International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

AN ENHANCED MACHINE LEARNING FRAMEWORK FOR ACCURATE MEDICARE FRAUD DETECTION

Manal Mohammed Afzal[1], Dr. Muhammed Pasha[2]
[1]PG Scholar, Department of AI & DS, Shadan Women's College of Engineering and Technology, Hyderabad
manal.afzal1804@gmail.com
[2]Professor, Department of CSE, Shadan Women's College of Engineering and Technology, Hyderabad
muhammed.pasha@gmail.com

ABSTRACT
Due to unbalanced datasets, which frequently lead to less-than-ideal model performance, healthcare fraud detection is a crucial endeavour that faces many difficulties. Prior research has mostly used conventional machine learning (ML) methods, which have problems with overfitting from Random Oversampling (ROS), noise from the Synthetic Minority Oversampling Technique (SMOTE), and important information loss from Random Undersampling (RUS). With an emphasis on the Medicare Part B dataset, we present a novel solution to the imbalanced data issue in healthcare fraud detection. The first step in our methodology is to carefully extract the categorical feature "Provider Type,"
SMOTE-ENN generates synthetic data points while eliminating noisy, irrelevant occurrences by combining the Synthetic Minority Oversampling Technique (SMOTE) with Edited Nearest Neighbours (ENN). In addition to balancing the dataset, this combination approach lessens the possible negative consequences of imbalanced data. Using popular assessment metrics including accuracy, F1 score, recall, precision, and the AUC-ROC curve, we assess how well the logistic regression model performs 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 that are unbalanced. The experimental results show that logistic regression outperforms other approaches and validates the effectiveness of our suggested method for identifying healthcare fraud in an unbalanced dataset, achieving an astounding 98% accuracy.
Keywords: Machine learning, Random Forest, social anxiety.

1. INTRODUCTION

In order to protect healthcare systems from financial exploitation, which can result in needless expenses, subpar patient treatment, and a decline in public confidence in medical establishments, healthcare fraud detection is essential. Healthcare costs are greatly impacted by fraudulent claims, such as invoicing for services that were never provided or offering needless services to boost compensation. However, because of the enormous amount and complexity of healthcare data, identifying these fraudulent actions is becoming more difficult. Fraudulent claims are frequently a small percentage of genuine claims in huge datasets with millions of claims, resulting in an unbalanced distribution that makes fraud detection more difficult. Healthcare fraud detection has been tackled using conventional machine learning approaches including Random Forests, Decision Trees, and Logistic Regression.
However, they do have some limits. The imbalanced dataset problem, in which the majority class (non-fraudulent claims) outnumbers the minority class (fraudulent claims), causes many of these models to produce biased predictions. Due to the algorithms' propensity to favour the majority class and produce a high proportion of false negatives—fraudulent cases mistakenly classified as non-fraudulent—this imbalance may lead to models that are unable to identify fraud.

Several strategies, such as Random Oversampling (ROS), Synthetic Minority Oversampling Technique (SMOTE), and Random Under sampling (RUS), have been put forth to address the problems caused by unbalanced data. By replicating already-existing instances, ROS expands the number of minority class examples, but by adding redundant data, it runs the danger of overfitting. SMOTE generates artificial examples of the minority class, Although these techniques aid in balancing the dataset, they have drawbacks of their own that may compromise the fraud detection algorithms' overall effectiveness. By using a hybrid resampling technique known as SMOTE-ENN (Synthetic Minority Oversampling Technique - Edited Nearest Neighbours), we provide a novel approach to healthcare fraud detection in this study that directly solves the problems of dataset imbalance. SMOTE-ENN combines Edited Nearest Neighbours (ENN) to eliminate noisy and superfluous examples from the dataset with the advantages of SMOTE for creating synthetic minority class instances.

2. EXISTING SYSTEM OF THE PROJECT

- In order to identify instances of fraudulent activity in healthcare datasets, like the Medicare Part B dataset, healthcare fraud detection systems mostly rely on conventional machine learning (ML) approaches. These systems frequently employ Support Vector Machines (SVMs), Random Forests,

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Decision Trees, and Logistic Regression. Based on characteristics extracted from healthcare claim data, these models are trained to differentiate between fraudulent (minority class) and non-fraudulent (majority class) actions. The algorithms' ability to correctly identify fraudulent transactions while keeping a low false positive rate for real transactions is a major factor in their performance.

- Class disparity, however, continues to be a serious problem. The majority of healthcare datasets have a situation where the minority class (fraudulent cases) is under-represented since fraudulent instances are uncommon in comparison to the number of legitimate claims. This mismatch can result in biased model training, whereby conventional machine learning algorithms have a tendency to accurately forecast the majority class (non-fraudulent claims) while ineffectively identifying fraudulent instances. This is especially troublesome when it comes to healthcare fraud detection, when the capacity to spot false claims is essential for minimising losses and maintaining system integrity.

EXISTING SYSTEM DISADVANTAGES:

- ROS can lead to overfitting, as the minority class is oversampled by duplicating existing instances.
- It can cause the synthetic instances to overlap with the majority class
- By reducing the number of instances from the majority class, RUS can result in an underrepresented and uninformative majority class, leading to underfitting.

3. RELATED WORK

Researchers and practitioners have paid close attention to healthcare fraud detection because of its social and economic ramifications [11], [12]. The existing research is reviewed in this part and divided into three categories: explainable AI techniques, deep learning (DL) methodologies, and traditional machine learning approaches. We also go over recent developments in hybrid models and how they relate to the suggested framework.

Healthcare fraud detection has made substantial use of traditional machine learning models [13], [14], which use structured claims data to find irregularities and fraudulent activity.

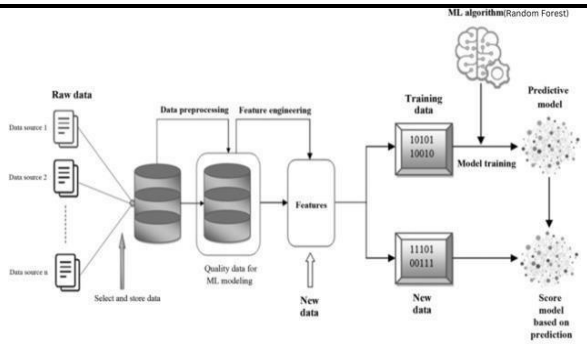
A unique hybrid deep learning system for healthcare fraud detection is proposed in a recent paper by S Agarwal [15]. It integrates CNNs, Transformers, and XGBoost to achieve high accuracy and resilience. Temporal patterns and provider-patient interaction graphs are examples of domain-specific elements included in the framework. The model is evaluated using two datasets: the Healthcare Providers dataset and the Medicare Provider Fraud dataset. It achieves AUC-ROC values greater than 0.96 and F1-scores of 0.92 and 0.94, respectively. The use of different machine learning techniques to identify fraud in health insurance claims is investigated by Kasaraneni et al. [6]. The study divides machine learning techniques into three categories: anomaly detection, unsupervised learning, and supervised learning. For supervised tasks, it emphasises well-known models like LR, random forests (RFs), and gradient boosting; for unsupervised learning, it examines clustering techniques and isolation forests. In addition to discussing real-world case studies showing the practical application of ML-based fraud detection systems in various healthcare contexts, the study highlights the crucial role feature engineering plays in improving model performance.

4 PROPOSED SYSTEM

- We suggest a unique method to enhance model performance, specifically for the Medicare Part B dataset, in response to the 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 Under sampling) have been problems for the current systems. By concentrating on enhancing the dataset's balance and making sure the model is trained on more varied, pertinent data, our suggested method 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. We improve the representativeness of the dataset by generating synthetic cases that provide more variety to the minority class by copying existing "Provider Type" values.
- The classification algorithm used by the system is logistic regression, which works well for identifying fraud in datasets that are unbalanced. We assess the model's performance using a variety of metrics, such as accuracy, F1 score, recall, precision, AUC-ROC, and the Area Under the Precision-Recall Curve (AUPRC), which is important in situations involving imbalanced datasets. In order to guarantee good detection performance, the suggested approach efficiently balances the dataset, improves model training, and offers a thorough evaluation.

5 SYSTEM ARCHITECTURE:

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper



Modules :

1. Data Collection

Gathering the pertinent data—in this case, the Medicare Part B dataset—is the first step. This dataset contains facts about healthcare claims, including patient and provider information, billing amounts, and whether or not the claim is false. The machine learning models are trained and assessed using this data as their basis.

2. Data Preparation

After collecting the data, the next step is to prepare it for analysis. This includes cleaning the data by handling missing values, correcting inconsistencies, and transforming categorical variables into numerical formats. Feature engineering is also an essential part of this step, where key features like "Provider Type" are extracted, and additional synthetic features may be created to improve the model's performance**.**

3 .Data Splitting

Following preparation, the data is divided into training and testing datasets. The model is typically trained on 70–80% of the data, with the remaining 20–30% set aside for testing and validation. This division guarantees that the model's capacity for generalisation can be assessed using data that hasn't been observed.

4. Model Training A machine

learning model—in this case, logistic regression—is trained on the training dataset during this stage. The dataset's class imbalance is addressed by applying the resampling approach SMOTE-ENN, which guarantees that the model learns to accurately classify both the majority (non-fraudulent) and minority (fraudulent) classes. During this stage, hyperparameter adjustment may also be carried out to maximise the model's performance.

5. Model Evaluation

Following training, the model's efficacy is assessed using a variety of performance criteria. The model's capacity to identify fraudulent claims is evaluated using important metrics like accuracy, precision, recall, F1 score, AUC-ROC, and AUPRC, particularly in light of the dataset's extreme imbalance. These metrics aid in determining the model's performance on the majority and minority classes.

6. Model Prediction

Lastly, predictions are made on fresh, untested data using the trained model. This makes it possible to detect possibly fraudulent medical claims in real-time situations. The accuracy and dependability of the model's predictions in identifying fraud are then evaluated by comparing them to the actual results.

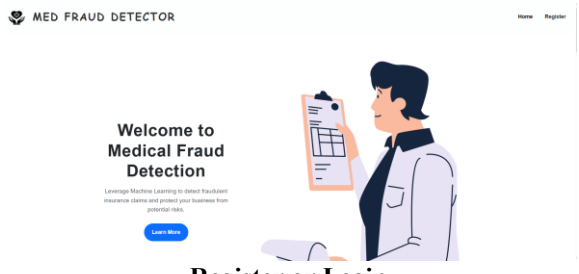
6. TECHNIQUES USED IN PROJECT

- **Random Oversampling (ROS)**
- **SMOTE (Synthetic Minority Oversampling Technique)**
- **Random Under sampling (RUS)**
- A straightforward method to address class imbalance is Random Oversampling (ROS), which increases the number of instances of the minority class. To do this, current instances of the minority class are randomly duplicated and added to the dataset. By doing this, it ensures that the majority and minority classes have an equal number of instances, so balancing the distribution of classes. ROS is frequently applied as a preprocessing step prior to machine learning model training in order to enhance the model's capacity to identify instances of minority classes.

SMOTE is a sophisticated oversampling strategy that creates synthetic instances for the minority class instead of replicating existing ones in order to address class imbalance. SMOTE operates by choosing an instance of a minority class and locating its closest neighbours. The chosen instance and its neighbours are then interpolated to create new synthetic examples. In order to increase the diversity of the minority class and lower the possibility of overfitting, the objective is to produce new data points that are comparable to but distinct from current cases**.**

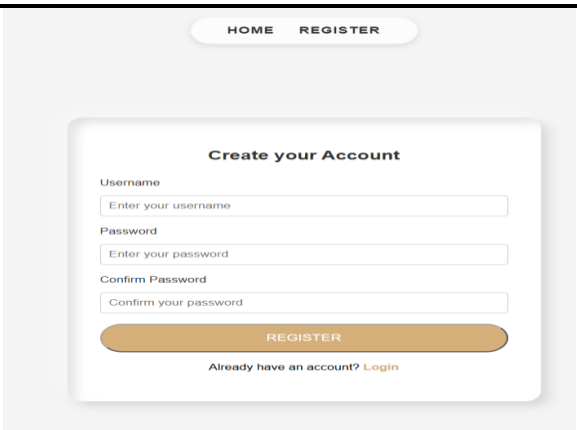
Random Under sampling (RUS) addresses class imbalance by reducing the number of instances from the majority class. This technique randomly selects and removes instances from the majority class until the dataset is balanced between the majority and minority classes. RUS is simple to implement and can be used as a preprocessing step before training the machine learning model to prevent the model from being biased towards the majority class.

7. RESULT

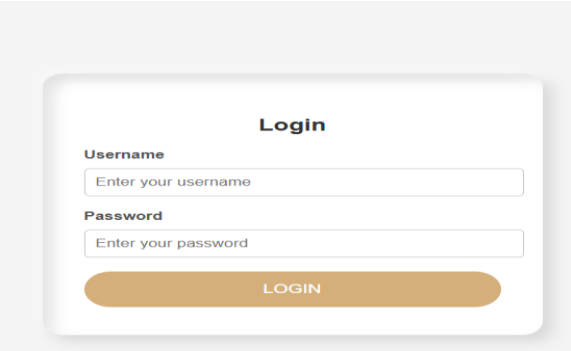


Register or Login

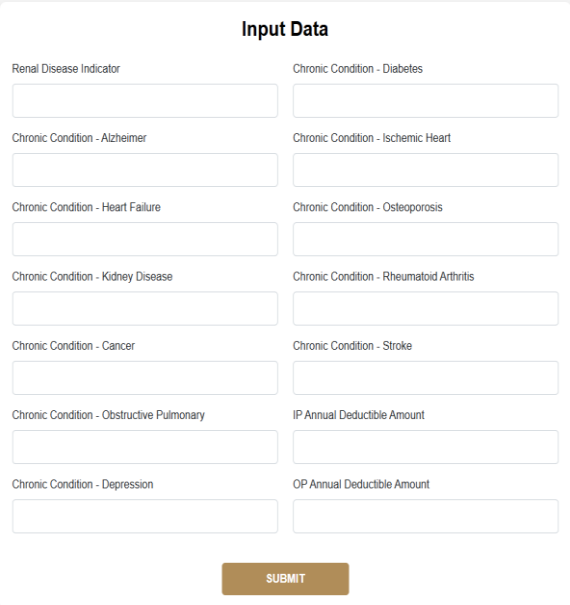
International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper



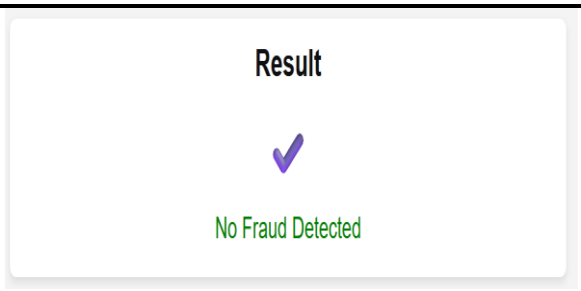
Registration Page



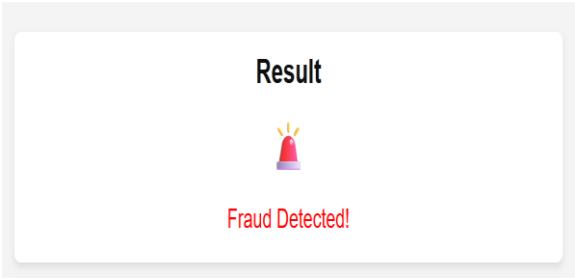
Login Page



Enter Data



Predicted Result-No Fraud Detected



Predicted Result- Fraud Detected

7. CONCLUSION

The urgent necessity to solve the ongoing issues with imbalanced datasets in healthcare fraud detection is highlighted by this study. Effective fraud detection systems are essential because fraudulent actions in healthcare systems result in large financial and operational costs. By combining the creation of synthetic instances with the elimination of noisy and unnecessary data, the suggested framework—which is based on the SMOTE-ENN hybrid resampling method—has effectively addressed the problem of data imbalance. This dual strategy increases the overall performance and dependability of machine learning models in addition to improving dataset quality.

This framework's use of logistic regression demonstrated its capacity to manage binary classification issues while maintaining computing efficiency and interpretability. Additionally, a full evaluation of the model's capabilities was ensured by the inclusion of comprehensive evaluation metrics like Accuracy, F1 Score, Precision, Recall, AUC, and AUPRC. AUPRC, which captures the balance between recall and precision in forecasting the minority class (fraudulent cases), was particularly useful for evaluating performance in imbalanced circumstances. The suggested framework's robustness and efficacy are validated by the achieved accuracy of 98%, indicating its potential as a useful tool for real-world applications in healthcare fraud detection. This study offers a solid basis for further research and innovation in this field in addition to its technical qualities. Researchers and practitioners can develop sophisticated fraud detection systems that are scalable and flexible enough to accommodate a variety of

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

healthcare datasets and changing fraud patterns with the help of the approaches and conclusions offered.

8. FUTURE ENHANCEMENT

Future research can investigate the integration of cutting-edge AI approaches, such as deep learning architectures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to build on the findings of this study. These methods work effectively for identifying intricate, non-linear patterns in medical data that conventional models could overlook. Furthermore, 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 can handle a variety of fraud detection scenarios more skilfully and increase accuracy. Future research must also focus on generalisation and scalability. To guarantee the SMOTE-ENN framework's resilience and flexibility across a range of healthcare fraud scenarios, it should be tested on bigger and more varied datasets. Furthermore, it will be crucial to verify the model's performance in real-time fraud detection situations to make sure the system can react quickly to fraudulent activity, reduce financial losses, and boost operational effectiveness. In dynamic healthcare settings, this real-time capability is especially crucial for the quick detection and prevention of fraudulent claims.

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.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

[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.
[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.