Application of Machine Learning in Insurance Fraud Detection: Achievements and Future Prospects

Yourui Guo

Department of Mathematics, University of Illinois, Urbana IL, 61801, USA
morh@usf.edu

Abstract. Insurance is a crucial component of modern society and insurance fraud inflicts significant financial losses on candid customers, insurance companies, and the entire economy. The insurance industry constantly searches for measures to confront the challenge of fraudulent claims. Traditional ways to identify insurance fraud were not effective enough in today's digital world. Recent technological development in artificial intelligence and machine learning brings revolution to insurance fraud detection methods. This paper examines various applications of artificial intelligence that detect fraudulent behaviors in insurance claims. More specifically, this study identifies key challenges for fraud detection in different types of insurance, and explains how to overcome these obstacles with the use of several particular machine learning algorithms. The findings show that application of machine learning in various insurance fraud detection systems significantly improves prediction accuracy and overall efficiency. Artificial intelligence powers new methods to confront fraudulent behaviors that are previously unimaginable in the insurance industry. This paper also discusses other advantages and limitations of existing machine learning based fraud detection methods. The result demonstrates that applications of artificial intelligence and machine learning have already made huge contributions in fighting insurance fraud. The analysis also points out several potential directions of future researches for utilizing machine learning in insurance fraud detection systems.

Keywords: Machine learning, artificial intelligence, neural network, fraud detection, insurance.

1 Introduction

Insurance has become one of the fundamental building blocks of modern society and economy. People all around the world today rely on governmental healthcare support or corporate medical benefits to pay for hospitalization and treatment costs [1]. Automobile insurance is legally mandated in most countries. However, the insurance industry suffers huge annual losses due to fraudulent claims. The Federal Bureau of Investigation (FBI) estimated that insurance related fraud is responsible for approximately $80 billion economic losses each year in the U.S alone, making it the second-largest white-collar crime [2]. To offset the excessive expenses caused by

Y. Wang (ed.), *Proceedings of the 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024)*, Advances in Intelligent Systems Research 185,
https://doi.org/10.2991/978-94-6463-512-6_65

fraud, insurance providers are forced to increase premiums dramatically, which in turn hurt people who actually need support and severely hinder the optimization of the whole industry. Therefore, it is necessary to develop accurate, effective and efficient insurance fraud detection methods.

Traditionally, insurance providers rely on manual work by trained professionals to examine and identify fraudulent claims. However, this approach is often complex and time-consuming [2]. It also poses additional drawbacks such as increased possibility of auditing error and rising cost in human resources [3]. Furthermore, the large amount of work required by traditional fraud detection methods takes longer time to process. This delay reduces the efficiency for legitimate claim payouts. Insurance providers might also incur extra losses in actual fraud cases due to late decisions [2]. Fortunately, technological revolutions in Artificial Intelligence (AI), machine learning, and big data in recent years shed new light on the insurance fraud issue. AI solutions have already proved their astonishing power in the financial services industry. The integration with new technological advances presents advantages in many ways. On one hand, it can drastically increase process automation and reduce operational costs. On the other hand, AI can transform and restructure existing practices, creating new innovative business models [4]. The application of AI technologies in insurance fraud detection presents a very promising future.

In the past few years, researches of new fraud detection methods powered by AI, machine learning and data analysis have already produced some significant results. For example, Dhieb et al. developed a secure and automated framework based on block chain to identify fraudulent claims, alert potentially risky customers, and minimize economic losses for insurance providers [2]. Sun et al. established a new method using abnormal group mining to distinguish fraudsters from normal customers [5]. These models utilizing cutting edge technologies deliver much better accuracy for detecting fraudulent claims in contrast with traditional methods. AI powered insurance fraud detection methods have already begun to make big improvements to the industry. This paper aims to analyze different methods proposed in recent researches, discuss their strengths and weaknesses, and highlight potential directions for future research.

The remainder of this paper is structured as follows. Section 2 first gives a general explanation for machine learning based fraud detection process, and then introduces different applications of fraud detection methods in two major types of insurance. Section 3 discusses the advantages and drawbacks of AI fraud detection solutions. A foresight of future research direction is also given based on the analysis. Finally, Section 4 sets conclusion for this paper.

2 Method

2.1 General Introduction of AI/ML Based Fraud Detection Methods

Currently, there are a wide variety of AI and machine learning techniques available for building insurance fraud detection models. Most existing researches focus their

study on one or two particular approach. However, combining several machine learning methods into one solution package may also work in appropriate situations. For example, an integrated model may use unsupervised learning to uncover strange behaviors, then use supervised learning to further identify potential fraudulent cases [6].

While individual models can be created with vastly different approaches, there is a general process shown in Fig. 1 for establishing AI/ML based insurance fraud detection solutions. This process can typically be described using the following steps. Phase 1 builds a training dataset for the algorithm to learn from. A large amount of sample customer data, both normal and fraudulent, were collected. These data will go through preliminary processing to become regularized. Appropriate labels may also be attached to each training data sample in this step. Phase 2 train the algorithm with the dataset. The algorithm will process training data fed into it, and generates predictions for each case. In supervised learning, predictions of normal or fraud are then compared with the given label. The training process continues until the performance and accuracy of the algorithm reaches a satisfactory level. Lastly, phase 3 will be applying the well-trained algorithm with new customer data, allowing it to help identify potentially fraudulent claims.

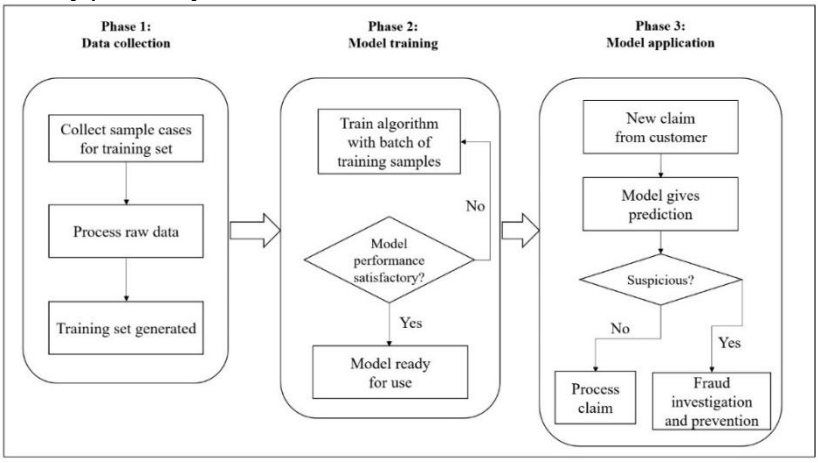

Fig. 1. Generalized fraud detection model training and application process (Photo/Picture credit : Original).

2.2 Fraud Detection Methods in Property & Casualty Insurance Services

In Property & Casualty insurance claims, two important issues must be carefully examined: Cause of the damage to the insured property, and severity of the damage. The application of Computer Vision may provide great assistance in analyzing these two key elements. Pictures of damaged properties were processed with an appropriate model to assess the scope of damage. Other key features can also be extracted to help determine the type of accident. For example, Sahni et al. proposed a model that integrated multiple algorithms to identify fraudulent insurance claims regarding field fires. First, Semantic Segmentation will map out areas of crop with a bird's eye view

image of the field. Various sensors continuously monitor the temperature, humidity, and infrared readings. They function both as alarms and as recorders for the environmental parameters of the fire. When an accident occurs, drone will take images for the crop regions. These images of field fire are then processed with multiple classification algorithms. In the end, the model will determine whether the fire is a natural disaster or a deliberate human act, based on parameters from all sensors joined with fire classifiers from images [7]. Such models have the potential to automate the process of claim validation for Property & Casualty insurances.

Automobile insurance is another major component of the Property & Casualty insurance sector. Fraudulent claims for automobile insurance can be very hard to identify. This is largely due to the complexity of car accidents and intricate cover-up measures of fraudsters. Many researches on fraud detection algorithms for automobile insurance have been done within the past five years. A wide variety of techniques have been proposed, explored, and tested [8]. Amongst these researches, supervised machine learning is the most commonly used methodology for classifying automobile insurance claims and locating potential frauds. From the year 2019 to 2023, more than twenty new solutions have been proposed in this category. These studies involve ensemble learning, neural network, natural language processing, and other supervised learning algorithms [8]. However, some researchers favor unsupervised machine learning methods. This is largely because not enough data of automobile insurance are properly labeled for the training purpose of supervised fraud detection approaches [8].

2.3 Fraud Detection Methods in Health and Medical Care Insurance Services

The issue of fraudulent claims is also critical in the Health and medical care insurance sector. Unlike Property & Casualty insurance claims, Healthcare insurance frauds can be even more complex. Because aside from the customer and insurance company, medical service providers such as hospital, physician and pharmacy also play a crucial role in the claim or reimbursement procedures. Healthcare frauds can occur in many forms: Customers may create false claims; providers may fake charges for services with a high price tag that were never actually performed [9]. Due to the complexity of relationships between all parties involved in a Healthcare insurance claim, simple conventional machine learning algorithm alone may not be enough to comprehend the whole picture and generate the best fraud identification results. Therefore, some researchers began to explore building Healthcare fraud detection models with graph neural networks (GNN). This new type of machine learning algorithms has the ability to learn connected relationships directly from a graph-structured dataset. Which makes GNN algorithms good candidates for identifying collusion relationships in Medicare frauds [10]. However, GNN also has its drawback: it requires significant computational resources when learning from a big dataset, greatly reducing the time efficiency for the entire fraud detection model. Another approach to tackle this issue is incorporating graph centrality measures together with conventional machine learning algorithms. Graph information is first extracted from the complex relationship between insurance company, hospital, and customer. Then a traditional machine learning model was built with the extracted information as one feature.

Efficiency of this alternative method proves to be better than GNNs while maintaining similar levels of performance and accuracy [10].

In Health and medical care insurance services, the issue of security must also be treated with close attention. Electronic health records hold sensitive data that can be vulnerable to unintentional security leaks or deliberate attacks. A fraud detection solution that combines blockchain technology and machine learning algorithm can be used to confront this challenge. For example, Alnuaimi et al. proposed such a system that separates registration of entities and validation of transactions in a pair of smart contracts [11]. The registration section is responsible for confirming identities of the patient, physician, pharmacy, and insurance company agent. Registration process is oversighted and can only be initiated by regulatory authorities, ensuring its authenticity. Once properly registered, each entity can then obtain appropriate access to the approval smart contract bases on their identity. They can then execute functions within that contract, completing necessary steps for validating a health insurance claim, such as assigning prescription, pharmacy verification, insurance claim request, and payment approval [11].

3 Discussion

3.1 Efficiency and Accuracy

Based on the methods introduced in the previous part of this paper, AI and machine learning technologies significantly changed the dynamics of insurance fraud detection solutions. With the aid of these new methods, insurance companies now have the ability to search and identify potential fraud cases in mountains of data [12]. Among the vast variety of machine algorithms and their unique applications, one main advantage they share in common is significantly better efficiency and prediction accuracy over traditional non-AI fraud detection methods. When used properly, AI technology have the power to make great achievement in fighting frauds for both Property & Casualty insurance and Healthcare insurance services.

However, there are still some aspects that can be studied and improved to make these algorithms even better. Selecting the appropriate algorithm for a particular task can have significant impact on the overall performance of the fraud detection model. Establishing a training set with proper labels and sample sizes also plays a key role in efficiency. More training data typically yields higher prediction accuracy, but the rate of improvement may plummet when the size of training set exceeds certain thresholds. From that point on, the downsides of additional cost of computational resources may outweigh benefits of small accuracy gains.

3.2 Data Interpretability

The process of algorithm training and model evolution also raises important issues. The first is data interpretability. In real application of AI based insurance fraud detection methods, the model will be trained with enough data so that its accuracy

converges to a satisfactory level. Then, new customer data is processed with the well-trained model to generate predictions or weights, these numbers will identify potential fraudulent claims or transactions. However, the meaning of these numerical results from the model output are ultimately interpreted by insurance company analysts. For example, the algorithm compares new data with existing patterns established from the training dataset, then mark people who displayed abnormal behaviors as fraud suspects. However, there may be special occasions that cause ordinary people to behave outside such patters. This led to false positives in the model [5]. In such scenarios, it is up to the insurance company analysts to make the final decision on how to interpret algorithm outputs. Therefore, improving data interpretability is necessary to make the model application easier and more user friendly.

3.3 Distribution Difference

Another issue worth studying is distribution difference. In real applications, the training dataset is typically drawn from either public databases, or from the insurance company's own historical customer data. However, the customer base for each individual insurance company are somewhat unique from the others. This diversity can be caused by demographic characteristics, advertising strategies, pricing ranges, and other relevant factors. These differences can create bias between the training dataset and the test dataset of real new customers. For example, AI models trained with data from a large diverse public dataset may not serve to be the perfect fit for an insurance company operating only in a small region or serves only certain customer groups. Algorithm developed for one company may also perform with accuracy less than expected when applied to other companies. How to resolve or remedy such distribution differences is an interesting future prospect for the development of AI fraud detection algorithms.

3.4 Privacy and Security

Data use in insurance fraud detection can be sensitive, particular in the Healthcare insurance sector. Sensitive data includes health records, credit card payments, and other private information. In the application process of AI fraud detection method, such sensitive data may pass through the hands of each entity in the system. Therefore, it is essential to protect the privacy of sensitive information, and enforce strict access restrictions. These preventative restrictions must be considered from the very beginning of algorithm development, and built into the model itself. Security measures such as digital signature and data encryption should be used in the fraud detection system whenever necessary. This is to protect sensitive data from privacy leaks or malicious cyber-attacks. Preservation of privacy and security should always be one of the top priority for AI fraud detection algorithm developers, and for insurance companies that use these models.

4 Conclusion

This paper studied different application of AI and machine learning technologies for insurance fraud detection. Various methods and approaches of AI-based fraud detection models are explained, analyzed, and discussed. This paper revealed that new fraud detection methods empowered by AI can achieve significantly higher accuracy and efficiency compared to traditional methods. The paper also highlighted potential limitations of existing AI-based models in data interpretability and distribution difference. Finally, this research pointed out that future prospect for AI-based fraud detection methods can be focused on relieving such limitations and on better preservation for privacy and security.

References

1. Matloob, I., Khan, S.A., Rukaiya, R., Khattak, M.A.K., Munir, A.: A Sequence Mining-Based Novel Architecture for Detecting Fraudulent Transactions in Healthcare Systems. IEEE Access 10, 48447-48463 (2022).
2. Dhieb, N., Ghazzai, H., Besbes, H., Massoud, Y.: A Secure AI-Driven Architecture for Automated Insurance Systems: Fraud Detection and Risk Measurement. IEEE Access 8, 58546-58558 (2020).
3. Kapadiya, K. et al.: Blockchain and AI-Empowered Healthcare Insurance Fraud Detection: An Analysis, Architecture, and Future Prospects. IEEE Access 10, 79606-79627 (2022).
4. Cosma, S., Rimo, G.: Redefining insurance through technology: Achievements and perspectives in Insurtech. Research in International Business and Finance 70, Part A, 102301 (2024).
5. Sun, C., Yan, Z., Li, Q., Zheng, Y., Lu, X., Cui, L.: Abnormal Group-Based Joint Medical Fraud Detection. IEEE Access 7, 13589-13596 (2019).
6. Saddi, V.R., Gnanapa, B., Boddu, S., Logeshwaran, J.: Fighting Insurance Fraud with Hybrid AI/ML Models: Discuss the Potential for Combining Approaches for Improved Insurance Fraud Detection. In: 2023 4th International Conference on Communication, Computing and Industry 6.0 (C216), Bangalore, India, pp. 01-06 (2023).
7. Sahni, S., Mittal, A., Kidwai, F., Tiwari, A., Khandelwal, K.: Insurance Fraud Identification using Computer Vision and IoT: A Study of Field Fires. Procedia Computer Science 173, 56-63 (2020).
8. Schrijver, G., Sarmah, D.K., El-hajj, M.: Automobile insurance fraud detection using data mining: A systematic literature review. Intelligent Systems with Applications 21, 200340 (2024).
9. Matloob, I., Khan, S.A., Rahman, H.U.: Sequence Mining and Prediction-Based Healthcare Fraud Detection Methodology. IEEE Access 8, 143256-143273 (2020).
10. Yoo, Y., Shin, J., Kyeong, S.: Medicare Fraud Detection Using Graph Analysis: A Comparative Study of Machine Learning and Graph Neural Networks. IEEE Access 11, 88278-88294 (2023).
11. Alnuaimi, A., Alshehhi, A., Salah, K., Jayaraman, R., Omar, I.A., Battah, A.: Blockchain-Based Processing of Health Insurance Claims for Prescription Drugs. IEEE Access 10, 118093-118107 (2022).

12. Byrapu Reddy, S., Kanagala, P., Ravichandran, P., Pulimamidi, R., Sivarambabu, P.V., Polireddi, N.S.A.: Effective fraud detection in e-commerce: Leveraging machine learning and big data analytics. Measurement: Sensors 33, 101138 (2024).