ISBN: 978-628-96613-1-6. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2025.1.1.1758

Detecting Medicare Fraud: A Machine Learning Approach

Rodolfo Andrés Rivas Matta
Department of Electrical Engineering and Computer Science
Florida Atlantic University
Boca Raton, United States
rodolforivasmatta@gmail.com

***Abstract*— The present paper is a survey that started with the analysis of a baseline paper by Bauder and Khoshgoftaar [1], exploring the issue of Medicare fraud detection and the effect of varying class distribution on models' performance. It continued with the analysis of other studies that address the same issues in the process of creating a machine learning approach to detecting Medicare fraud; those include challenges in data collection, handling imbalance, inclusion of proper evaluation metrics, and selection of learners. However, the focus of this study is to understand the issue of Medicare fraud and current solutions to that problem along with their shortcomings, and it includes suggestions for future works.**

The present study includes highlights to keep in mind for future implementations and suggestions unexplored by other works. It revealed that there is an existing method for gathering the data for this problem that is more reliable than the other—namely, using the Centers for Medicare and Medicaid Services (CMS) yearly published data of claims along with the List of Excluded Individuals/Entities (LEIE) published by the Office of Inspector General (OIG). It also pointed out that using proper sampling methods, like RUS, and more adequate class distributions, like 65:35 or 75:25 instead of 50:50, is more beneficial for the task of training machine learning models. Finally, it goes over the existing solutions considering their best-performing models, and it argues that work must be done to map the vast number of solutions and perform a proper comparative analysis of the learners.

Finally, it included suggestions for future works. In addition to hints provided during the analysis of existing methods, it suggests performing certain tasks not covered in reviewed works. For example, it explains the possibility of analyzing ensemble methods, like Random Forests, to interpret their decision-making despite the complexity of their systems; it also strongly exhorts researchers to include companies and experts in the field of detecting fraud to work with more appropriate data and better handler interpretability of results, respectively.

Keywords—Medicare, fraud, datasets, imbalance, models, shortcomings, opportunities

I. Introduction

The present study investigated the issue of Medicare fraud detection, current machine learning solutions, their limitations, and possibilities for future work. It started with a baseline paper by Bauder and Khoshgoftaar [1]. That paper introduced the issue, discussed methods to compile a dataset that would work to solve the problem, and focused on how different levels of class distribution in the dataset affected the machine learning models' performance. Therefore, this study focuses on the issue of fraud detection, methods of compiling a proper dataset for the problem, and different machine learning approaches, along with their strengths and limitations.

Medicare fraud is the main problem this study focuses on. Medicare is an insurance program for people who are 65 years or older, but it makes exceptions for younger people with certain disabilities or conditions [2]. Considering the people who use this program and its function, its importance is clear. Another fact that supports this claim is the increase in spending [3], showing how more people need it every year. Fig. 1 shows, for example, the federal spending in billions of dollars of the Medicare program from 1970 to 2023. Not only is it increasing, but it has gone over a trillion dollars in the last few years.

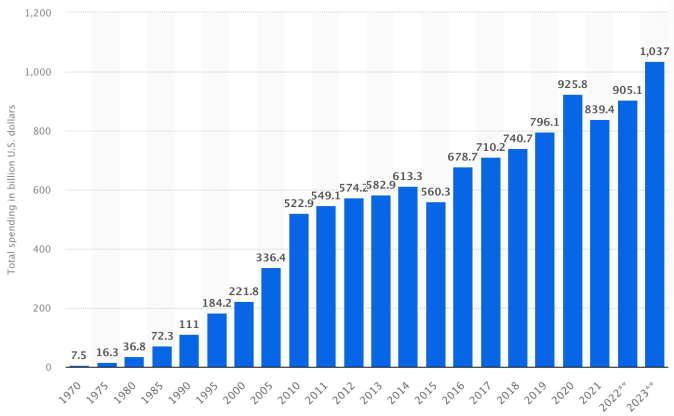

Fig. 1. Chart of Medicare spending from 1970 to 2023 in billions of dollars. [4]

A considerable number of people need this program, but there are other factors that create pressure to ensure Medicare works efficiently. On one side, the population's age distribution is straining the program. People younger than 65 are the major contributors to taxes, which fuels this program. However, the share of elderly population is increasing [5]. In other words, the percentage of people who contribute to this program is decreasing while the percentage of people using it is increasing. Nonetheless, there is another big issue causing pressure on the efficiency of Medicare: fraud.

Fraud has been detected manually, but it is a process where machine learning can help significantly. In 2024, more than 1 billion dollars were lost due to Medicare fraud [6]. Similarly, arguments in [1] point to the fact that there is more money being lost, which we do not know about; the FBI estimates the real number to be much higher. That is another very negative aspect that is causing stress in the program's quality, which affects its reputation and threatens its future. However, fraudulent claims are a problem that could be detected using machine learning classifiers.

II. Current Solutions

Interestingly, there is vast research in Medicare fraud detection using machine learning. Different solutions use different frameworks, data collection approaches, models, hyperparameters configuration, evaluation metrics, and deal with challenges in different ways. Therefore, this study considered the pipeline and main contributions by [1] when considering other papers and their approach to the same stages and problems.

A. Data Collection and Preprocessing

The first problem any machine learning project faces is usually the data. If there was a clear correlation between the input and the expected output, then machine learning would not be necessary. Such correlation does not exist. Similarly, if sample data was easily available, clean, and with no unexpected issues, it would be a matter of applying the data to multiple learners and using the best one for the task. However, no such dataset exists for this problem. There is no public record of Medicare claims with labeled fraudulent entries, so researchers used two main approaches to gathering the sample data in the papers reviewed.

One approach was to use a dataset by Gupta from Kaggle [7]. The source of this dataset is unclear. Nonetheless, the page recognizes fraudulent claims as being a major issue affecting Medicare, and they argue it is an issue all insurance companies suffer as well. Therefore, the goal for those using this dataset, according to the page, must be to develop a machine-learning approach to detect fraudulent claims. That fits perfectly with the goal of the researchers using this dataset; two examples of studies using this approach are [8] and [9].

Preprocessing that Kaggle dataset is simple, and it does not present many issues. The data is organized in four separate tables.

1. Inpatient claims.
2. Outpatient claims.
3. Beneficiaries.
4. Providers' ID and fraudulent flag.

Researchers using this dataset must aggregate all four tables to create samples where each claim has a flag of being fraudulent or not. In fact, in [8], the process is explained and consists of three main steps. Fig. 2 shows this process. First, the outpatient and inpatient claims are merged, which is not complex since they have almost the same columns. Then, the output of the merged sets is aggregated with the beneficiaries table using the beneficiary's ID. Finally, the providers' table only has the providers' ID and whether each one is flagged as fraudulent or not; those using this approach must flag each claim as fraudulent using the provider's ID.

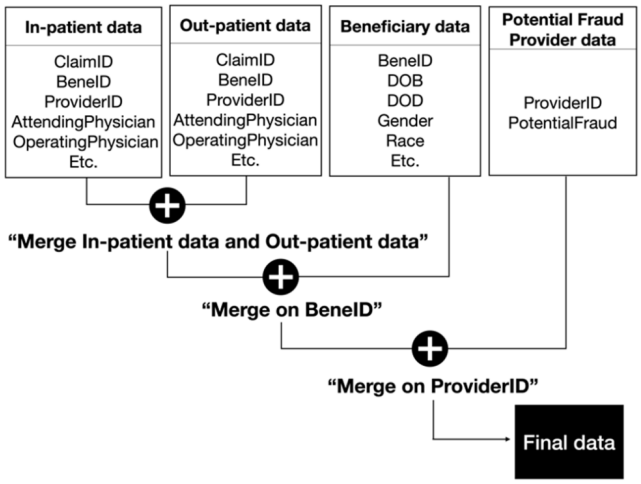

Figure 2. Dataset merging process, extracted from [8].

The second approach consists of combining real public data from different sources to create a dataset of claims with a label that flags fraudulent claims. Examples of papers using this approach are [1], [10], [11], [12], and [13]. Yet, this approach has more difficult challenges when it comes to preprocessing, but studies using this approach have been innovative when coming up with solutions.

The second method extracts data from three different sources first. Fortunately, the Centers for Medicare and Medicaid Services (CMS) publish data annually on the claims for the different parts of the Medicare program—Medicare has different parts where each covers different types of claims, like inpatient care, services from doctors or providers, drug coverage, and more. One such place where that data is available is [14]. However, that data is not usable for fraud detection, as it has no label that identifies fraudulent claims.

Researchers must properly flag fraudulent claims and handle other underlying issues. Similarly to the Kaggle dataset, claims could be flagged using an external set of providers flagged as fraudulent. The Office of Inspector General (OIG) has a public dataset of providers that have been excluded from the Medicare program along with the reasons behind the exclusion; that is, the List of Excluded Individuals/Entities (LEIE). Of course, there are many challenges in using such data to flag fraudulent claims as well.

The process of flagging claims requires attention to detail and an adequate process to avoid introducing errors. In addition to listing the excluded providers, the LEIE dataset also has other relevant information, such as the period of exclusion and reasons. In [1], the process is clearly explained. The claims only have the year in which they were made, while the LEIE's exclusions include the specific month and day in addition to waiver if applicable—a date in which the providers have been reinstated if it is earlier than the original end period of the exclusion. Those using this approach must take all those variables into account and flag claims accordingly.

Additionally, studies using the second approach also considered other factors that could improve the learners' ability to learn the underlying patterns. For example, the CMS datasets use codes for the procedure listed in the claims—a number. However, Johnson and Khoshgoftaar [15] found a significant improvement in using the Healthcare Common Procedure Coding System (HCPCS) to get the description of those procedures and encode them using methods that capture the semantic meaning, like the Word2Vec algorithm.

Both approaches to get the data require work, but the second approach using the CMS dataset seems to require significantly more time preprocessing the data with the benefit of a much more realistic training set. The final dataset produced by either method still has underlying challenges that researchers must consider.

B. Imbalance

Regardless of the approach used to gather the data, whether it was using the Kaggle dataset or the combination of many real sources, the data for this problem always revealed severe imbalance. In fact, one of the studies using the second method stated that their dataset had "only 0.04% of instances being labeled as fraud" [1]. In fact, the effect of class imbalance on learners' performance was a major focus in [1].

High-class imbalance causes issues in learners with major repercussions that cannot be easily identified without proper metrics. It is common to evaluate machine learning models using accuracy—the correct number of samples classified—but that approach could lead to the conclusion that a model is highly efficient when it really is just unusable. For example, using a testing set with a class distribution like the dataset in [1], with a minority class (fraudulent claims) of 0.04%, a model that does not perform any smart decision-making and flags all entries as non-fraudulent could achieve an accuracy of 99.96%. Such a model, of course, is useless in the real world.

However, using the proper evaluation metrics can only catch the issue. Bauder and Khoshgoftaar's work points out that sampling the original dataset to reach a different class distribution is a major contributor to solving the problem [1]. Otherwise, it is difficult for models to learn the underlying patterns to classify fraudulent claims. Interestingly, there are many approaches to sampling the dataset. Most of them were covered by [1], as their focus was to understand the effect of varying class imbalance on learners' performance.

- Random Oversampling (ROS): This approach consists in randomly duplicating instances of the minority class to increase its representation in the dataset. However, due to duplication, it can lead the learners to suffer from overfitting.
- Random Undersampling (RUS): Similarly to ROS, it uses random selection, but it removes instances of the majority class instead. This method seems to be the most effective for big data, as highlighted in [16]. The drawbacks of this method are not significant for the problem of Medicare fraud detection given the available datasets.
- Hybrid (ROS-RUS): Combining both approaches of ROS and RUS to randomly remove the majority class and randomly duplicate the minority class is also possible. However, the given problem seems to benefit more from RUS alone, as ROS requires a significant number of duplications to make an impact; unfortunately, such duplication increases the effects of overfitting. One study that particularly used this approach was [11].
- Synthetic Minority Oversampling Technique (SMOTE): Another approach to increasing the number of samples in the minority class is to generate synthetic samples of the minority class by interpolating existing ones. In other words, it looks at samples of the minority class and creates a sample that would exist in between the two instances, which at the same time are the nearest neighbor of each other. The problem with this approach is that it can introduce noise.
- Wilson's Editing (WE): Finally, WE is a method that also uses the nearest neighbors' approach in its decision making. It works by identifying noisy samples, using a nearest-neighbor classifier, and removes them from the dataset. One sample study that used it was [16]. Unfortunately, this method does not seem to be as effective for the problem of Medicare fraud detection.

Bauder and Khoshgoftaar's work [1] pointed out that RUS is the most effective method for the CMS dataset. Additionally, they looked at the varying class distributions—namely, 99.9:0.1, 99:1, 95:5, 90:10, 75:25, 65:35, and 50:50. Their results show that RUS is effective, possibly due to the large number of samples in the original dataset. Even though RUS could remove a significant number of samples, there is enough data for the model to learn the underlying patterns. Nonetheless, the more information there is available to learners, the better; their conclusions also pointed out that a class ratio of 50:50 is not optimal, the best being 65:35 and 75:25. That is possible because the last two ratios have more information that is available to learners during their training.

C. Existing Machine Learning Approaches

Just like there were two approaches to gathering the data for this problem, most papers analyzed in this study demonstrated two paths to solving the issues. Most works use traditional machine learning models, like decision trees, logistic regression, and more; meanwhile, the rest seem to have taken an interest in neural networks. Just like data collection, each machine learning approach had its challenges.

1) Traditional Machine Learning

Both [1] and [13] seem to list the most traditional machine learning methods used for this problem. Those are:

- Naïve Bayes (NB)
- Logistic Regression (LR)
- K-Nearest Neighbor (KNN)
- Local Outlier Factor (LOF) (specifically in [13])

Identify applicable funding agency here. If none, delete this text box.

- Support Vector Machine (SVM)
- Decision Tree (C4.5 or J48)
- Random Forest (RF) (best model as listed in [1])

There are some major aspects of interest among these learners. For example, learners like LR and J48 had the benefit of being easily interpretable. That is beneficial for companies implementing a fraud detection system since they could understand how these machine learning models are making their decisions and point them to possible solutions that go beyond detecting fraud. Similarly, most studies show ensemble methods being more effective. For example, RF was the best model according to the comparison made in [1]; unfortunately, ensemble methods are complicated to interpret. The RF in [1] consisted of 100 trees, and it would not be enough to look at the nodes in those trees to understand why a sample received a specific classification.

Another interesting research was done in [13], where they looked at unsupervised approaches. They used KNN and LOF, for example, to classify models without using the fraudulent label in the training stage. They explained that they had created a cluster of all instances and selected a distance threshold so that instances either fell inside the cluster or outside. The threshold was chosen in such a way that, for example, only 1% of instances would exist outside the cluster. That minority outside the cluster represents outliers, which are considered fraud. Interestingly, this approach performed reasonably well. However, binary-class classifiers (BCC), those that have both labels, tend to perform much better. Again, that may be due to the extra information available to models that allows them to better learn the underlying patterns.

2) Neural Networks

The second major approach to solving the issue of fraud detection is using neural networks. Works like [8], [9], and [11] follow this path. However, it has not been as thoroughly studied as the traditional machine learning methods. The studies analyzed seem to follow two approaches when using neural networks: one is using a relatively simple multilayer architecture [11], and the other approach is to use graph neural networks (GNN) as in [8] and [9]. Works on both approaches provided significant contributions.

The study using a multilayer architecture [11] experimented with models having between 2 and 4 hidden layers, but most of their work was in evaluating the effect of data manipulation in such models. Their neural architecture, for example, also contained multiple normalization and dropout layers in addition to exploring multiple class distributions, as in [1]. Their results highlighted the importance of data preprocessing even when using neural networks.

On the other hand, [8] and [9] organized the data in a novel structure using graphs where nodes represent beneficiaries and providers while links represent the claims or services provided. Their argument behind this approach was that fraud usually involves more than one party. Their work—in addition to demonstrating that neural networks can be as effective in this task—lays the ground for others to work on the interpretability of neural networks in this task. Although it was not the focus of their work, preparing the data as a relationship between beneficiaries and providers means that neural networks learn to classify claims starting from the relationship that beneficiaries and patients have.

D. Evaluation Metrics and Analysis

Regardless of the learners used or methods involved in gathering the data, researchers must select proper testing methods, performance metrics, and analysis tools. There are two validation methods in the studies reviewed; it is a trade-off between computational resources and correctness. On the other hand, proper performance metrics are essential to reach valid conclusions; as aforementioned, accuracy alone gives a bad metric when evaluating models. Finally, there are also analysis tools, like Analysis of Variance (ANOVA) and Tukey's HSD test, that allow researchers to objectively measure the significance of their results.

There are two major approaches in the papers analyzed in this study for the validation method. The first and most correct method is using stratified k-fold cross-validation. It consists of separating the dataset in k-folds and using one fold for testing and the rest for training, repeating the process k times for each of the folds, and aggregating the results to get a final report of the full dataset. This method is great because it uses the whole dataset for training and testing, avoiding issues like overfitting. Studies like [1]. [10], [12], [13], [15], and [16] use this approach. However, it is the most computationally expensive. The other method consists of separating the dataset into a training and testing set only once. The second method is mostly used for computationally intensive models, e.g., neural networks.

However, the most important aspect when measuring the effectiveness of models lies in the performance metrics used. All studies analyzed seem to be aware of the issue of using accuracy as a reliable metric. Nonetheless, one study tried a similar metric called balanced accuracy, but their results demonstrated that it also lacks reliability. Therefore, the most reliable and most used methods for this problem are:

- Area Under the Receiver Operating Characteristic (ROC) Curve (AUC): The AUC is a commonly used metric for classification tasks because it takes into consideration both positive and negative classifications, generating a curve where the area under the curve is a single reliable number to summarize the efficiency of models.
- Area Under the Precision-Recall Curve (AUPRC): Like AUC, AUPRC also reduces the results to a single number extracted from a curve. However, the curve is the result of looking the precision and recall values of a model.
- True Positive Rate (TPR) & True Negative Rate (TNR): Instead of summarizing the results, one could also look at the TPR and TNR values of a test. The goal is to achieve a balance between the two, while both numbers being as high as possible.
- False Positive Rate (FPR) & False Negative Rate (FNR): The FPR and FNR complement the TPR and TNR. However, they could be seen as type I and type

II error rates—where type II error rate is the metric of most interest. In the case of Medicare fraud, type II errors are fraudulent claims classified as non-fraudulent. The goal is to minimize type II errors as much as possible keeping a balance with the type I errors, just as in the previous metrics.

E. Overview of Major Results

When analyzing the results of each of the studies, an issue came to light. Even when using the same metrics and models, there are many variables that make the comparison inappropriate. One such factor was the highly different hyperparameter values used for the same models. Nonetheless, the most significant findings are discussed below, keeping the limitations just mentioned and looking always at the AUC of their best models.

A. In [1], a Random Forest with 100 trees reached an AUC of 87.30%, the highest value across the studies analyzed. The same study analyzed various traditional machine learning models. However, no neural networks were included.

B. One of the first neural networks analyzed appeared in [8], which used a graph neural network (GNN), more specifically GraphSAGE and the Kaggle dataset. They were the earliest known study to use GNNs, and they achieved a remarkable performance of 71% in AUC.

C. There was another study following the previous one that applied some tuning to the Graph Attention Network (GAN) [9]. They performed a more extensive evaluation across models, and they reached 74% of AUC. Interestingly, they compared the results of the neural network model with those of classical methods. Two of the best classical methods were RF and XGBoost, which scored with AUCs of 66% and 68%, respectively. Nonetheless, they provided only limited information on the hyperparameters used for those models—for example, they did not explain how many trees made up the RF.

D. The third study using the neural network consisted of the multilayer architecture mentioned before [11]. Interestingly, they achieved a performance of 85.09% of AUC. However, they attribute their performance to using ROS & RUS for preprocessing the data rather than to their various modifications to the models' architecture.

III. Shortcomings of Existing Solutions

Some papers do have shortcomings that others point out. For example, [1] discusses the importance of understanding the process of collecting the data, something that [8] and [9] do not consider when using the Kaggle dataset. This section focuses instead on shortcomings not found in the works analyzed.

A. Dataset Limitations

Both approaches, using the Kaggle dataset or the CMS dataset with the LEIE and HCPCS codes, have their limitations.

The Kaggle dataset seems to have the most obvious limitations that make it unusable for real-case scenarios. First, in the dataset's website [7], there is no clear information on the source of the data. It could be artificial or adapted from a source, but it does not match any existing standard. For example, the CMS datasets usually include the providers' ID as a unique 10-digit number, while the Kaggle dataset has each provider start with the "PRV" prefix followed by a 5-digit number.

In fact, Kaggle itself points to the low dataset quality, as visible in Fig. 3.

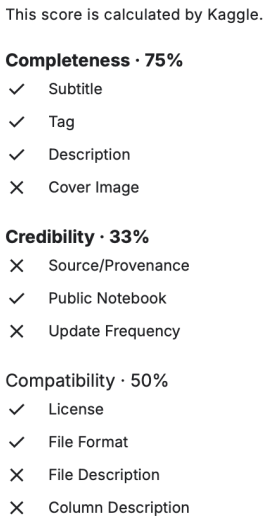

Figure 3. Screenshot taken from the Kaggle website in [7].

However, the Kaggle dataset has realistic challenges. For example, the CMS datasets do not contain a fraud label for their claims. The second approach to collecting the data was to use the LEIE's dataset to flag claims as fraudulent. Researchers using the Kaggle dataset face a similar challenge, but their results may not be as relevant if the data is purely artificial. Even if a specific method or tunning could achieve astonishing performance with this data, it could perform less than optimal in a real dataset.

The second approach to gathering the data, using the CMS datasets aggregated with the LEIE's dataset and the HCPCS code set, faces other issues that also put into question the resulting metrics of the models using it. One of the most obvious issues is the noise introduced by using the LEIE list of providers—something that is also present in the Kaggle dataset. The LEIE set has only the period of exclusion of providers. It does not necessarily mean that providers' claims during that period are all fraudulent; similarly, claims outside of that period may be fraudulent—for example, consider the frauds that led to the exclusion.

Other less obvious shortcomings are the missing fraudulent labels and the bias introduced by using historical data. First, as [1] points out, there are many fraudulent activities that are not caught by authorities; that means that many non-fraudulent claims in this approach may be fraudulent. That highlights the importance of looking for a method of crafting the data that could improve that issue. Similarly, a common problem with using historical data is that models learn patterns that will not always be true in the future. Of course, using the CMS approach fights back this issue by updating the datasets with the latest

publications by the CMS—they publish their claims yearly. However, there is an unexplored area of analyzing how claims change over time—mostly if claims are available for many years.

B. Interpretation of Models

In [1], the interpretation of models is discussed as a relevant factor. However, they also explained that models like decision trees are interpretable, while RF and other ensemble methods are too complex to interpret. Nonetheless, programs could be written to get relevant data from those ensemble methods—for example, a program could navigate through the 100 trees of a RF looking for the most common nodes and their distance to the root, taking also into account duplicates in each tree.

Studies like [8] and [9] also proved that it is possible to analyze the data fed to a neural network and its format to understand how the NNs make their predictions. Nonetheless, interpreting neural networks is considerably more complicated.

On another hand, a strong suggestion for those working in the field of Medicare fraud detection with machine learning is to involve experts in the field. Including those that are currently part of the system that detects fraud, along with companies that have more realistic data, can significantly support research groups to come up with more realistic and better-performing systems.

C. Other Opportunities for Future Work

All the issues discussed so far lay the ground for future work. For example, future works should consider involving experts or companies in the process. However, another evident opportunity for future work is to create a comprehensive map of existing solutions. This study analyzed many works related to Medicare fraud detection, but there are many more that also fit within this category and were not analyzed. A comprehensive survey that analyzes the various methods in a proper setup would be highly beneficial to plan the direction of future work that could bring more fruitful work. Such a study should not only include the various machine learning models, but it should ensure that all have adequate hyperparameter tunning. For example, [9] did compare a GNN with more traditional approaches, but it focused on tuning the GNN and most probably used a limited setup for the traditional methods—they do not provide details about those other models.

IV. Conclusions

This study started with [1] as the baseline paper and analyzed multiple works that aim to solve the same problem with similar challenges like data gathering, dealing with imbalance, and using appropriate performance metrics. The extensive search revealed interesting issues along with remarkable findings in the process of creating a proper model that can detect fraudulent claims.

The study analyzed two approaches mostly used to gather the data, the best methods to deal with imbalance, two main branches of tackling the program, and possibilities for future work. The study points out the benefits of using CMS with the LEIE dataset and HCPCS codes, but it draws attention to the need to create more reliable datasets—paying more attention to correctness to diminish noise. The study also revealed that using RUS to resample the datasets to a ratio of 65:35 or 75:25 is more beneficial for this task. Finally, it explored the different machine learning models used for this problem, but it also pointed to the need to have a clearer view of existing methods to prepare for future works.

Most importantly, this study provided relevant suggestions for future work. In addition to highlighting the importance of well-known practices, like using appropriate metrics for measuring classifiers' performance, there are suggestions like using interpreting models to gather more insights into the issue or involving experts and companies in the process of creating a machine learning fraud detection system.

References

[1] 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 Information Science and Systems*, vol. 6, no. 1, Sep. 2018, doi: https://doi.org/10.1007/s13755-018-0051-3.

[2] Medicare.gov, "Parts of Medicare." www.medicare.gov. Accessed: Nov. 23, 2024. [Online.] Available: https://www.medicare.gov/basics/get-started-with-medicare/medicare-basics/parts-of-medicare

[3] Centers for Medicare & Medicaid Services, "NHE Fact Sheet | CMS," www.cms.gov. Accessed: Nov. 23, 2024. [Online.] Available: https://www.cms.gov/data-research/statistics-trends-and-reports/national-health-expenditure-data/nhe-fact-sheet

[4] "Total Medicare spending from 1970 to 2023 (in billion U.S. dollars)." Statista. Accessed: Nov. 21, 2024. [Online.] Available: https://www.statista.com/statistics/248073/distribution-of-medicare-spending-by-service-type/

[5] Z. Caplan, "U.S. older population grew from 2010 to 2020 at fastest rate since 1880 to 1890," United States Census Bureau, May 25, 2023. Accessed: Nov. 21, 2025. [Online.] Available: https://www.census.gov/library/stories/2023/05/2020-census-united-states-older-population-grew.html

[6] "Office of Public Affairs | National Health Care Fraud Enforcement Action Results in 193 Defendants Charged and Over $2.75 Billion in False Claims | United States Department of Justice," www.justice.gov. Accessed: Nov. 21, 2024. [Online.] Available: https://www.justice.gov/opa/pr/national-health-care-fraud-enforcement-action-results-193-defendants-charged-and-over-275-0

[7] R. A. Gupta. *Kaggle Healthcare Provider fraud detection Datasets*. Available: https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis

[8] Y. Yoo, D. Shin, D. Han, Sunghyon Kyeong, and J. Shin, "Medicare fraud detection using graph neural networks," *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, Jul. 2022, doi: https://doi.org/10.1109/icecet55527.2022.9872963.

[9] S. Mardani and H. Moradi, "Using Graph Attention Networks in Healthcare Provider Fraud Detection," in *IEEE Access*, vol. 12, pp. 132786-132800, 2024, doi: 10.1109/ACCESS.2024.3425892.

[10] J. L. Leevy, Z. Salekshahrezaee and T. M. Khoshgoftaar, "A Review of Unsupervised Anomaly Detection Techniques for Health Insurance Fraud," *2024 IEEE 10th International Conference on Big Data Computing Service and Machine Learning Applications (BigDataService)*, Shanghai, China, 2024, pp. 141-149, doi: 10.1109/BigDataService62917.2024.00028.

[11] J. M. Johnson and T. M. Khoshgoftaar, "Medicare fraud detection using neural networks," *Journal of Big Data*, vol. 6, no. 1, Jul. 2019, doi: https://doi.org/10.1186/s40537-019-0225-0.

[12] R. Bauder and T. Khoshgoftaar, "A Survey of Medicare Data Processing and Integration for Fraud Detection," *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, Salt Lake City, UT, USA, 2018, pp. 9-14, doi: 10.1109/IRI.2018.00010.

[13] R. A. Bauder and T. M. Khoshgoftaar, "Medicare Fraud Detection Using Machine Learning Methods," *2017 16th IEEE International Conference*

on Machine Learning and Applications (ICMLA), Cancun, Mexico, 2017, pp. 858-865, doi: 10.1109/ICMLA.2017.00-48.

[14] "Open Payments Dataset Downloads | CMS," *www.cms.gov*. https://www.cms.gov/priorities/key-initiatives/open-payments/data/dataset-downloads

[15] J. M. Johnson and T. M. Khoshgoftaar, "Hcpcs2Vec: Healthcare Procedure Embeddings for Medicare Fraud Prediction," *2020 IEEE 6th International Conference on Collaboration and Internet Computing (CIC)*, Atlanta, GA, USA, 2020, pp. 145-152, doi: 10.1109/CIC50333.2020.00026.

[16] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano, "Evaluating the Impact of Data Quality on Sampling," Journal of Information & Knowledge Management, vol. 10, no. 03, pp. 225–245, Sep. 2011, doi: https://doi.org/10.1142/s021964921100295x.