Securing Healthcare Finances: AI Approach to Insurance Fraud Detection

Ashwini Surjuse

Mrs. Sneha Deshmukh
Assistant Professor

Department of Computer Engineering,
Dhole Patil college of Engineering

Abstract- **Now a days, Health insurance has become very important issue in terms of security as the number of health issues are increasing. Healthcare insurance policies and services are facing security issue since few years as policies and other services contain lots of private and individual data. Fraud and abuse are continuously increasing in the healthcare system. So, it is essential for the system to secure the data with respect to the payment and amount. With each and every new policy, the system gets new and personal data of the policy holder, therefore it is mandatory to detect the fraud and not allow to steal the data by third party. Inspired of that, we provide a comprehensive survey of methods applied to health care fraud detection, with focuses on classifying fraudulent behaviors and major sources, also discussing the data preprocessing, as well as feature extraction and comparing fraud detection methods. Multiple artificial intelligence methods are collectively illustrated for fraud detection. Finally, the paper introduces the future research areas and concludes with an overall discussion.**

Keywords- Artificial Intelligence, Fraud detection, Healthcare Insurance, Machine Learning.

1. INTRODUCTION

Healthcare fraud is indeed a serious issue that can have far-reaching consequences. Not only does it impact the financial stability of healthcare systems, but it also undermines the trust between patients and providers, which is essential for effective healthcare delivery. The staggering cost associated with healthcare fraud underscores the urgency to address this issue comprehensively. To combat healthcare fraud effectively, a multi-faceted approach is necessary. This approach may include implementing robust detection systems leveraging advanced technologies like artificial intelligence and data analytics to identify suspicious patterns and anomalies in billing and claims. Additionally, stringent regulatory measures and enforcement actions can serve as deterrents against fraudulent activities.

The choice of domain is crucial in AI development because it determines the specific challenges, requirements, and nuances that the AI system needs to address. AI models trained in one domain may not perform optimally in a different domain without further adaptation or training. Understanding and defining the domain is essential for creating effective and specialized AI solutions tailored to specific industries or applications. This research operates at the intersection of computer science, artificial Intelligence, healthcare domain: AI systems focused on medical diagnosis, drug discovery, patient care, and healthcare analytics, financial domain: AI applications for fraud detection, algorithmic trading, credit scoring, and financial forecasting and machine learning to detect the frauds which can be happens now a days. In this study, we present the technique, which can deal with the methods applied to health care fraud detection, with focuses on classifying fraudulent behaviors and major sources, also discussing the data preprocessing, as well as feature extraction and comparing fraud detection methods. Multiple artificial intelligence methods are collectively illustrated for fraud detection. Finally, the paper introduces the future research areas and concludes with an overall discussion. The rest part of the paper is arranged as follows. In Section 2, we present previous work. In Section 3, we present the proposed model and multiple algorithms. The evaluation of the model with result conducted are discussed in Section 4. The paper with a summary and discussion explained in Section 5.

2. RELATED WORK

In this era of data-driven decision-making and technological advancements, healthcare fraud

detection has gained prominence. 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. Nonetheless, the healthcare industry is increasingly turning to advanced technologies, particularly machine learning, and they are supposed to enhance its fraud detection capabilities with analyze via two cascaded checks for the detection of insurance claim related frauds [4].

Machine learning models offer improvement when it comes to the accuracy and efficiency of fraud detection in healthcare. Data analysis and machine learning are best ways used to address many problems regarding any automated system. To overcome such types of problems there is a model that can flag these suspicious fraudulent claims for the insurance companies to help them out in saving money and time and helping them become more efficient in reacting to these fraudulent claims [2]. 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. Machine learning enables continuous learning from new data, and it allows the novel representation learning approach, which translates diagnosis and procedure codes into Mixtures of Clinical Codes (MCC) [3]. Lots of 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 which are available easily.

Detecting healthcare insurance fraud using AI involves leveraging advanced technologies to identify irregularities, patterns, or anomalies in healthcare claims data. The domain of healthcare insurance fraud detection using AI encompasses various techniques and approaches. Data analytics and machine learning: Utilizing machine learning algorithms to analyze historical claims data to identify patterns associated with fraudulent activities, employing clustering algorithms to group similar claims and detect anomalies or outliers in each cluster and defines a list of features of a best-in-class healthcare insurance fraud detection application solution [14]. Developing predictive models to assess the likelihood of a claim being fraudulent based on various features such as patient history, provider behavior, and claim details.

3. THE PROPOSED MODEL

In the proposed system the techniques can be used for detecting the healthcare insurance fraud are as follows:

3.1 Decision tree Algorithm

A decision tree is a popular machine learning algorithm used for both classification and regression tasks. It works by recursively partitioning the data into subsets based on the most significant attribute at each level. A decision tree is a flowchart-like where each internal node denotes the feature, branches denote the rules and the leaf nodes denote the result of the algorithm. It is a versatile algorithm, which is used for both classification and regression problems. It is one of the very powerful algorithms. Here's an overview of the decision tree algorithm:

1) Selecting the root node: Choose the attribute that provides the best split. This is often determined using metrics such as Information Gain or Gini Impurity for classification tasks and Mean squared Error for regression tasks.

2) Splitting data: Divide the dataset into subsets based on the chosen attribute in the root node. Each subset corresponds to a specific value or range of the selected attribute.

3) Repeating the process: For each subset, repeat the process recursively. Choose the best attribute for splitting the subset, creating child nodes.

4) Stopping criteria: Define stopping criteria to decide when to stop the tree-growing process. This may include a maximum depth, a minimum number of samples per leaf node, or other criteria to prevent overfitting.

5) Leaf node assignments: Assign a class label (for classification) or a numerical value (for regression) to each leaf node based on the majority class or mean value of the samples in that node.

3.2 Random Forest Algorithm

Random Forest is an ensemble learning algorithm that belongs to the family of tree-based methods. It is widely used for both classification and regression tasks in machine learning. The Random Forest algorithm is an ensemble of decision trees, where each tree is trained on a random subset of the data and makes a prediction. The final prediction is then determined by aggregating the predictions of all the individual trees. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting. The Working process can be explained in the below steps and diagram:

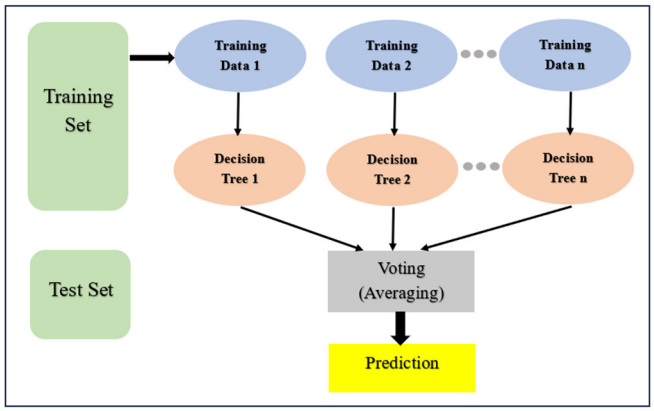

FIGURE 1 Working of Random Forest algorithm

Step-1: Select random K data points from the training set.
Step-2: Build the decision trees associated with the selected data points (Subsets).
Step-3: Choose the number N for decision trees that you want to build.
Step-4: Repeat Step 1 & 2.
Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase. The below diagram explains the working of the Random Forest algorithm, see figure

3.3 Support Vector Machine Algorithm

The Support Vector Machine (SVM) algorithm involves several steps to train a model for classification tasks. Here's a breakdown of the key steps:

1. Data Collection and Preprocessing: Gather a dataset with labeled examples where each example belongs to one of two classes (binary classification) or multiple classes (multiclass classification). Preprocess the data by handling missing values, scaling features, and encoding categorical variables if necessary.

2. Feature selection or Extraction: Identify relevant features that are informative for distinguishing between different classes. Optionally, perform feature extraction techniques to reduce dimensionality or transform the data into a more suitable representation.

3. Model Selection: Choose the appropriate SVM variant based on the problem requirements: Linear SVM: Suitable for linearly separable data. Kernel SVM: Applicable for nonlinear classification tasks, utilizing kernel functions like polynomial, radial basis function (RBF), or sigmoid to map data into higher-dimensional space.

4. Parameter Selection: Tune hyperparameters such as the regularization parameter (C), kernel type, and kernel parameters (e.g., gamma for RBF kernel) through cross-validation or grid search to optimize model performance.

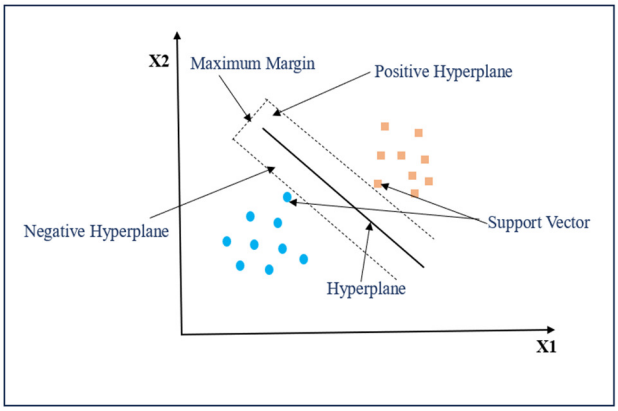

FIGURE 2 Working of Support Vector Machine algorithm

5. Model Training: Given the labeled training data and selected parameters, the SVM algorithm finds the optimal hyperplane that maximizes the margin between different classes. For linear SVM, this involves solving a convex optimization problem using techniques like gradient descent or quadratic

programming. For kernel SVM, the algorithm computes the decision function in the higher-dimensional feature space without explicitly mapping the data points.

6. Model Evaluation: Assess the performance of the trained SVM model using evaluation metrics such as accuracy, precision, recall, F1-score, or area under the ROC curve (AUC). Validate the model's generalization ability using techniques like cross-validation on a hold-out validation set or through techniques like k-fold cross-validation.

7. Model Testing and Deployment: Evaluate the SVM model on unseen test data to estimate its performance in real-world scenarios. If the model meets the desired performance criteria, deploy it to make predictions on new, unseen data. Monitor the model's performance over time and update it as needed to adapt to changes in the data distribution or problem requirements.

These steps outline the typical workflow for training and deploying an SVM model for classification tasks. Adjustments and fine-tuning may be required based on the specific characteristics of the dataset and the problem domain.

Following figure shows the working of the Support Vector machine algorithm with support vectors.

3.4 System Architecture

Designing a healthcare insurance fraud detection system involves various components and stages. Below is a simplified architecture diagram (Figure 3) for a healthcare insurance fraud detection system using AI. The system architecture has following stages:

a) Input Dataset

Raw data from insurance claims, including patient information, procedures, diagnosis codes, provider details, and billing information. Additional sources such as public health records, social determinants of health, and other relevant datasets to enrich the analysis.

b) Data Preprocessing

Initial processing steps, such as resizing or normalization, to prepare the image. Create relevant features from raw data, such as aggregating claim amounts, calculating frequencies, or extracting patterns. Scale numerical features to a standard range for model training.

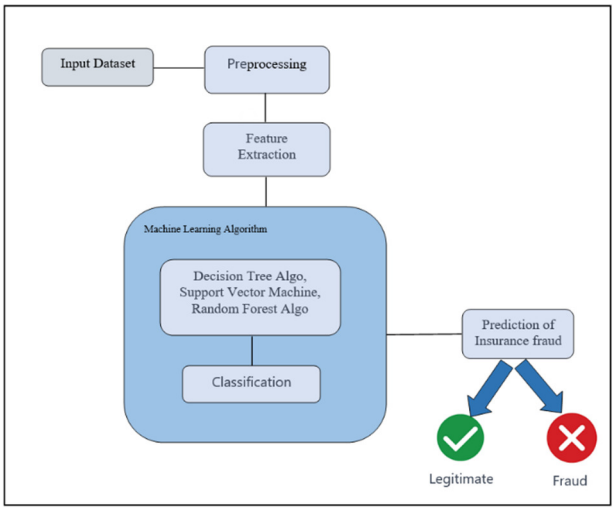

FIGURE 3 System Architecture

c) Fraud Detection model or Machine Learning Algorithm

Ensemble models like Decision Tree Algo, Support Vector Machine, Random Forest Algorithm for fraud detection. Analyze feature importance to understand the contribution of different variables to fraud detection. Implement a rule-based system to flag suspicious claims based on predefined rules and thresholds. Integrate machine learning models into the decision engine for dynamic and adaptive fraud detection.

d) Prediction Workflow

Notify fraud investigators and relevant stakeholders about potentially fraudulent activities. Create dashboards and reports to provide insights into fraud detection performance, trends, and key metrics. Track and manage fraud investigation cases through a centralized system. Incorporate feedback from investigators to improve model performance and refine rules.

4 RESULTS AND DISCUSSIONS

In this discussion, the user dataset is input to the system, then it will preprocess the dataset such as cleaning, then features are extracted and then Machine learning algorithm will apply to classify the particular dataset and according to that the system will predict if the user is fraud or legitimate. In this

way the system will detect fraud user or data and help to healthcare insurance company, bank as well as the customer.

5 CONCLUSION AND FUTURE WORK

In this paper, we present holistic approach that combines advanced technology, human expertise, and a commitment to ethical and transparent practices. By leveraging the strengths of both AI and human investigators, the healthcare industry can build resilient systems that effectively identify and mitigate fraudulent activities, ultimately contributing to a more secure and sustainable healthcare ecosystem. The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set is higher accuracy score will be found out. This application can help to find the Prediction of Insurance Claim Fraud.

Healthcare insurance fraud has been depleting medical finances, but conventional, manual fraud detection methods require time and effort. Machine and deep learning methods offer a practical, cost-effective solution that detects healthcare insurance fraud effectively. We built a model that aimed to detect fraud in healthcare claims. This model successfully used random forest, decision tree, and SVM to detect fraud with optimal accuracy and good evaluation metrics. Furthermore, each model revealed the significant features causing the outcome. Policy type, education, and age were identified as the most significant features that contributed to fraudulent acts. However, further studies with larger datasets, more variables, and various healthcare providers are advised for better generalization.

6 REFERNCES

[1] Shamitha S.K, V Ilango, " A time-efficient model for detecting fraudulent health insurance claims using Artificial neural networks", IEEE ICSCAN, ISBN 978-1-7281-62027, 2020.

[2] Arif Ismail Alrais, "Fraudulent Insurance Claims Detection Using Machine Learning" Information Technology and Systems. ICITS 2019, Springer, Cham, vol 918, 2022.

[3] Md Enamul Haque, Mehmet Engin Tozal, "Identifying Health Insurance Claim Frauds Using Mixture of Clinical Concepts" Procedia Comput. Sci., vol. 64, pp. 713–720, 2021.

[4] Matloob, Shoab Ahmed Khan, Habib Ur Rehman, " Sequence Mining and Prediction based Healthcare Fraud Detection Methodology Irum", Perspect Health Inf Manag;6:1g. pp. 1–24, 2020.

[5] Y.Lakshmi Shreejha, Ch. Purnima Gayatri, G. Bhavana, N. Teja Nayana, P.Sandhya Krishna, “Medicare Fraud Detection Using Supervised Learning”, Springer London, vol. 8, 2020.

[6] A novel optimized GA–Elman neural network algorithm. Neural Comput & Applic 31, 449–459 2019.

[7] X. Jiang, S. Pan, G. Long, F. Xiong, J. Jiang and C. Zhang, "Cost Sensitive Parallel Learning Framework for Insurance Intelligence Operation," in IEEE Transactions on Industrial Electronics, vol. 66, no. 12, pp. 9713-9723, Dec. 2019.

[8] G. G. Sundarkumar and V. Ravi, "A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance," Eng. Appl. Artif. Intell., vol. 37, pp. 368–377, 2015.

[9] Benchaji I., Douzi S., El Ouahidi B. (2019) Using Genetic Algorithm to Improve Classification of Imbalanced Datasets for Credit Card Fraud Detection, AIT2S 2018, Springer Lecture Notes in Networks and Systems, vol 66, 2018.

[10] C. Phua, D. Alahakoon, and V. Lee, "Minority report in fraud detection," ACM SIGKDD Explore. News., vol. 6, no. 1, p. 50, 2004.

[11] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE : Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[12] Shamitha S K, V Ilango "A Hybrid Technique for Health Insurance Fraud Detection on Highly Imbalanced Dataset", International Journal of Innovative Technology and Exploring Engineering (IJITEE), pp. 3498– 3501, vol 8, no 11,2019.

[13] Yufeng Kou, Chang-Tien Lu, S. Sirwongwattana and Yo-Ping Huang, "Survey of fraud detection techniques," IEEE International Conference on Networking, Sensing and Control, 2004, Taipei, Taiwan, pp. 749-754 Vol.2, 2004.

[14] Jasmine Kaur Gill and Shaun Aghili, "Health Insurance Fraud Detection", International Journal of Engineering & Technology, [S.l.], v. 7, n. 4, p. 5862-5868, Apr. 2020.