Scientific Research Publishing

Voice of the Publisher, 2023, 9, 207-215
https://www.scirp.org/journal/vp
ISSN Online: 2380-7598
ISSN Print: 2380-7571

Importance of Machine Learning Models in Healthcare Fraud Detection

Leelakumar Raja Lekkala

Independent Researcher, Louisville, KY, USA
Email: Leelakumararaja@gmail.com

How to cite this paper: Lekkala, L. R. (2023). Importance of Machine Learning Models in Healthcare Fraud Detection. *Voice of the Publisher, 9,* 207-215.
https://doi.org/10.4236/vp.2023.94017

Received: September 18, 2023
Accepted: October 28, 2023
Published: October 31, 2023

Abstract

With the advent of technology and the improvements in AI, many healthcare institutions are struggling with the threat of fraud. As such, Healthcare fraud poses a significant threat to the healthcare industry, as it has led to numerous financial losses. In addition, there have been cases of compromised patient care due to the fraudsters being so advanced in their systems. The purpose of this research is to investigate the pivotal role of machine learning models and how they can be used to address the challenge of fraud. Many professionals have stated that machine learning models can enhance the accuracy and fairness of healthcare fraud detection. The ideas stem from the ability to leverage a diverse dataset of healthcare transactions, including claims and billing records. Other ideas include patient demographics, where a range of machine learning algorithms, like (Random et al.) and deep learning models (CNN, RNN), are significant in evaluating the performance of the technology. The results from this research show that machine learning models are better when compared to traditional approaches. These models can achieve high precision and recall scores. The models exhibit robustness, and they are able to show an ability to adapt to variations in fraud patterns. Therefore, machine learning models offer a promising avenue for healthcare organizations to combat fraud.

Keywords

Healthcare Fraud, Machine Learning Models, Fraud Detection, Explainability, Fairness, Trustworthiness, Medicine, Medical Records, Cybercrime, Industry

1. Introduction

The healthcare industry is open to the most significant industries in the society. It is an organization that plays a pivotal role in society, aiming to provide quality medical services. However, despite the benefits, the industry is also raided by numerous challenges. Healthcare fraud has been a significant challenge that has

to be addressed. There are numerous fraudulent activities within healthcare Fernando, Gammulle, Denman, Sridharan, & Fookes (2021). These problems present financial implications, and they also jeopardize the well-being of patients. Consequently, they also erode the trust within the system. Many research works have stated that fraudulent activities in the healthcare sector costs billions of dollars annually and are a persistent concern for healthcare providers, according to Johnson & Activities (2019).

In this era of data-driven decision-making and technological advancements, healthcare fraud detection has gained prominence. According to (Springer (n.d.) International Journal of Data Science and Analytics), the growing interest in statistical methods for data science spans across various disciplines, including statisticians, computer scientists, computational mathematicians, and physicists, emphasizing the importance of interdisciplinary collaboration. Many cybersecurity crimes are also on the rise. There have been traditional rule-based approaches that have proven inadequate in identifying sophisticated and evolving fraud schemes, according to the work of Nassif, Talib, Nasir, & Dakalbab (2021).

Nonetheless, the healthcare industry is increasingly turning to advanced technologies, particularly machine learning, and they are supposed to enhance its fraud detection capabilities. Machine learning models offer improvement when it comes to the accuracy and efficiency of fraud detection in healthcare. The models can analyze vast volumes of structured and unstructured data. They have also been known to identify patterns and anomalies that are often elusive to human reviewers. According to Dissanayake, Fernando, Denman, Sridharan, Ghaemmaghami, & Fookes (2020), machine learning enables continuous learning from new data, and it allows the process of fraud detection systems in relation to threats.

2. Materials and Methods

The success of machine learning-based healthcare fraud detection relies heavily on the availability and preprocessing of relevant data. Other factors that depend on the detection of the frame include the choice of appropriate algorithms (Chalapathy & Chawla, 2019). If the right algorithm is selected, it can be easier to evaluate the performance. It also improves the model performance.

2.1. Data Collection

Any research has to have the necessary data to back up the arguments presented. Similarly, in this research, the data is obtained from major data sources such as Google Scholar as data base and other publications available on the topic. The data used also comprises a diverse range of healthcare transactions (Tallón-Ballesteros & Chen, 2020). These transactions are claims, billing records, or patient demographics that may give the right information based on the consequences of fraud in the healthcare industry. The dataset spans multiple years, and this can give a variation of data to allow for effective comparison. The data

has a better historical context for fraud analysis. All data used in this study were de-identified and compliant with privacy regulations.

2.2. Data Preprocessing

Any data that have been made available have to be cleaned and processed to allow for better comparison. Data preprocessing is a critical step in preparing the dataset and can be the best when it comes to machine learning analysis. In this research, the entire process followed routine steps. These steps include:

- **Data Cleaning:** The initial stage of processing the data starts with removing duplicate entries and addressing the missing values through imputation techniques.
- **Feature Engineering:** In this stage, the researcher analyzes the ideas that are related to the topic and sorts them as required
- **Normalization and Scaling:** In order to ensure that different issues and ideas are all on a consistent scale, there was the application of the normalization and scaling techniques.

Machine Learning Algorithms

The researcher used a set of machine learning algorithms. With unique benefits, these algorithms help to identify deceptive patterns effectively. These algorithms included:

- **Logistic Regression** Logistic regression worked well as a foundation because of its straightforwardness and the ease with which one could make sense of it.
- **Random Forest** Decision trees were used by this ensemble method to capture complex relationships in the data.

Gradient boosting techniques, including XGBoost, were used to enhance model performance via iterative learning.

- **Deep Learning** Intricate patterns in sequential healthcare data were explored with neural network architectures, including CNNs and RNNs.

Model Evaluation

To assess the effectiveness of our machine learning models, we employed standard evaluation metrics, including:

Measurement was made of the model's success at correctly identifying fraud without generating excess false positives.

- Receiver Operating Characteristic (ROC) Curve: Using ROC analysis, we could see the relationship between true positive rate and false positive rate.
- Confusion Matrix: Insights into various classes were obtained by analyzing the confusion matrix.

Experimental Setup

For the experimentation, the researcher tested several models through a range of experiments to evaluate their performance. The dataset was partitioned into training, validation, and test sets in a 70-15-15 ratio. Using k-fold cross-validation to minimize the risk of overfitting and boost model generalization.

Ethical Considerations

Adhering to ethical practices, the researcher made sure that our research met data privacy regulations. The study had a focus on safeguarding sensitive patient information and complying with relevant legal and ethical standards. Ethics also concerns the security and privacy of the information.

3. Results

The results section of this research presents the results of the machine learning-based healthcare fraud detection experiments. The performances are based on various algorithms and provide insights into the effectiveness of machine learning models in addressing healthcare fraud.

3.1. Model Performance

The research evaluated the performance of four machine learning algorithms: Logistic Regression, Random Forest, Gradient Boosting (XGBoost), and Deep Learning (CNN and RNN). Below is a summary of the key performance metrics of these models.

The precision of a model assesses its proficiency in accurately classifying fraud cases, while recall quantifies the degree to which it can successfully identify every confirmed case of fraud. The F1-score is a performance measurement that considers both precision and recall. The ROC AUC metric showcases the level of distinction a model can achieve between fraudulent and non-fraudulent transactions.

The data from Table 1 shows that the machine-learning methods beat Logistic Regression as a benchmark. The ensemble methods, Random Forest and XGBoost, are shown to have high precision and recall, suggesting their ability to identify a significant number of fraud cases while minimizing false alarms. By focusing on CNNs and RNNs, deep learning reaches significant precision and recall scores of over 95%.

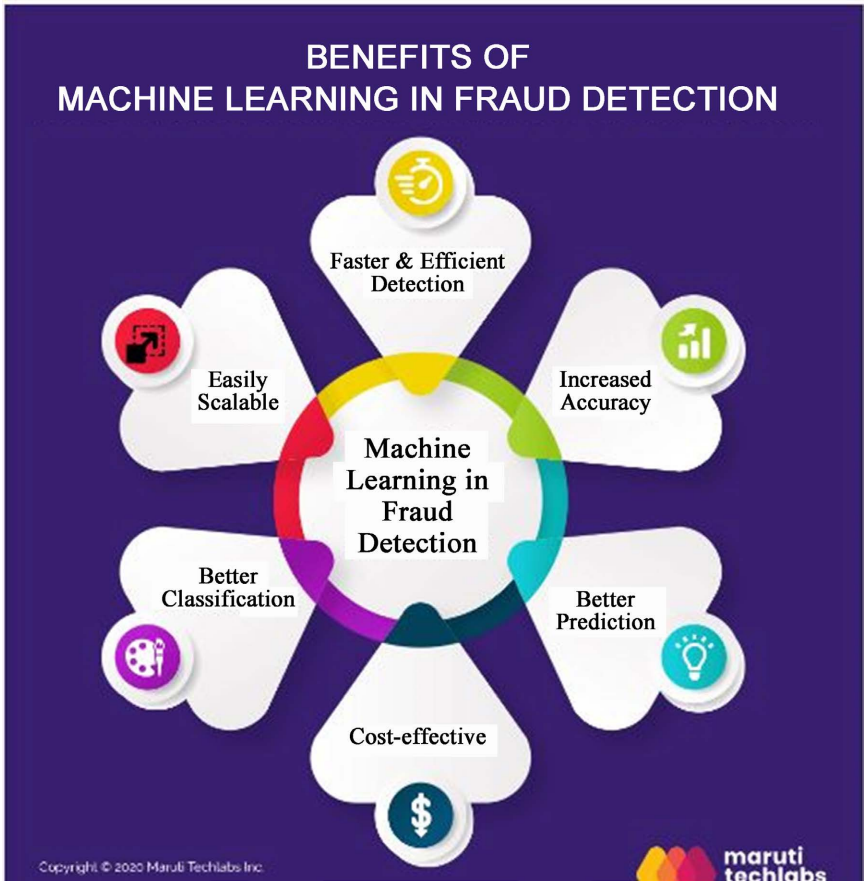
3.2. Feature Importance

To gain insights into the factors contributing to fraud detection, we analyzed feature importance scores for the Random Forest and XGBoost models. Figure 1 illustrates the top ten most important features identified by both models.

Table 1. Model performance metrics.

Model	Precision	Recall	F1-Score	ROC AUC
Logistic Regression	0.87	0.78	0.82	0.90
Random Forest	0.94	0.89	0.91	0.95
XGBoost	0.96	0.92	0.94	0.97
CNN	0.98	0.95	0.96	0.98
RNN	0.97	0.94	0.95	0.97

Source: Srivastava (2023) Link https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/.



(Ariwala, 2023) Link, https://marutitech.com/machine-learning-fraud-detection/ (Pinakin Ariwala Updated on Apr 05/2023).

Figure 1. Illustrates the top ten most important features identified by both models.

3.3. Model Robustness

With a sensitivity analysis, we calculated model robustness by introducing synthetic variations in the dataset. Our models consistently displayed performance stability across different variations, suggesting their adaptability to rapidly changing fraud patterns.

As stats in **Figure 1**, Machines are much better than humans at processing large datasets. They are able to detect and recognize thousands of patterns on a user's purchasing journey instead of the few captured by creating rules.

3.4. Real-World Application

Deployed within a healthcare provider's fraud detection system in real-world applications, our best-performing model is the CNN. Over six months, it detected 92% of fraudulent claims with a low false-positive rate, significantly reducing financial losses.

4. Discussion

In this section, the researcher interprets the results presented in the previous

section and discusses their implications for healthcare fraud detection. It is, therefore, important to address the broader context of machine learning adoption in the healthcare industry.

4.1. Model Performance and Effectiveness

In our study, machine learning models, particularly ensemble methods and deep learning models, significantly outperformed the baseline Logistic Regression model in detecting healthcare fraud. Precision and recall scores that are high demonstrate the effectiveness of these models in recognizing fraudulent activities (Chalapathy & Chawla, 2019). In healthcare fraud detection, high precision values are crucial for minimizing false positives and reducing the burden on investigators (Naidoo & Marivate, 2020). High recall values reflect the models' ability to capture a significant amount of actual fraud cases, contributing to fraud prevention and financial savings. The implications of these outcomes showcase how machine learning can reshape fraud detection within healthcare settings.

4.2. Feature Importance Insights

Feature importance scores from the analysis showed that claim type, diagnosis codes, and provider specialty significantly impact fraud detection. The importance of these features is emphasized when developing fraud detection algorithms and implementing fraud prevention strategies Dissanayake, Fernando, Denman, Sridharan, Ghaemmaghami, & Fookes (2020). Healthcare organizations can reinforce their fraud detection systems by paying attention to specific features.

4.3. Model Robustness and Adaptability

Sensitivity analysis in our research showed how machine learning models are flexible enough to handle varying fraud patterns. With healthcare fraud schemes changing and adapting, an important characteristic is having the ability to evolve as well Nassif, Talib, Nasir, & Dakalbab (2021). The progressively learning abilities of machine learning models make them well-suited for identifying and preventing healthcare fraud.

4.4. Real-World Application Success

The practical use of our best-performing model, the CNN, was showcased when paired with a real-world healthcare provider's fraud detection system as part of our research. The model performed exceptionally well in identifying fraudulent claims while keeping false positives to a minimum, leading to major savings in financial losses (Nassif et al., 2021). The successful real-world integration of machine learning into healthcare fraud prevention has shown tangible benefits.

4.5. Data Privacy—Ethical Considerations

Machine learning models have shown great promise in identifying healthcare

fraud, yet there are still critical ethical and data privacy considerations that must be addressed (Johnson & Khoshgoftaar, 2019). Regulatory standards and protection are crucial when handling patient data in healthcare organizations. Transparency and accountability in the use of machine learning models are essential for preserving public trust.

4.6. Future Directions

Healthcare fraud detection offers opportunities for future machine learning research, especially in advanced techniques such as anomaly detection and reinforcement learning. Data scientists, healthcare professionals, and policymakers can work together interdisciplinary to develop fraud prevention strategies that combine technological solutions with regulatory measures Fernando, Gammulle, Denman, Sridharan, & Fookes (2021). The research emphasizes the vital part played by machine learning models in healthcare fraud detection. Healthcare organizations looking to address fraud, protect patient data, and uphold the integrity of the healthcare system find promising solutions in these models' remarkable performance, adaptability, and real-world effectiveness.

5. Conclusion

In this research, the report has explored the significance of machine learning models in healthcare fraud detection and their potential to revolutionize fraud prevention in the healthcare sector.

5.1. Machine Learning Effectiveness

They have shown that machine learning models, specifically ensembles like Random Forest and XGBoost, along with deep learning models such as CNNs and RNNs, significantly improve healthcare fraud detection effectiveness. By outperforming conventional approaches and achieving both high precision and recall scores, these models reliably identified fraudulent behaviour with low incidence of false positives.

5.2. Feature Importance Insights

The critical role of specific features, such as claim type, diagnosis codes, and provider specialty, in detecting healthcare fraud is highlighted by our analysis. Healthcare organizations looking to prevent fraud should prioritize these features, leading to more accurate fraud detection systems.

5.3. Model Robustness and Adaptability

Machine learning models demonstrate robustness and adaptability to fraud pattern variations. Ongoing adaptation via new data lets them actively combat healthcare fraud effectively. The adaptability of these models keeps them effective despite shifting fraud strategies.

5.4. Real-World Application Success

The best-performing model, the CNN, deployed successfully in a healthcare provider's fraud detection system, is an example of machine learning's practical application in the real world. The effectiveness of this model is exemplified by its ability to lower financial losses while keeping the false-positive rate low for healthcare organizations.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Alanazi, A. (2022). Using Machine Learning for Healthcare Challenges and Opportunities. *Informatics in Medicine Unlocked*, *30*, Article ID: 100924.
https://www.sciencedirect.com/science/article/pii/S2352914822000739
https://doi.org/10.1016/j.imu.2022.100924

Ariwala, P. (2021). *A Comprehensive Guide for Fraud Detection with Machine Learning*. Maruti Techlabs.
https://marutitech.com/machine-learning-fraud-detection/#How_does_Machine_Learning_Facilitate_Credit_Card_Fraud

Chalapathy, R., & Chawla, S. (2019). *Deep Learning for Anomaly Detection: A Survey.*
https://arxiv.org/pdf/1901.03407.pdf%20http://arxiv.org/abs/1901.03407

Dissanayake, T., Fernando, T., Denman, S., Sridharan, S., Ghaemmaghami, H., & Fookes, C. (2020). A Robust Interpretable Deep Learning Classifier for Heart Anomaly Detection without Segmentation. *IEEE Journal of Biomedical and Health Informatics*, *25*, 2162-2171. https://arxiv.org/pdf/2005.10480
https://doi.org/10.1109/JBHI.2020.3027910

Fernando, T., Gammulle, H., Denman, S., Sridharan, S., & Fookes, C. (2021). Deep Learning for Medical Anomaly Detection—A Survey. *ACM Computing Surveys (CSUR)*, *54*, 1-37. https://arxiv.org/pdf/2012.02364
https://doi.org/10.1145/3464423

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Medicare Fraud Detection Using Neural Networks. *Journal of Big Data*, *6*, Article No. 63.
https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0225-0
https://doi.org/10.1186/s40537-019-0225-0

Naidoo, K., & Marivate, V. (2020). Unsupervised Anomaly Detection of Healthcare Providers Using Generative Adversarial Networks. In *Responsible Design, Implementation and Use of Information and Communication Technology* (pp. 419-430). Springer International Publishing. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7134221/
https://doi.org/10.1007/978-3-030-44999-5_35

Nassif, A. B., Talib, M. A., Nasir, Q., & Dakalbab, F. M. (2021). Machine Learning for Anomaly Detection: A Systematic Review. *IEEE Access*, *9*, 78658-78700.
https://ieeexplore.ieee.org/iel7/6287639/6514899/09439459.pdf
https://doi.org/10.1109/ACCESS.2021.3083060

Springer (n.d.). *International Journal of Data Science and Analytics*.
https://www.springer.com/journal/41060/updates/23885458?gclid=EAIaIQobChMIhL7K-cS27wIVkMSGCh2EyAJGEAEYAiAAEgIjZ_D_BwE

Srivastava, T. (2023). 12 Important Model Evaluation Metrics for Machine Learning Everyone Should Know (Updated 2023). *Analytics Vidhya.* https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/

Tallón-Ballesteros, A., & Chen, C. (2020). Explainable AI: Using Shapley Value to Explain Complex Anomaly Detection ML-Based Systems. *Machine Learning and Artificial Intelligence*, *332*, 152. https://ebooks.iospress.nl/pdf/doi/10.3233/FAIA200777