasized the need for effective data preprocessing and balancing techniques.Suesserman et al. (2023) explored unsupervised deep learning techniques to detect anomalies in healthcare claims data. Their study used deep autoencoder models to identify unusual procedure codes that could indicate fraudulent behavior. The experimental results showed that deep learning models are capable of identifying abnormal patterns in large-scale healthcare datasets.

Although these studies have made significant contributions to healthcare fraud detection, many existing methods still struggle with the problem of highly imbalanced datasets. Fraudulent claims typically represent only a small percentage of the total healthcare data, which can reduce the effectiveness of traditional machine learning models. Therefore, this study proposes a hybrid resampling approach using SMOTE-ENN combined with logistic regression to improve fraud detection performance and address the imbalance problem in Medicare fraud datasets.

4. PROPOSED METHODOLOGY

To address the challenges associated with imbalanced datasets in healthcare fraud detection, this study proposes a machine learning-based framework that improves the identification of fraudulent Medicare claims. The proposed system focuses on improving dataset balance and model performance by combining feature extraction, hybrid resampling techniques, and classification algorithms.

4.1 Data Collection

The first step of the proposed system involves collecting the Medicare Part B dataset, which contains healthcare claim information such as provider details, billing amounts, procedure codes, and claim status. This dataset serves as the primary source for training and evaluating the fraud detection model.

4.2 Data Preparation

In the data preparation stage, the collected dataset is cleaned and processed to improve its quality and usability. Missing values are handled, inconsistent data entries are corrected, and categorical variables are transformed into numerical formats suitable for machine learning models. Feature engineering is also performed by extracting important attributes such as Provider Type, which helps enhance the diversity of the minority class in the dataset.

4.3 Data Splitting

After preprocessing, the dataset is divided into training and testing subsets. Typically, 70–80% of the data is used for training the model, while the remaining 20–30% is reserved for testing. This division ensures that the model is evaluated using unseen data to measure its generalization capability.

4.4 Data Balancing using SMOTE-ENN

Healthcare fraud datasets are highly imbalanced, where fraudulent claims represent only a small portion of the total data. To address this issue, a hybrid resampling technique called SMOTE-ENN is applied. SMOTE generates synthetic samples for the minority class, increasing the number of fraudulent instances, while Edited Nearest Neighbors (ENN) removes noisy and irrelevant data points. This combination helps create a cleaner and more balanced dataset for model training.

4.5 Model Training

Once the dataset is balanced, a **Logistic Regression** model is trained using the processed data. Logistic regression is selected because of its simplicity, interpretability, and effectiveness in binary classification tasks such as fraud detection. The model learns patterns and relationships between the dataset features and the target variable to identify fraudulent claims.

4.6 Model Evaluation

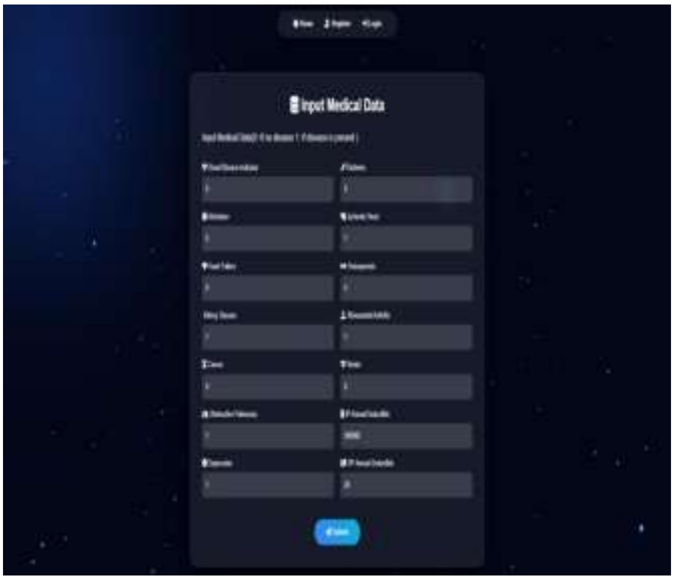
The performance of the trained model is evaluated using multiple metrics, including **accuracy, precision, recall, F1-score, AUC-ROC, and AUPRC**. These evaluation metrics provide a comprehensive assessment of the model's performance, particularly in handling imbalanced datasets.

4.7 Fraud Prediction

Finally, the trained model is used to predict fraudulent healthcare claims from new or unseen data. By analyzing claim features, the model can classify each claim as either fraudulent or legitimate, enabling healthcare organizations to detect and prevent fraudulent activities more effectively.

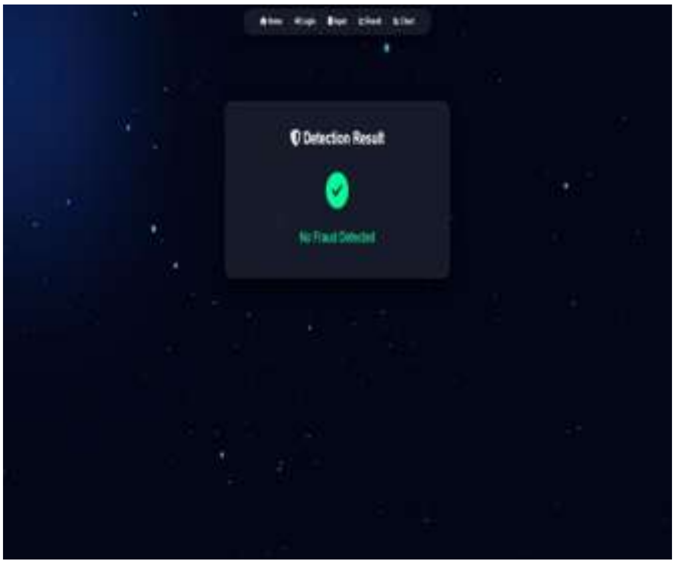
5. RESULTS AND DISCUSSION

The experimental evaluation of the proposed healthcare fraud detection framework demonstrates its effectiveness in handling imbalanced datasets and identifying fraudulent Medicare claims. The system applies the hybrid resampling technique SMOTE-ENN to balance the dataset and uses Logistic Regression as the classification model to detect fraudulent healthcare activities.Initially, the dataset undergoes preprocessing and feature extraction to improve data quality and ensure that the model can effectively learn from the available information. The Provider Type feature is extracted as an important attribute for generating synthetic instances of the minority class. After preprocessing, the dataset is balanced using the SMOTE-ENN technique, which generates synthetic minority samples while removing noisy or irrelevant data instances. This process improves the dataset quality and reduces the bias toward the majority class.The balanced dataset is then used to train the Logistic Regression model. The model is evaluated using several performance metrics, including accuracy, precision, recall, F1-score, AUC-ROC, and AUPRC. These metrics provide a comprehensive evaluation of the model's ability to detect fraudulent claims, especially in the presence of imbalanced data.



Input Medical Data

Fig. 1. Medical Data Input Interface of the Proposed Fraud Detection System.

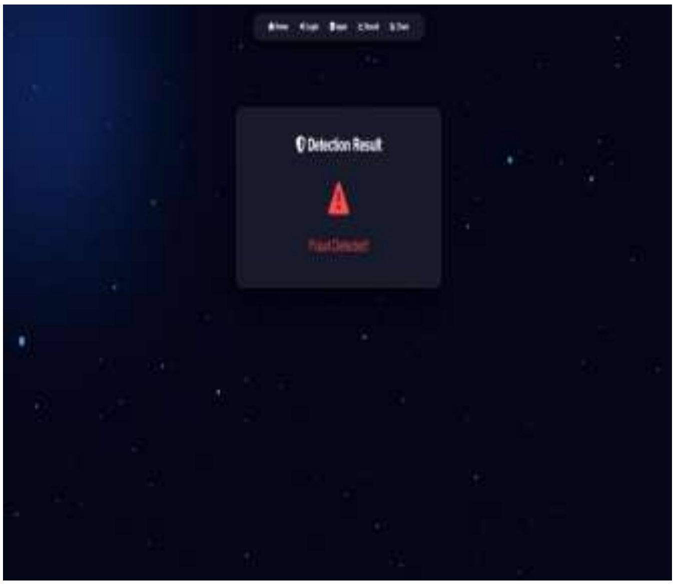


Detection Result

No Fraud Detected

Fig. 2. Fraud Detection Result Interface of the Proposed System

The experimental results show that the proposed model achieves an accuracy of approximately 98%, indicating a strong ability to distinguish between fraudulent and legitimate healthcare claims. The use of the SMOTE-ENN hybrid resampling technique significantly improves the detection of minority class instances by generating synthetic samples and removing noisy data points. This approach helps reduce false negatives and improves the overall reliability of the fraud detection system.Furthermore, the evaluation metrics demonstrate that the model maintains a good balance between precision and recall, which is essential for effective fraud detection. The inclusion of AUPRC as an evaluation metric provides a more accurate representation of model performance in imbalanced datasets. Overall, the results confirm that the proposed approach improves fraud detection performance and provides a reliable solution for identifying fraudulent activities in healthcare systems.



Detection Result

Fraud Detected

Fig. 3. Fraud Detection Alert Result of the Proposed System

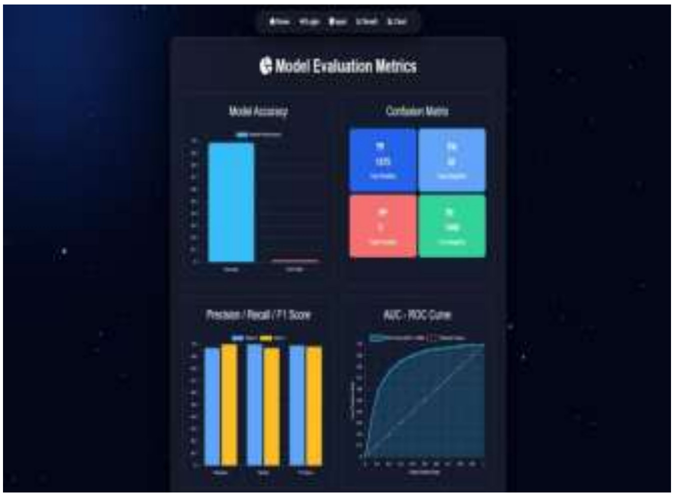


Fig. 4. Model Evaluation Metrics of the Proposed Fraud Detection System

6.CONCLUSION

Healthcare fraud detection has become an important challenge due to the increasing number of fraudulent medical claims and the large volume of healthcare data generated in modern healthcare systems. Identifying fraudulent claims is particularly difficult because healthcare datasets are highly imbalanced, where fraudulent cases represent only a small portion of the total data. Traditional machine learning techniques often struggle to accurately detect fraud under such conditions.In this research, a machine learning-based approach for healthcare fraud detection was proposed using the Medicare Part B dataset. The study addressed the issue of data imbalance by applying the SMOTE-ENN hybrid resampling technique, which combines synthetic sample generation with noise removal to improve dataset quality. A Logistic Regression model was then trained on the balanced dataset to classify healthcare claims as fraudulent or legitimate.The experimental results demonstrated that the proposed model achieved high accuracy and strong classification performance, as evaluated using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. The confusion matrix and ROC curve further confirmed the effectiveness of the model in detecting fraudulent healthcare claims while maintaining a low error rate.Additionally, the proposed system was implemented as a web-based application that allows users to input medical claim details and obtain real-time fraud detection results. This practical implementation improves usability and supports efficient monitoring of healthcare claims.

Overall, the proposed fraud detection framework provides an effective and reliable solution for identifying fraudulent healthcare claims. The results indicate that combining hybrid resampling techniques with machine learning models can significantly improve fraud detection performance in imbalanced healthcare datasets.

7. FUTURE SCOPE

Although the proposed healthcare fraud detection system demonstrates strong performance in identifying fraudulent Medicare claims, several improvements can be explored in future work to enhance the system's effectiveness and scalability.One possible extension of this work is the use of advanced deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which can capture more complex patterns and relationships in large healthcare datasets. These models may further improve fraud detection accuracy and enable the system to detect more sophisticated fraud patterns.Another potential improvement is the application of ensemble learning techniques, such as Random Forest, Gradient Boosting, or XGBoost. Combining multiple machine learning models can increase prediction accuracy and provide more robust fraud detection results compared to a single classification model.Future research can also focus on real-time fraud detection systems by integrating the proposed model with large-scale healthcare databases and streaming data platforms. This would allow healthcare organizations to monitor medical claims continuously and detect fraudulent activities immediately.In addition, incorporating Explainable Artificial Intelligence (XAI) techniques could improve the transparency of fraud detection systems. Explainable models can help healthcare administrators understand the reasoning behind fraud predictions, which is important for decision-making and regulatory compliance.

Finally, the system can be expanded into a cloud-based healthcare fraud monitoring platform that can handle large volumes of medical claim data and support multiple healthcare organizations. Such improvements would make the system more scalable, efficient, and suitable for real-world healthcare fraud detection applications.

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.