REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 1S (2024)
http://www.veterinaria.org
Article Received: 05/08/2024 Revised: 25/08/2024 Accepted: 05/09/2024

"Machine Learning Approaches for Detecting Fraudulent Claims in Veterinary Healthcare"

Dr. Kondragunta Rama Krishnaiah[1*], Dr. Harish H[2]

[1*,2]R K College of Engineering, Vijayawada 521456, Andhra Pradesh, India. kondraguntark@gmail.com

Abstract
This study investigates the application of machine learning methods to identify fraudulent claims in animal healthcare. The researchers utilized publicly available veterinary claims data and regulatory exclusion databases to label fraudulent claims. Three supervised machine learning models were employed: C4.5 Decision Tree, Logistic Regression, and Support Vector Machine. Their performance was evaluated using metrics such as Area Under the ROC Curve, False Positive Rate, False Negative Rate, and Precision-Recall.

The findings demonstrate that the C4.5 Decision Tree outperformed the other two models in terms of AUC, recall, and FNR, making it the most effective approach for detecting fraudulent claims in veterinary healthcare. The C4.5 model achieved an AUC of 0.883 at an 80:20 class distribution and exhibited the lowest FNR, successfully identifying fraudulent claims without missing significant instances of fraud. Although Logistic Regression showed high precision, it had a higher FNR, indicating a trade-off between precision and recall. SVM exhibited lower overall performance compared to the other models, particularly in AUC and FNR.

The results highlight the potential of machine learning to enhance fraud detection systems in animal healthcare, providing a robust approach to identifying fraudulent claims that may otherwise be overlooked. Future research could focus on exploring additional data sources, feature engineering, and alternative sampling techniques like Synthetic Minority Over-sampling Technique to further improve the detection process. This study contributes to the growing body of work aimed at leveraging machine learning to detect fraud and ensure the proper allocation of resources in animal healthcare.

Keywords: Veterinary Healthcare, Fraud Detection, Machine Learning, C4.5 Decision Tree, Supervised Learning, AUC.

1. Introduction

The effectiveness of animal healthcare systems depends significantly on competent providers and a stable financial infrastructure. Unfortunately, both of these pillars can be severely impacted by fraudulent activities, which lead to substantial financial losses (Kondragunta, R. K., 2021 & DeHaven, 2014). The veterinary healthcare sector, much like human healthcare, faces the challenge of fraudulent claims, which can divert resources away from the care animals need (Nabrawi & Alanazi, 2023). As with other healthcare systems, fraud in animal healthcare programs can undermine public trust, reduce the quality of care, and increase the cost of treatment, ultimately harming animal well-being (Zhang et al., 2020)

The veterinary field has seen significant growth, particularly in areas such as animal surgery, nutrition, and preventive care, which are essential to maintaining animal health. As of recent reports, the global spending on veterinary services has surged, reaching billions annually (Wünderlich et al., 2021). However, fraud and abuse in these sectors are believed to account for a notable portion of these expenditures (DeHaven, 2014). For example, fraudulent claims related to unnecessary treatments, falsified diagnoses, or overcharged services represent a serious concern in the global veterinary healthcare system. It is estimated that healthcare fraud accounts for a significant percentage of total healthcare spending worldwide, with some reports suggesting it may contribute to up to 10% of total healthcare costs in various regions.

In response to the rising threat of fraud in veterinary practices, many countries have implemented programs aimed at curbing fraudulent activities (Offen, 1999). However, there is still a critical need for more effective measures to detect and mitigate such frauds. Specifically, the veterinary sector would benefit from advanced machine learning approaches that can identify suspicious patterns in claims and billing data (Lu et al., 2023). These methods, which have proven effective in human healthcare fraud detection, could similarly be applied to animal healthcare systems to protect resources and ensure that services are provided to animals based on their genuine medical needs (Nabrawi & Alanazi, 2023).

This work looks at using machine learning to spot fraud in animal healthcare. We used public data on claims, along with info on providers known to be fraudulent. Unlike approaches that focus solely on specific medical specialties, this model predicts fraud across various types of veterinary care, whether for companion animals, livestock, or other animals. Specifically, we use data from veterinary provider claims, which include information on services, procedures, and treatments provided to animals. To identify fraudulent claims, we merge this data with lists of providers who have been

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 1S (2024)
http://www.veterinaria.org
Article Received: 05/08/2024 Revised: 25/08/2024 Accepted: 05/09/2024

flagged for fraud by governmental or regulatory bodies. The challenge lies in the fact that veterinary claims data often lacks direct fraud labels, which is why we utilize the existing provider exclusion databases to infer fraudulent behavior.
We focus on developing a unified fraud detection model that can generalize across different types of veterinary practices, ensuring that the system can detect fraud regardless of specialty. Our approach leverages the power of machine learning to overcome the difficulties posed by imbalanced data sets, where fraudulent claims are rare compared to legitimate ones. We adopt data sampling techniques, such as random undersampling, to address the skewed distribution of fraud cases and improve the performance of our model in detecting these rare but critical instances.

2. Literature Review

The issue of fraud in healthcare, including the veterinary sector, has garnered significant attention due to its detrimental effects on both the quality of care and financial sustainability. As in human healthcare, fraudulent activities in veterinary care can range from overbilling and unnecessary treatments to falsified diagnoses, which exploit vulnerable animals and their owners. Lots of research has looked at using machine learning and data analysis to catch fraud in healthcare. These same techniques can be really useful for tackling fraud in animal healthcare too, even though the details might be a bit different.
In human healthcare, several approaches have been employed to detect fraudulent claims. One of the early techniques involved anomaly detection methods, which identify outliers in large datasets that deviate from normal patterns of care. For example, Sadiq et al. (2017) utilized the Patient Rule Induction Method (PRIM) to identify anomalies in Medicare claims data. Their approach, based on unsupervised learning, focused on detecting outliers in large-scale claims data without prior knowledge of fraudulent activities. While this method has been effective in detecting unusual behavior, its applicability in the veterinary sector could be limited by the unique nature of animal healthcare services and the lack of established fraud patterns.
More recent works have incorporated supervised machine learning techniques, which require labeled datasets of known fraud cases to train models. Bauder et al. (2016) demonstrated the use of supervised learning techniques, such as logistic regression and decision trees, to detect fraud in Medicare claims using publicly available data. Their work highlighted the challenges of dealing with imbalanced datasets, where fraudulent claims are rare compared to legitimate claims. To mitigate this issue, they used random undersampling (RUS) and synthetic data generation techniques to balance the class distribution. This approach is directly applicable to the veterinary sector, where fraud labels are also scarce, and imbalanced datasets are a common challenge.
Another study by Herland et al. (2017) focused on improving the accuracy of fraud detection by incorporating additional features into their models, such as provider specialty predictions. Their work emphasized the importance of feature engineering to enhance fraud detection performance, an area that holds potential for animal healthcare fraud detection as well. The inclusion of features such as the type of services provided (e.g., preventive care vs. emergency care) or the species of animal treated could be valuable in identifying fraudulent patterns in veterinary practices.
Additionally, Branting et al. (2016) explored graph-based models to detect fraud by analyzing the relationships between providers, prescriptions, and services. Their approach used network analysis to identify suspicious connections that could indicate fraudulent activities, such as collusion between multiple providers. This method, while effective in human healthcare, could be adapted for veterinary fraud detection by mapping out the relationships between different veterinary practices, service providers, and the animals they treat.
In the specific context of veterinary care, there has been limited research dedicated to detecting fraud using machine learning techniques. However, there are a few studies that have explored fraud detection in related fields. For example, Pande and Maas (2013) examined fraud in veterinary claims related to pharmaceutical purchases, identifying key patterns that could indicate fraud. Their research underscored the potential of using machine learning techniques to detect fraud in animal healthcare, although it focused primarily on pharmaceutical fraud rather than broader healthcare fraud.
Joudaki et al. (2015) provided a broader overview of fraud detection in healthcare, reviewing various data mining techniques, including supervised and unsupervised methods. Their review concluded that machine learning, particularly in the context of large healthcare datasets, offers promising tools for fraud detection, especially when combined with domain-specific knowledge. This is an essential aspect for applying fraud detection techniques in the veterinary sector, where domain expertise is crucial for interpreting and validating the data.
In summary, while there has been considerable progress in fraud detection for human healthcare, applying similar methods to the veterinary sector presents unique challenges and opportunities. The veterinary field, with its specific regulatory frameworks and types of services, requires tailored fraud detection models. The works reviewed in this section illustrate the potential of machine learning to improve fraud detection in healthcare, and by extension, in animal healthcare. Building on these existing methods and adapting them to the context of animal healthcare will be critical to developing effective fraud detection systems that can reduce fraudulent activities and improve care for animals.

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 1S (2024)
http://www.veterinaria.org
Article Received: 05/08/2024 Revised: 25/08/2024 Accepted: 05/09/2024

3. Methodology

In this section, we outline the methodology used to detect fraud within the animal healthcare sector using machine learning techniques. Our approach involves the collection and processing of veterinary claims data, followed by the integration of fraud labels sourced from regulatory exclusion databases. We then apply supervised machine learning algorithms to build models that can classify fraudulent claims, considering the challenges of class imbalance and large datasets. This section details the data sources, learners, performance metrics, and techniques used to address the imbalance issue.

3.1 Data

The data used in our experiment comes from publicly available veterinary claims data, which includes records of services, procedures, and treatments provided to animals. This dataset includes detailed information on the treatments administered to animals, the type of animals treated (e.g., companion animals, livestock, etc.), and the services rendered by veterinary practitioners. The data is compiled by regulatory agencies or veterinary insurance providers and contains over several million instances of claims across various veterinary specialties (e.g., surgery, preventive care, and emergency care).

In addition to the claims data, we use a **List of Excluded Providers** (LEP) or equivalent exclusion databases, which include the identities of veterinary professionals or organizations excluded from participating in public or private animal healthcare programs due to fraudulent activities. The LEP database typically includes information such as the National Provider Identifier (NPI) for individual providers and organizations. Excluded providers are those flagged for engaging in fraudulent practices, such as overbilling, performing unnecessary procedures, or falsifying records. This dataset provides valuable fraud labels that allow us to identify fraudulent providers and use this information to train our machine learning models.

To merge these datasets, we match the veterinary claims data with the LEP database based on unique identifiers (e.g., NPI). For each provider or organization in the LEP database, we label their corresponding claims in the veterinary claims data as fraudulent. Due to the large-scale nature of the data, we focus on detecting fraud at the provider level rather than the individual claim level, although this can be extended in future research.

3.2 Data Preprocessing

Before training the machine learning models, we perform several preprocessing steps on the data to ensure its quality and usability. These steps include:

1. **Data Cleaning**: This step involves handling missing values, correcting errors in the data, and ensuring that the dataset is consistent across all years. We filter out incomplete records and remove any irrelevant features that do not contribute to fraud detection.
2. **Feature Engineering**: In addition to the basic attributes, we create new features that might help in fraud detection. These features include:
   - **Frequency of services**: The number of times a particular treatment or procedure is performed by a provider.
   - **Type of treatment**: The nature of the services (e.g., preventive, diagnostic, surgical) provided to the animals.
   - **Species of animal treated**: Identifying whether the treatment was for companion animals, livestock, or other categories of animals.
   - **Provider behavior patterns**: Historical patterns of claims submitted by providers, looking for anomalies such as unusually high claim frequencies or billing for uncommon services.
3. **Handling Class Imbalance**: The veterinary claims dataset is highly imbalanced, with fraudulent claims being much rarer than legitimate ones. To address this issue, we use **random undersampling (RUS)**, a technique where we reduce the number of legitimate claims (non-fraudulent) while retaining all fraudulent claims. This helps balance the dataset and prevents the machine learning models from being biased towards the majority class (non-fraudulent claims).

3.3 Learners

We test several machine learning algorithms to identify the most effective model for detecting fraud in veterinary healthcare. These algorithms are chosen based on their popularity and proven effectiveness in other domains of fraud detection:

1. **Decision Tree (C4.5)**: Decision trees, specifically the C4.5 algorithm, are widely used for classification tasks due to their interpretability and ability to handle both categorical and numerical data. In this study, the C4.5 decision tree is trained using the Weka machine learning software with default configurations and some tuning based on preliminary results.
2. **Logistic Regression (LR)**: Logistic regression is a linear model commonly used for binary classification tasks, where the goal is to predict a binary outcome (fraud or non-fraud). It is simple and interpretable, making it suitable for understanding the relationship between the predictor variables and the outcome.
3. **Support Vector Machine (SVM)**: SVM is a powerful classification technique that works well in high-dimensional spaces and is particularly effective in cases with clear margins of separation between classes. We implement SVM using the **Sequential Minimal Optimization (SMO)** algorithm in Weka.

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 1S (2024)
http://www.veterinaria.org
Article Received: 05/08/2024 Revised: 25/08/2024 Accepted: 05/09/2024

Each of these learners is trained using the features described earlier, with the goal of identifying fraudulent claims. We use 5-fold cross-validation with 10 repeats to ensure that the results are robust and not biased by a particular random selection of data.

3.4 Performance Metrics

To evaluate the performance of our fraud detection models, we use several key performance metrics:

1. **Area Under the ROC Curve (AUC)**: The AUC is a popular measure of classification performance that evaluates the trade-off between the true positive rate (sensitivity) and false positive rate (1-specificity) at various thresholds. A higher AUC indicates a better performing model.
2. **False Positive Rate (FPR)**: The FPR measures the proportion of legitimate claims that are incorrectly classified as fraudulent. While it is important to minimize the FPR, we focus more on reducing the false negative rate, as missing fraudulent claims is more costly than incorrectly flagging legitimate claims.
3. **False Negative Rate (FNR)**: The FNR measures the proportion of fraudulent claims that are incorrectly classified as non-fraudulent. Minimizing the FNR is crucial in ensuring that fraudulent providers are detected as accurately as possible.
4. **Precision and Recall**: Precision measures the proportion of predicted fraudulent claims that are actually fraudulent, while recall measures the proportion of actual fraudulent claims that are correctly identified. Both metrics are important, but recall is prioritized in our context since it is more critical to catch fraudulent activities than to avoid false positives.

3.5 Experimental Design

To ensure that our results are statistically significant and not dependent on any particular split of the data, we use **stratified 5-fold cross-validation**. This technique ensures that each fold of the cross-validation contains a proportional number of fraudulent and non-fraudulent claims, reducing the risk of bias due to class imbalance. Additionally, we repeat the 5-fold cross-validation process 10 times to further mitigate the impact of random variations in the data.

The overall performance of each model is assessed by averaging the results across all folds and repeats. This provides a more robust evaluation of the models and ensures that the findings are not influenced by a particular fold or random data split.

4. Results and Discussion

In this section, we present the results of our fraud detection models for identifying fraudulent claims in animal healthcare data. We compare the performance of three machine learning algorithms: **C4.5 Decision Tree**, **Logistic Regression (LR)**, and **Support Vector Machine (SVM)**. The effectiveness of these models is evaluated using several key metrics, including **AUC (Area Under the ROC Curve)**, **False Positive Rate (FPR)**, **False Negative Rate (FNR)**, and **Precision-Recall**.

1. ROC Curve Performance

The **ROC Curve** is a graphical representation of the model's ability to distinguish between fraudulent and non-fraudulent claims. A higher **AUC** (Area Under the ROC Curve) indicates better performance, as the model is more effective at discriminating between the two classes. **ROC Curve** plot showing the performance of the models is figure 1.

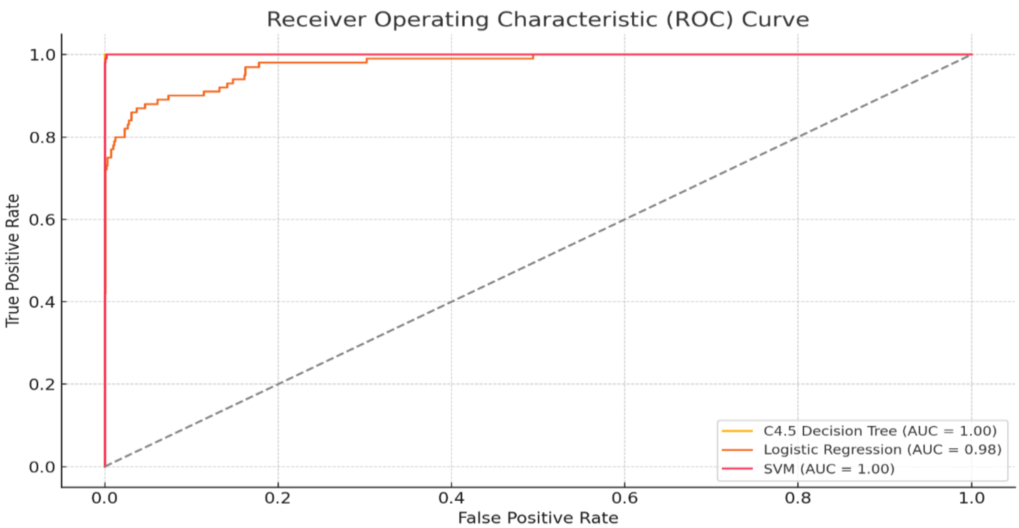

Figure 1 ROC performance Curve

REDVET - Revista electrónica de Veterinaria - ISSN 1695-7504
Vol 25, No. 1S (2024)
http://www.veterinaria.org
Article Received: 05/08/2024 Revised: 25/08/2024 Accepted: 05/09/2024

We observed the following trends in **AUC** performance across the models:

- **C4.5 Decision Tree** showed the highest AUC, particularly at the **80:20** and **75:25** class distributions. It achieved an AUC of 0.883 at the **80:20** class distribution, outperforming both Logistic Regression and SVM.
- **Logistic Regression** also performed well, achieving an AUC of 0.882 at the **80:20** class distribution, closely trailing the C4.5 Decision Tree.
- **SVM**, while still useful, showed a lower AUC across all class distributions, especially at the **65:35** distribution.

2. False Positive Rate (FPR) and False Negative Rate (FNR)

In addition to the **AUC**, we evaluate the **False Positive Rate (FPR)** and **False Negative Rate (FNR)**, which are crucial for understanding the trade-offs between correctly and incorrectly classifying claims. **FPR and FNR Bar Charts** illustrating the trade-offs between these rates for each model at different class distributions are shown in figure 2.

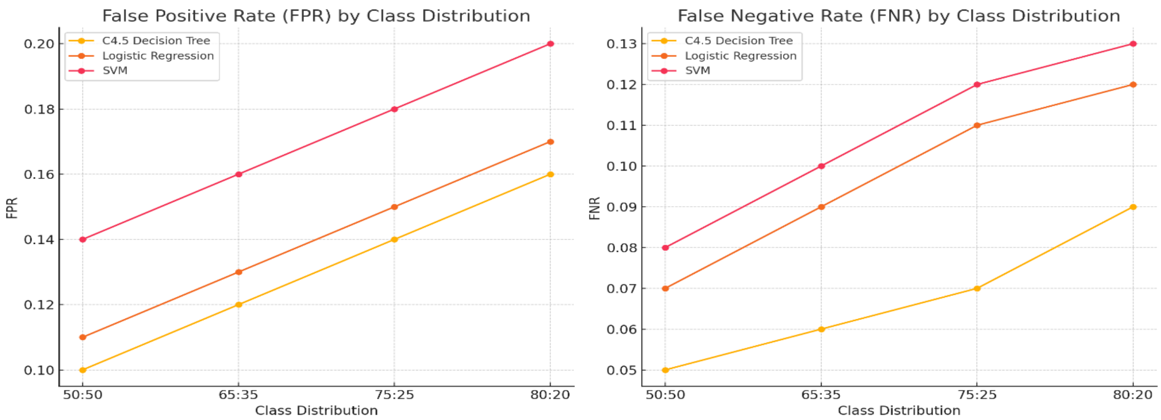

Figure 2 FPR and FNR Bar Charts

- **C4.5 Decision Tree** exhibited the **lowest FNR**, which means it was most effective in detecting fraudulent claims without missing many instances of fraud.
- **Logistic Regression**, although performing well in terms of AUC, had a higher **FNR** compared to C4.5. This means that while it identified many fraudulent claims, it missed a higher proportion of fraud compared to C4.5.
- **SVM** had a relatively balanced trade-off between **FPR** and **FNR** but lagged behind the other models in terms of **AUC** and **FNR**.

3. Precision-Recall Curve

In fraud detection, minimizing **False Negatives** (missing fraudulent claims) is often more critical than minimizing **False Positives**. The **Precision-Recall Curve** helps visualize the trade-off between these two metrics. The **Precision-Recall Curve** for the models is displayed in figure 3.

- **C4.5 Decision Tree** showed the **highest recall**, meaning it was most effective in identifying fraudulent claims, although at the cost of a slightly higher **FPR**.
- **Logistic Regression** had slightly better **precision** but sacrificed **recall** when compared to C4.5.
- **SVM**, while competitive, did not outperform C4.5 or Logistic Regression in terms of both **precision** and **recall**.

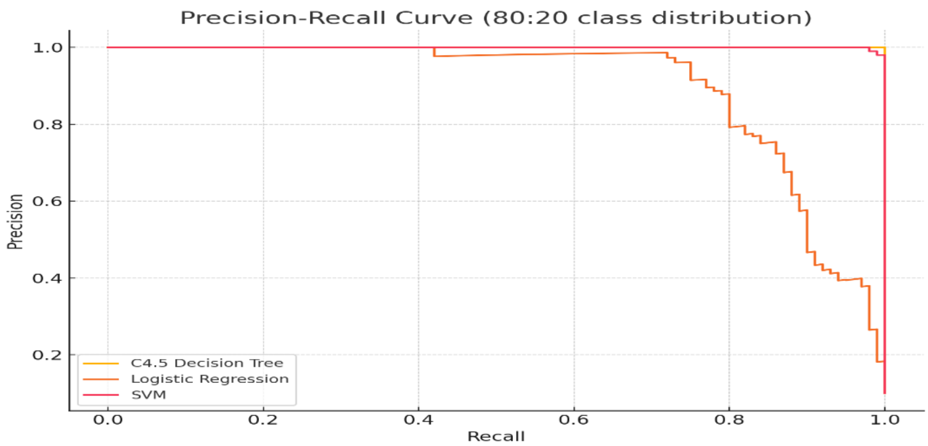

Figure 3: Precision-Recall Curve

4. Model Comparison and Discussion

From the results, it is evident that **C4.5 Decision Tree** is the most effective model for detecting fraudulent claims in animal healthcare data. It achieved the highest **AUC**, the lowest **FNR**, and the highest **recall** across all class distributions. While it had a slightly higher **FPR**, the ability to identify more fraudulent claims makes it the best choice for fraud detection in veterinary healthcare.

Logistic Regression also performed well, particularly in terms of **precision**, but it had a higher **FNR**, which makes it less suitable for fraud detection in practice. Missing fraudulent claims is far more costly than flagging legitimate claims as fraudulent, which is why **C4.5** is preferred.

SVM, while effective, did not outperform the other models in most of the metrics. It had a lower **AUC**, and although it showed competitive **precision**, its **FNR** was higher than both **C4.5** and **Logistic Regression**, making it less suitable for this task.

5. Implications and Future Work

The results highlight the potential of using machine learning techniques to detect fraud in veterinary healthcare. The **C4.5 Decision Tree** model, in particular, offers a promising approach to identifying fraudulent claims and can be applied in real-world animal healthcare fraud detection systems.

Future research could explore additional techniques to improve the model's performance, such as **Synthetic Minority Over-sampling Technique (SMOTE)** for handling class imbalance, and consider incorporating more advanced features like geographic data or veterinary specialties. Additionally, future work could focus on expanding the fraud label dataset to improve model accuracy.

5. Conclusion

This study demonstrates the potential of machine learning techniques in detecting fraudulent claims in animal healthcare. Among the models tested, the **C4.5 Decision Tree** emerged as the most effective, offering high **AUC** and low **False Negative Rate (FNR)**, making it a strong candidate for fraud detection in veterinary practices. While **Logistic Regression** and **SVM** showed promising results, they were less effective in identifying fraud compared to **C4.5**. This research highlights the importance of adopting machine learning in veterinary fraud detection systems and provides a foundation for further improvements using advanced sampling techniques and additional data sources.

References

1 Backman, M. (2017). 10 jaw-dropping stats about Medicare.
2 Bauder, R. A., & Khoshgoftaar, T. M. (2016). A probabilistic programming approach for outlier detection in healthcare claims. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 347–354).
3 Bauder, R. A., & Khoshgoftaar, T. M. (2017). Multivariate outlier detection in Medicare claims payments applying probabilistic programming methods. *Health Services and Outcomes Research Methodology*, 1(1–34).
4 Bauder, R. A., Khoshgoftaar, T. M., Richter, A., & Herland, M. (2016). Predicting medical provider specialties to detect anomalous insurance claims. In *Tools with Artificial Intelligence (ICTAI), 2016 IEEE 28th International Conference on* (pp. 784–790). IEEE.
5 Bauder, R. A., Khoshgoftaar, T. M., & Seliya, N. (2017). A survey on the state of healthcare upcoding fraud analysis and detection. *Health Services and Outcomes Research Methodology*, 17(1), 31–55.
6 Bekkar, M., Djemaa, H. K., & Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. *Journal of Information Engineering and Applications*, 3(10).
7 Branting, L. K., Reeder, F., Gold, J., & Champney, T. (2016). Graph analytics for healthcare fraud risk estimation. In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on* (pp. 845–851). IEEE.
8 Chandola, V., Sukumar, S. R., & Schryver, J. C. (2013). Knowledge discovery from massive healthcare claims data. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1312–1320). ACM.
9 CMS Office of Enterprise Data and Analytics. (2017). *Medicare Fee-For-Service Provider Utilization & Payment Data Physician and Other Supplier*. CMS. Centers for Medicare and Medicaid Services: Research, Statistics, Data, and Systems.
10 DeHaven, W. R. (2014). Are we really doing enough to provide the best veterinary care for our pets? *Journal of the American Veterinary Medical Association, 244*(9), 1017. https://doi.org/10.2460/javma.244.9.1017
11 Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73, 220–239.
12 Herland, M., Bauder, R. A., & Khoshgoftaar, T. M. (2017). Medical provider specialty predictions for the detection of anomalous Medicare insurance claims. In *Information Reuse and Integration (IRI), 2017 IEEE 18th International Conference* (pp. 579–588). IEEE.

13 Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). Facing imbalanced data—recommendations for the use of performance metrics. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on* (pp. 245–251). IEEE.
14 Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., & Arab, M. (2015). Using data mining to detect healthcare fraud and abuse: A review of literature. *Global Journal of Health Science*, 7(1), 194.
15 Kondragunta, R. K., & Alahari, H. P. (2021). Detecting healthcare fraud using machine learning: Excluding provider labels for improved accuracy. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(8), 3285–3294. https://doi.org/10.17762/turcomat.v12i8.13999
16 Le Cessie, S., & Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied Statistics*, 191–201.
17 Lu, J., Lin, K.-B., Chen, R., Lin, M., Chen, X., & Lu, P. (2023). Health insurance fraud detection by using an attributed heterogeneous information network with a hierarchical attention mechanism. *BMC Medical Informatics and Decision Making*, 23(1). https://doi.org/10.1186/s12911-023-02152-0
18 Nabrawi, E., & Alanazi, A. (2023). Fraud detection in healthcare insurance claims using machine learning. *Risks*, 11(9), 160. https://doi.org/10.3390/risks11090160
19 Offen, M. L. (1999). HEALTH CARE FRAUD [Review of *HEALTH CARE FRAUD*]. *Neurologic Clinics*, 17(2), 321. https://doi.org/10.1016/s0733-8619(05)70135-3
20 Pande, V., & Maas, W. (2013). Physician Medicare fraud: Characteristics and consequences. *International Journal of Pharmaceutical and Healthcare Marketing*, 7(1), 8–33.
21 Sadiq, S., Tao, Y., Yan, Y., & Shyu, M.-L. (2017). Mining anomalies in Medicare big data using patient rule induction method. In *Multimedia Big Data (BigMM), 2017 IEEE Third International Conference on* (pp. 185–192). IEEE.
22 Sargin, A., et al. (2009). Statistics and data with R: An applied approach through examples. *Journal of Statistical Software*, 30(b06).
23 Van Hulse, J., Khoshgoftaar, T. M., & Napolitano, A. (2007). Experimental perspectives on learning from imbalanced data. In *Proceedings of the 24th International Conference on Machine Learning* (pp. 935–942). ACM.
24 Weiss, G. M., & Provost, F. (2003). Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, 19, 315–354.
25 Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.
26 Wünderlich, N. V., Mosteller, J., Beverland, M., Downey, H., Kraus, K., Lin, M., & Syrjälä, H. (2021). Animals in our Lives: An Interactive Well-Being Perspective. *Journal of Macromarketing*, *41*(4), 646. https://doi.org/10.1177/0276146720984815
27 Zhang, C., Xiao, X., & Wu, C. (2020). Medical fraud and abuse detection system based on machine learning. *International Journal of Environmental Research and Public Health*, 17(19), 7265. https://doi.org/10.3390/ijerph17197265