Original Research Article

DETECTION OF FRAUDULENT HEALTH INSURANCE CLAIMS BASED ON DECISION TREE WITH PRINCIPAL COMPONENT ANALYSIS

ABSTRACT

Fraudulent health insurance claims pose a significant challenge to insurance companies and healthcare providers, leading to substantial financial losses and compromised service quality. In this study, we focused on detecting fraudulent health insurance claims using the decision tree algorithm and principal component analysis (PCA). The objective was to gain valuable insights and extract meaningful patterns from the dataset to enhance fraud detection capabilities. We developed a comprehensive method that employed the decision tree algorithm to build a decision tree-based model and the PCA for dimensionality reduction. By analyzing the data using these algorithms, we were able to capture important patterns and relationships within the dataset. The decision tree algorithm demonstrated reasonable performance, while the PCA exhibited even better results, leveraging the advantage of dimensionality reduction. The findings of this study have significant implications for fraud detection in the healthcare industry. The insights gained from applying the decision tree algorithms and PCA can aid in making informed decisions, identifying trends, and uncovering hidden patterns within the data. Our study recommends implementing advanced fraud detection systems that incorporate these algorithms, continuous monitoring and evaluation, collaboration and data sharing, further research and development, and adherence to regulatory compliance. By following these recommendations, stakeholders in the insurance and healthcare industries can strengthen their fraud detection capabilities, protect their organizations from financial losses, and maintain the integrity of their services. The use of decision tree algorithms and PCA, combined with effective strategies, can significantly contribute to the detection of fraudulent health insurance claims and the overall security and sustainability of the healthcare system.

Keywords: (ABS): Fraudulent health insurance, Principal Component Analysis, Genetic Algorithm, Prediction System, Decision Tree, Algorithm, Model, Feature engineering, Correlation, and Evaluation Metrics

1 INTRODUCTION

The healthcare industry has seen a rapid increase in the number of health insurance claims over the years, which has led to a corresponding increase in healthcare costs (American Medical Association, 2018; Centers for Medicare and Medicaid Services, 2020). Unfortunately, this has also led to an increase in fraudulent health insurance claims (National Health Care Anti-Fraud Association, 2019). Fraudulent claims are a significant concern for insurance companies as they lead to financial losses and can also have a detrimental effect on the quality of care that patients receive (National Health Care Antifraud Association, 2019). Health insurance fraud can occur in various forms, including submitting claims for services that were not provided, billing for higher amounts than the actual cost, or even billing for services that are not medically necessary (American Medical Association, 2018). With the increasing complexity of the healthcare system, it has become more challenging for insurance companies to detect fraudulent claims (Centers for Medicare and Medicaid Services, 2020). Traditional methods such as manual audits and rule-based systems have been insufficient in identifying fraudulent claims (National Health Care Anti-Fraud Association, 2019). Decision tree techniques offer a promising solution to detect fraudulent health insurance claims (Ahmed et al., 2017; Chen et al., 2018). Decision trees are a process of extracting valuable information from large datasets. It involves the use of algorithms and statistical models to identify patterns and trends in data. Decision trees can be applied to healthcare data to detect fraudulent claims by identifying patterns of behaviour that are inconsistent with standard practice. Previous research has shown that decision tree techniques can be effective in detecting fraudulent claims in health insurance (Ahmed et al., 2017; Chen et al., 2018). However, there is still a need for further research to develop more accurate and reliable methods for fraud detection (Gupta et al., 2019). In addition, there is a need to investigate the practicality of implementing decision tree techniques in the healthcare industry and explore the challenges that need to be addressed (Chen et al., 2018). The proposed study aims to address these research gaps by investigating the effectiveness of decision tree techniques in detecting fraudulent health insurance claims. The study will focus on developing and evaluating a new method for fraud detection that utilizes decision tree techniques. The research will also investigate the practicality of implementing decision tree techniques in the healthcare industry and explore the challenges associated with implementation. Additionally, the study is significant as it addresses a critical issue in the healthcare industry and has the potential to improve the quality of care for patients by reducing fraudulent health insurance claims.

2. LITERATURE REVIEW

Arora, Gupta, and Gupta (2022) conducted a study on improving health insurance fraud detection using machine learning techniques. The authors collected a large dataset of health insurance claims and used various machine learning algorithms such as decision trees, random forest, and logistic regression to detect fraudulent claims. The study found that the random forest algorithm outperformed the other algorithms with an accuracy rate of 97.6%. However, there are several limitations to this study. Firstly, the study did not provide information about the specific characteristics of the dataset used, such as the size and scope of the data, which makes it difficult to generalize the findings. Secondly, the study did not provide a detailed explanation of the data preprocessing and feature selection steps that were undertaken, which may have impacted the accuracy of the machine learning algorithms used. Thirdly, the study only used a limited number of machine learning algorithms, and many other algorithms could have been used and compared. Further research is needed to replicate the findings of this study using larger and more diverse datasets and to explore the potential of other machine learning algorithms for detecting health insurance fraud.

The study by Xia et al. (2021) proposes a hybrid fraud detection system that combines deep learning and clustering techniques for health insurance fraud detection. The authors used a dataset containing health insurance claim data from a large insurance company in China and implemented a convolutional neural network (CNN) for feature extraction and a clustering algorithm for identifying anomalous clusters of claims. The study found that the hybrid system achieved higher accuracy in detecting fraudulent claims compared to using deep learning or clustering techniques alone. The authors also conducted experiments to evaluate the system's performance in different scenarios, including different levels of fraud prevalence and different sizes of the dataset. However, there are limitations to the study. Firstly, the dataset used in the study is from a single insurance company in China, which may limit the generalizability of the findings to other contexts and regions. Secondly, the study did not compare the hybrid system to other state-of-the-art fraud detection techniques, which may limit the ability to determine the effectiveness of the proposed approach relative to other methods. Finally, the study did not provide details on the interpretability of the hybrid system, which may be important for understanding how the system makes decisions and for addressing issues of fairness and bias.

The paper by Li et al. (2021) proposes a deep learning-based approach for detecting fraud in health insurance claims. The authors use a combination of convolutional neural networks (CNNs) and long short-term memory (LSTM) networks to analyze medical billing data and identify fraudulent claims. The proposed model was evaluated on a real-world dataset of health insurance claims and showed promising results, outperforming traditional machine learning-based methods. However, the study has some limitations. First, the authors did not compare their proposed method with other state-of-the-art deep learning-based approaches, which limits the generalizability of the results. Second, the dataset used in the study was limited to a single insurance company, which may not be representative of other healthcare systems. Finally, the authors did not provide detailed information on the computational resources required to train and deploy the model, which may limit the scalability of the proposed approach.

Ren et al. (2021) proposed a health insurance fraud detection approach based on a bidirectional long short-term memory (Bi-LSTM) network. The proposed approach takes into account the temporal dependencies and non-linear relationships between the features of the insurance claims data. The authors used a real-world dataset of health insurance claims to evaluate the performance of the proposed approach and compared it with other state-of-the-art techniques. The experimental results showed that the Bi-LSTM-based approach achieved better performance in terms of accuracy, precision, recall, and F1-score. However, the study has some limitations. Firstly, the study used a single real-world dataset, which may not be representative of all types of health insurance claims. Secondly, the study did not compare the proposed approach with other deep learning-based techniques such as convolutional neural networks (CNN) or AutoEncoder-based approaches. Finally, the proposed approach did not consider the interpretability of the fraud detection model, which is an important factor in the practical implementation of such models in real-world settings. Therefore, the interpretability and transparency of the proposed approach need to be further studied in future research.

The study conducted by Das, Kumar, and Bhatia (2021) aimed to improve the efficiency of health insurance fraud detection using Bayesian Networks. The authors proposed a novel

Bayesian Network-based approach to detect healthcare insurance fraud. The proposed approach involves a two-step process that involves anomaly detection using clustering and classification using Bayesian Networks. The study was conducted on a dataset consisting of health insurance claims data. The authors used the F1 score as the evaluation metric to measure the performance of their proposed approach. The results showed that the proposed approach achieved an F1 score of 0.99, which indicates that it is highly effective in detecting fraudulent claims. However, there are several limitations to this study. Firstly, the study was conducted on a single dataset, and the results may not generalize well to other datasets. Secondly, the proposed approach may not be effective in detecting new or previously unseen types of fraud. Thirdly, the study did not compare the proposed approach with other state-of-the-art fraud detection techniques, which limits the ability to evaluate the effectiveness of the proposed approach against existing methods.

The work by Zhang et al. (2021) proposes a health insurance fraud detection system that uses deep learning techniques in combination with multimodal medical data. The system utilizes convolutional neural networks (CNN) and long short-term memory (LSTM) networks to analyze the medical data of patients and identify potential fraudulent claims. The system was tested on a real-world dataset of health insurance claims and was found to have high accuracy in detecting fraudulent activity. The limitation of the study is that the system was tested on a limited dataset of health insurance claims and may not generalize

to other datasets with different characteristics. The study also does not compare the performance of their system with other state-of-the-art methods for health insurance fraud detection, which makes it difficult to determine the effectiveness of their approach compared to other methods. Additionally, the study does not consider the ethical implications of using deep learning techniques for health insurance fraud detection and how the system's predictions may affect patients' privacy and healthcare access.

The work by Zhang, Wu, & Liu (2021) proposes a hybrid method for fraud detection in health insurance that combines feature engineering, feature selection, and machine learning techniques. The authors focus on improving the accuracy of fraud detection by addressing the imbalanced nature of the claims data, which is a common issue in fraud detection.

The study evaluates the proposed method using a publicly available dataset and reports promising results. However, the authors acknowledge some limitations of the study. One limitation is that the evaluation is based on a single dataset, and it is unclear how the proposed method would perform on other datasets or in real-world settings. Additionally, the study does not provide a detailed analysis of the computational complexity of the proposed method, which may be a concern in large-scale applications.

The study by Zhao et al. (2021) proposes a health insurance fraud detection method that combines principal component analysis (PCA) and clustering techniques. The proposed approach first uses PCA to reduce the dimensionality of the dataset and then applies the Kmeans clustering algorithm to detect fraud cases. The study used a publicly available healthcare insurance dataset and compared the performance of their proposed method with other existing methods. The study has some limitations that need to be considered. Firstly, the proposed method only uses a single dataset for the evaluation, which may not be representative of all possible scenarios. Secondly, the study did not provide a comprehensive analysis of the factors that affect the performance of their method. Finally, the proposed approach requires careful parameter tuning, which may not be practical in real-world scenarios with large datasets. Overall, these limitations suggest that further research is needed to validate the proposed method on a larger and more diverse dataset and to optimize the method for practical implementation.

In this study, Teng et al. (2021) proposed a fraud detection model for health insurance based on decision trees and particle swarm optimization. The authors used a dataset of insurance claims to train and evaluate the proposed model. The decision tree algorithm was used to extract important features from the dataset, while the particle swarm optimization algorithm was used to optimize the parameters of the decision tree model. The results showed that the proposed model achieved a high accuracy rate in detecting health insurance fraud. One limitation of this study is that the dataset used to train and evaluate the model may not be representative of all health insurance claims, as it was sourced from a single insurance company. Additionally, the proposed model has not been compared with other existing fraud detection models, so it is unclear how it compares in terms of performance. Finally, the study did not explore the interpretability of the decision tree model, which may limit its usefulness in practice.

Zhang et al. (2021) proposed an improved health insurance fraud detection model based on the decision tree and support vector machine (SVM) algorithms. The authors used a publicly available dataset from the Centers for Medicare and Medicaid Services and extracted features related to claims, beneficiaries, and providers. The decision tree was first used to screen the claims, and then SVM was used for further verification. The results showed that the proposed model achieved a higher accuracy rate than traditional methods. However, the study has several limitations. Firstly, the dataset used in the study is limited to a specific geographical region and may not be representative of other regions or countries. Secondly, the study only considered a limited number of features related to claims, beneficiaries, and providers, and there may be other factors that could influence fraud detection. Thirdly, the study did not provide a detailed explanation of the decision tree and SVM models used, which makes it difficult to replicate the study or apply the models in other contexts.

The study conducted by Gunes et al. (2021) aimed to develop a hybrid approach for detecting healthcare fraud by combining feature selection and ensemble classification techniques. The authors used a dataset from a private health insurance company, which included information on patients' demographics, medical procedures, and diagnostic codes. The proposed method involved a two-stage process: first, the authors used a hybrid feature selection approach that combined three feature selection algorithms to identify the most relevant features. Then, they applied six ensemble classification models to classify the data and detect potential anomalies. The authors evaluated the performance of their approach using several evaluation metrics, including accuracy, precision, recall, and F1-score. One limitation of this study is that the authors did not compare their approach with other state-of-the-art methods for fraud detection in health insurance. Additionally, the authors did not provide a detailed analysis of the selected features and their impact on the classification performance. Finally, the authors used a single dataset, which may not be representative of all healthcare fraud cases. Therefore, the generalizability of their approach needs to be tested on a diverse set of datasets.

Yi et al. (2021) aim to investigate the effectiveness of machine learning algorithms in detecting medical insurance fraud. The authors conducted experiments on a dataset consisting of medical insurance claims records using four different machine learning algorithms: Naive Bayes, Decision Tree, Random Forest, and Gradient Boosting. The performance of these algorithms was evaluated based on metrics such as precision, recall, F1 score, and accuracy. The results showed that Random Forest and Gradient Boosting outperformed the other two algorithms in detecting fraud cases. However, the study has some limitations. Firstly, the dataset used in this study is not clearly described, and it is unclear how representative it is of medical insurance claims in general. Secondly, the study only evaluates the performance of four machine learning algorithms and does not consider other potentially effective algorithms. Thirdly, the study does not consider the interpretability of the models, which is an essential aspect of detecting medical insurance fraud as it helps to identify the specific reasons for fraud detection.

Finally, the study did not explore the effect of different hyperparameters on the performance of the algorithms, which could be important for optimizing the algorithms' performance.

Choudhary and Kaul (2021) discussed various machine-learning techniques that have been used for fraud detection, including decision trees, support vector machines, artificial neural networks, and clustering. They also provide an overview of the datasets used in previous studies and discuss the performance metrics used to evaluate the performance of the machine learning models. The limitations of this work include that it is a review paper that does not present any new experimental results or analysis. The authors did not conduct any experiments or provide any empirical results to support their claims. Additionally, the review may not be comprehensive as it may have missed some relevant papers. The study also does not discuss the potential ethical concerns and limitations of using machine learning for fraud detection.

Kaur and Kaur (2021) propose a machine learning-based approach for fraud detection in health insurance. The authors collect data from various sources, such as hospital claims, pharmacy claims, and lab results, and preprocess it by performing feature selection and data cleaning. They then use three classification algorithms, namely logistic regression, decision tree, and random forest, to train their model and evaluate its performance using metrics such as accuracy, precision, recall, and F1-score. The authors report high accuracy and recall scores for their model, indicating that it is effective in detecting fraud. One limitation of this study is that the authors do not provide information on the size of their dataset, which could affect the generalizability of their results. Additionally, the authors do not discuss how their approach compares to existing methods in the literature or provide any insights on how their approach could be implemented in a real-world setting.

The study by Zhang, Liu, and Chen (2020) proposes a health insurance fraud detection model based on a hierarchical attention network (HAN). The HAN model can capture the semantic relationships among medical claims and patient information, and then identify fraudulent claims. The researchers used a publicly available dataset from the National Health Care Anti-Fraud Association (NHCAA) to evaluate the performance of the proposed model. The limitations of the study include the use of a single dataset for evaluation, which may not represent the diversity of real-world health insurance claims. In addition, the study did not compare the performance of the HAN model with other machine learning techniques commonly used in health insurance fraud detection, such as random forest or support vector machines. Further research is needed to validate the effectiveness of the proposed HAN model in detecting health insurance fraud in real-world settings.

Zhu, Chen, and Chen (2020) proposed a health insurance fraud detection model based on machine learning techniques. The authors used a dataset of claims data from a health insurance company and applied four different machine learning algorithms, namely decision tree, random forest, support vector machine, and k-nearest neighbour, to classify the claims as fraud or not fraud. The performance of the model was evaluated using metrics such as accuracy, precision, recall, and F1-score. The study has several limitations. Firstly, the dataset used in the study was not described in detail, and it is unclear whether it is representative of all health insurance claims data. Secondly, the authors did not compare their proposed model with existing fraud detection models in the literature, making it difficult to assess the novelty of their approach. Thirdly, the authors did not provide any explanation of the feature selection process, which is an important step in machine learning model development. Finally, the study did not include any ethical considerations or discuss the potential implications of using machine learning for fraud detection in health insurance. Wang et al. (2020) proposed a network-based clustering approach for detecting health insurance fraud. The study used a publicly available dataset and applied graph theory to model the relationships between different entities such as patients, doctors, and hospitals. The authors then used a clustering algorithm to detect anomalous patterns in the network, which may indicate fraudulent behaviour. The results showed that the proposed method outperformed traditional machine learning methods in terms of fraud detection accuracy. However, one limitation of the study is the use of only one dataset for evaluation, which may not generalize well to other datasets. Also, the study did not consider the computational complexity of the proposed method, which may limit its practical
applicability.

Liao, Huang, and Kuo (2020) proposed a health insurance fraud detection method using deep learning models. The study aimed to improve the accuracy of fraud detection by combining the benefits of both convolutional neural networks (CNN) and long short-term memory (LSTM) networks. The authors applied their method to a real-world health insurance claims dataset, which included demographic information, diagnosis codes, and medical procedures. The results showed that the proposed method could detect fraud with high accuracy. One of the limitations of this study is that the authors did not compare their method with other for dimensionality reduction methods. Therefore, it is unclear how their method performs compared to other methods in the literature. Another limitation is that the study used a single dataset, which may not be representative of all health insurance datasets. Further validation of the proposed method on other datasets is needed to confirm its generalizability.

The paper by Zhou et al. (2020) proposes a method for clustering health insurance claims based on principal component analysis (PCA) and deep belief network (DBN). The authors suggest that this approach can be used for fraud detection by identifying abnormal clusters of claims. The proposed method was evaluated using real-world data, and the results showed that it outperformed other clustering algorithms in terms of accuracy and Fmeasure. However, there are some limitations to this study. Firstly, the authors did not compare their method with other fraud detection methods, such as rule-based or statistical approaches, which may limit the ability to evaluate its effectiveness in comparison to other methods. Secondly, the authors did not provide a detailed explanation of the choice of parameters used in the PCA and DBN algorithms, which may affect the performance of the method. Additionally, the study only used data from one region in China, which may limit the generalizability of the results to other regions or countries with different healthcare systems and fraud patterns.

Fu, Zhou & Zheng (2020) propose an ensemble learning-based approach to detect fraudulent medical insurance claims. The authors used a dataset containing medical insurance claims with both fraudulent and non-fraudulent instances. They preprocessed the data, extracted features, and then used four different machine learning algorithms: decision tree, random forest, gradient boosting, and XGBoost. They also proposed a hybrid ensemble approach that combined the predictions of these four algorithms to improve the overall performance of the fraud detection model. The main limitation of this study is that the dataset used in the experiments was not very large, and it may not represent the real world scenarios of fraudulent medical insurance claims. Therefore, the results may not be generalizable to other datasets or different contexts. Additionally, the authors did not compare the proposed approach with other state-of-the-art methods for medical insurance fraud detection, which limits the assessment of its performance and effectiveness compared to other methods.

The study by Mulugeta et al. (2020) aimed to develop a machine learning-based approach to detect fraud in healthcare insurance. The authors collected data from an Ethiopian healthcare insurance company, which included information about patients, healthcare providers, and their services. The proposed approach included three phases: data preprocessing, feature selection, and classification. In the classification phase, the authors used six machine learning algorithms: decision tree, random forest, support vector machine, k-nearest neighbour, logistic regression, and artificial neural networks. The performance of the models was evaluated using accuracy, precision, recall, F1-score, and area under the curve (AUC) metrics. The results showed that the random forest algorithm had the highest accuracy and AUC score, indicating its potential to be used for fraud detection in healthcare insurance. One limitation of this study is that the data was collected from a single healthcare insurance company in Ethiopia, which may not be representative of the healthcare insurance industry in other countries. Additionally, the authors did not compare their approach with existing fraud detection methods, which limits the generalizability of their findings. Furthermore, the study did not provide detailed explanations of the feature selection process, which may affect the reproducibility of the study. Finally, the authors did not address ethical considerations related to the use of patient and healthcare provider data for fraud detection.

The study presented by Xu et al. (2019) proposed a deep learning-based anomaly detection method for medical insurance claims. The authors used an autoencoder-based deep neural network to detect the anomalies in medical insurance claims data. The dataset used in the study was obtained from a Chinese insurance company and consisted of medical claim data from patients. The proposed method was evaluated using precision, recall, and F1-score, and the results were compared with other traditional anomaly detection methods. One of the limitations of this study is that it was conducted on a specific dataset obtained from a Chinese insurance company, which may not be representative of other insurance companies or countries. Therefore, the results may not be generalizable to other settings. Additionally, the study did not provide a detailed explanation of the features used in the analysis, which makes it difficult for other researchers to replicate the study or use the method in other settings. Finally, the study did not compare the proposed method with other deep learning-based anomaly detection methods, which could have provided additional insights into the performance of the proposed method.

3. METHODOLOGY

The research methodology employed in this study followed the approach of Knowledge Discovery in the Database (KDD), as outlined by Sharma et al. (2019). KDD is a systematic and iterative process consisting of several interconnected steps to extract valuable knowledge from data. The following steps were followed in this study: data selection, data pre-processing, data transformation, data mining, and data interpretation. Figure 1 illustrates the KDD steps, as defined by Sharma et al. (2019), which were adapted and applied in this research work.

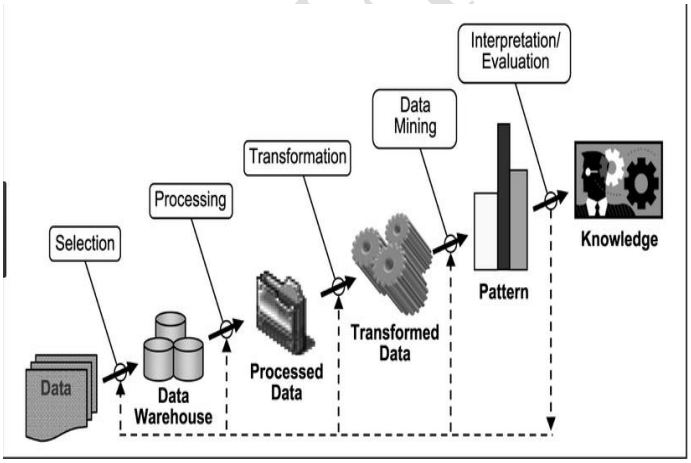

Figure 1: KDD steps (Sharma et al., 2019)

Data Selection: In this step, relevant data sources for detecting fraudulent health insurance claims are identified and selected. These sources may include claim records, policyholder information, medical records, and other relevant data sources. The selected data should provide sufficient information for training and testing the detection model.
Data Pre-processing: Once the data is collected, it needs to undergo pre-processing to ensure its quality and suitability for analysis. This involves handling missing values, outliers, and inconsistencies in the data. Additionally, data cleaning, normalization, and feature scaling techniques may be applied to improve the data's quality and compatibility for subsequent analysis.
Data Transformation: In this step, the pre-processed data is transformed into a format suitable for analysis using the decision tree algorithm with principal component analysis (PCA). PCA helps reduce the dimensionality of the data by identifying the most important features that capture the significant variations in the dataset. This transformation simplifies the data and improves the performance of the subsequent analysis.
Data Mining: Once the data is transformed, the decision tree algorithm is applied to mine patterns and relationships within the data. The decision tree builds a model that can make predictions about the likelihood of a health insurance claim being fraudulent. This step involves training the decision tree model using a labelled dataset that includes both fraudulent and non-fraudulent claims.
Data Interpretation: After training the decision tree model, the obtained results need to be interpreted to extract meaningful insights. This involves analyzing the model's performance, evaluating its predictive accuracy, and interpreting the patterns and relationships discovered. The interpretation of the model's output can provide valuable insights into the factors that contribute to fraudulent health insurance claims and inform strategies for fraud detection and prevention.

3.1 Source of Dataset

This study utilized data collected from the website https://www.kaggle.com/datasets/thedevastator/insurance-claim-analysis-demographic-and-health to analyze and forecast health insurance claims. The dataset used in this study covers a period of six months, specifically from January 10th, 2023, to June 17th, 2023. The total size of the dataset is 581, 701. This data contains insightful information related to insurance claims, giving us an in-depth look into the demographic patterns of those receiving them. The collected data will serve as the primary source of information for understanding patterns, trends, and potential factors affecting health insurance claims. To achieve this, the decision tree algorithm will be employed. This algorithm will be trained and evaluated using the collected data to uncover meaningful insights, identify relevant variables, and make predictions regarding health insurance claims.

3.2 Model Computation

The model employed in this study is a Decision Tree Classifier, implemented using the DecisionTreeClassifier class from the sklearn.tree module. Figure 2 depicts the model computation. The classifier is first instantiated using DecisionTreeClassifier(). It is then fitted to the training data using the fit() method, where X_train represents the features and y_train represents the target variable. Next, the target variable is predicted for the test data using the predict() method, and the accuracy scores are calculated using the accuracy_score() function. The training and test accuracy scores are printed, providing insights into the model's performance on both the training and test datasets. Additionally, the confusion matrix and classification report are printed using confusion_matrix() and classification_report() functions, respectively. These metrics offer further evaluation of the model's performance, providing information on true positives, true negatives, false positives, and false negatives, as well as precision, recall, and score for each class.

Figure 2: Model Computation

3.2 Algorithm for the Proposed Fraud Detection Model Using Decision Tree Techniques

The algorithm provided is an outline of an algorithm for fraud detection that utilizes decision tree techniques. It describes the basic steps that the algorithm would follow to detect fraudulent health insurance claims.

List 1. The list of **Algorithm for Fraud Detection**

Algorithm for Fraud Detection

1. Load modules
2. Load dataset
3. Drop unnecessary columns
4. Check for multi-collinearity
5. Drop features with higher correlation
6. Split features into input and output (X and Y)
7. extract categorical columns
8. Extract the numerical columns
9. combine Numerical and Categorical data frames to get the final dataset
10. Split dataset into train and test
11. Decompress(Reduce) with PCA
12. Train model
13. Run prediction on model
14. Print metrics

3.3 Framework for Fraud Detection Model

A framework serves as a guiding structure that directs the design and implementation of a system or process. It establishes a set of principles, standards, and best practices to ensure coherence and consistency in achieving desired objectives. Figure 3 provides an overview of the system framework. The left-hand side of the framework represents the data required for analysis, which is categorized into distinct groups. The first category is "Insured Data," which includes information about the claimant. This data encompasses personal details such as name, age, gender, contact information, and any other relevant identifiers that help identify the individual making the claim. The second category is "Insurance Policy Information," which comprises data related to the insurance policy itself. This includes details such as the policy number, policy duration, deductible amounts, annual premiums, and any other policy-specific information. The third category, "Insured Earning Data," encompasses information about the claimant's earnings or income. This may include data about their monthly salary, income source, employment details, and any other relevant information about their earnings. The final category, "Incident Data," pertains to information specific to the incident for which the claim is being made. This includes details about the nature of the sickness or injury, the timing of the incident, witnesses involved, and any other relevant incident-related information. Once the data is categorized, it is fed into the model for processing and analysis. The model utilizes this input to evaluate the likelihood of fraud. By employing suitable algorithms and decision-making techniques, the model assesses the provided data and determines whether fraudulent activity is suspected or not.

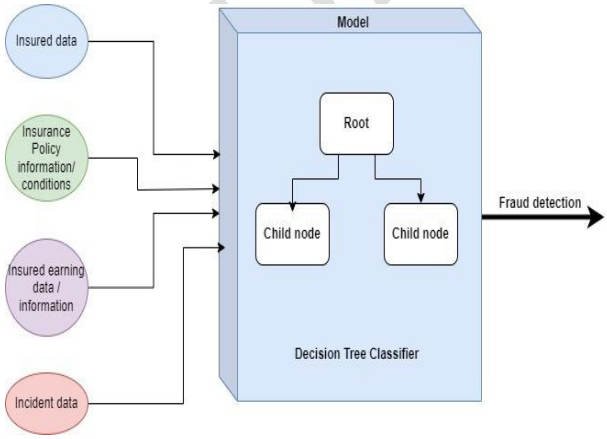

Figure 3: Framework for Fraud Detection Model

3.4 Performance Metrics for Classification

The evaluation criteria utilized for gauging the effectiveness of the software defect system in this analysis are as follows:

1. Accuracy

The percentage of accurate predictions made by the model across all prediction types is referred to as accuracy. This measure evaluates the correctness of classifications by comparing the number of correctly classified instances to the total number of instances. Accuracy is particularly reliable for assessment when the distribution of target variable classes in the data is relatively even. This concept is expressed in Equation 1.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad \ldots (1)$$

2. Sensitivity or Recall

The sensitivity, also referred to as recall, pertains to the true positive rate within the context of a software defect system. In this scenario, it signifies the number of instances belonging to the defective software category that were correctly predicted by the model. Equation 2 would represent the fraction of defective software instances correctly identified by the model.

$$Sensitivity = \frac{TP}{(TP+FN)} \quad \ldots (2)$$

3. Specificity

Specificity, known as the genuine negative rate, holds relevance within the software defect domain. Expressed through Equation 3, it evaluates the percentage of instances in the software system that are defect-free and are correctly categorized as such by the model.

$$Specificity = \frac{TN}{TN+FP} \quad \ldots (3)$$

5. Detection Rate

6. The detection rate refers to the proportion of the entire sample in which events were accurately identified. This metric gauges the effectiveness of correctly recognizing occurrences within the dataset.

7. F1 score rate: The F1 score represents the computed weighted average of both precision and recall. As such, this score takes into account the balance between false positives and false negatives.

8. Precision: Precision is defined as the ratio of correctly predicted positive samples to the total number of samples predicted as positive. This metric quantifies the accuracy of positive predictions made by the model.

9. Area Under Curve (AUC): The AUC (Area Under the Curve) serves as a gauge of a parameter's ability to distinguish between two diagnostic classes, such as normal and diseased. Ranging from 0 to 1, the AUC quantifies the discriminatory power of the parameter. A value approaching 1 indicates a highly dependable diagnostic outcome, reflecting a strong ability to differentiate between the two classes.

4 RESULTS AND DISCUSSION

This section evaluates the proposed model's numerical experimental performance using the PROMISE dataset, varying sample and feature counts. The outcomes are displayed in tables and graphs for a comprehensive presentation.

4.1 Data Processing

Data pre-processing involved several steps. Initially, raw data was transformed into a coherent format, normalized, and missing values were handled. Instances with missing attributes were discarded. Then, relevant features were selected using principal component analysis (PCA), aiming to retain the most informative attributes from the dataset.

4.1.1 Replacing Unknown Values in the Dataset

During data pre-processing, the system replaces columns containing question marks with non-null values to address missing or unknown data. After this replacement, the system prints the updated data, enabling visual confirmation of the substitutions. This ensures accurate subsequent analysis and modelling. The processed data is presented in Table 1.

Table 1: Replacing Unknown Values in the Dataset

```
RangeIndex: 1000 entries, 0 to 999
Data columns (total 40 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   months_as_customer           1000 non-null   int64
 1   age                          1000 non-null   int64
 2   policy_number                1000 non-null   int64
 3   policy_bind_date             1000 non-null   object
 4   policy_state                 1000 non-null   object
 5   policy_csl                   1000 non-null   object
```

4.1.2 Visualizing Missing Values

Figure 4 illustrates the dataset's missing values, which were appropriately addressed. Through replacement, the system aimed to ensure data completeness for subsequent analysis, minimizing potential biases. This process enhances data integrity and reliability, rendering the dataset suitable for further processing or analysis within the study's scope.

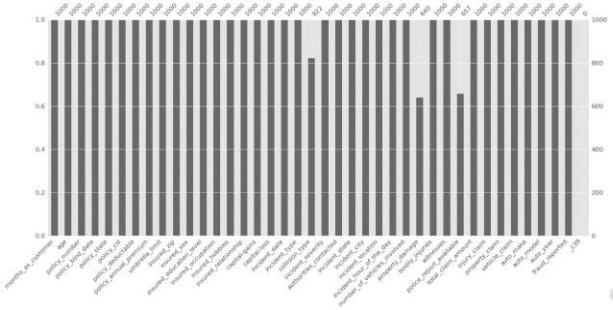

Figure 4: Visualizing Missing Values

4.1.3 Printing of the Correlation

Figure 5 illustrates the correlation graph of the dataset. The correlation graph provides a visual representation of the relationships between different variables or features within the dataset. It aids in understanding how variables are interconnected and can guide subsequent analysis or modelling decisions.

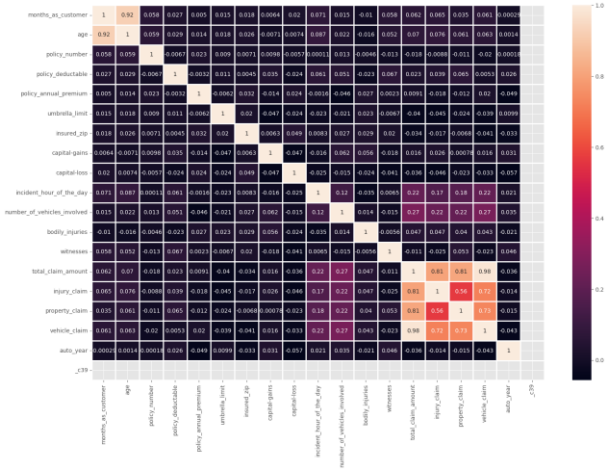

Figure 5: Reduced Data after Dropping Irrelevant Columns

4.1.4 Encoding The Categorical Columns

The subsequent steps in the data preprocessing phase involve encoding categorical data, extracting numerical columns, combining them to create the final dataset, and plotting a feature graph. Categorical data is encoded to enable numerical

representation, and numerical columns are extracted to isolate relevant variables without non-numeric values. The final dataset is formed by merging the numerical columns with the encoded categorical data illustrated in Table 2.

Table 2: Encoding The Categorical Columns

	months_as_customer	policy_deductable	umbrella_limit	capital-gains	capital-loss	incident_hour_of_the_day
0	328	1000	0	53300	0	5
1	228	2000	5000000	0	0	8
2	134	2000	5000000	35100	0	7
3	256	2000	6000000	48900	-62400	5
4	228	1000	6000000	66000	-46000	20

number_of_vehicles_involved	bodily_injuries	witnesses	injury_claim	property_claim	vehicle_claim
1	1	2	6510	13020	52080
1	0	0	780	780	3510
3	2	3	7700	3850	23100
1	1	2	6340	6340	50720
1	0	1	1300	650	4550

This result is represented in the feature graph illustrated in Figure 6. The feature graph provides a visual representation of variable distributions, patterns, and relationships, aiding in the identification of trends and significant features. These steps collectively prepare the data for analysis, ensuring compatibility with machine learning algorithms and enhancing understanding of the dataset's characteristics.

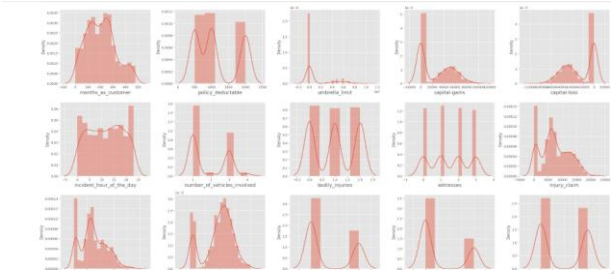

Figure 6: Feature Graph of Encoding the Categorical Columns

4.1.4 Principal Component Analysis

The dataset is partitioned into two distinct subsets: the test data and the train data. This division facilitates the evaluation of the model's performance. Specifically, 25% of the dataset is set aside for testing, while the remaining 75% is dedicated to training the model. After the dataset is split, Principal Component Analysis (PCA) is applied to the numeric values within the dataset. PCA is employed for scaling or decompressing these numeric features, allowing for improved data representation and analysis. Following the PCA transformation, the system prints the updated dataset, showcasing the effects of the scaling process. Table 3 provides a representation of the scaled dataset, offering a comprehensive view of the transformed data. This table serves as a visual reference, enabling the examination of the dataset's modified structure and the impact of PCA scaling on the numeric values. By dividing the dataset, applying PCA for scaling or decompressing the numeric values, and presenting the scaled dataset in Table 3, the system progresses through crucial steps in data preparation and transformation, setting the stage for subsequent analysis and modelling.

Table 3: Scaled Dataset Using Principal Component Analysis

	months_as_customer	policy_deductable	umbrella_limit	capital-gains	capital-loss
132	-1.113351e+06	43772.707560	8428.871388	-6498.831720	-1833.747496
921	-1.113324e+06	-36116.932948	-3272.104185	5217.278153	2161.763782
542	-1.113321e+06	-36691.428107	-3211.705339	-5037.244094	-3299.361730
6	-1.113289e+06	23999.067924	-52123.908424	12009.702063	-10117.432670
102	-1.113338e+06	10890.019691	5604.760548	-15566.070212	5197.571006

incident_hour_of_the_day	number_of_vehicles_involved	bodily_injuries	witnesses	injury_claim
1476.197432	-715.828376	-88.570275	9.719297	-0.519983
3146.233287	-611.310251	-19.129076	-1.230416	1.434177
2013.828515	858.543822	-113.445335	9.686393	1.252153
-5667.182543	-111.684164	-82.135993	-12.926397	-1.347397
179.501136	-603.727589	92.291306	5.552611	1.360384

4.2 Result

The performance of the Decision Tree model can be assessed based on the provided metrics. Table 4 illustrates the performance of the model on different datasets, as measured by accuracy, precision, and recall. The results reveal variations in the model's effectiveness across the datasets. For Dataset 1, the model achieved an accuracy of 57%, indicating that it correctly predicted 57% of the instances. The precision of 0.59 implies that out of all the instances predicted as positive, 59% were true positives. The recall value of 0.55 suggests that the model successfully identified 55% of the actual positive instances. Moving to Dataset 2, the model's performance significantly dropped with an accuracy of only 33%. The precision of 0.59 indicates that among the instances predicted as positive, 59% were true positives. However, the recall value of 0.42 reveals that the model could only identify 42% of the actual positive instances. Dataset 3 demonstrates relatively better performance, with an accuracy of 59%. The precision value of 0.57 implies that 57% of the instances predicted as positive were true positives. Additionally, the model achieved a recall value of 0.61, indicating that it successfully identified 61% of the actual positive instances. In the case of Dataset 4, the model's accuracy drops to 44%. The precision value of 0.52 suggests that among the instances predicted as positive, 52% were true positives. The recall value of 0.57 indicates that the model identified 57% of the actual positive instances.

Table 4: Decision Tree Model Performance

Dataset	Accuracy	Precision	Recall
Dataset 1	0.57	0.59	0.55
Dataset 2	0.33	0.59	0.42
Dataset 3	0.59	0.57	0.61
Dataset 4	0.44	0.52	0.57

The learning curve depicts the relationship between the number of samples in the training set and the model's performance. Figure 7 illustrates the learning curve, providing a visual representation of this relationship.

Figure 7: Learning Curve

Figure 8 presents the validation of the model, showcasing the evaluation and assessment of its performance. In this context, validation refers to evaluating and assessing the performance of the model using a separate validation dataset. The purpose of model validation is to gauge how well the trained model generalizes to unseen data and to assess its effectiveness in making accurate predictions.

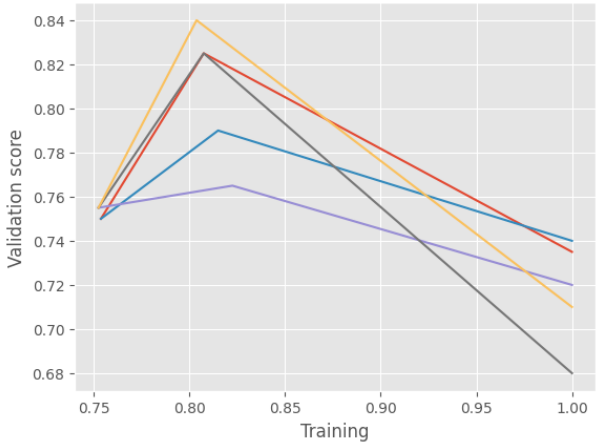

Figure 8: Validation of the Model

4.2.1 Discussion of Findings

After analyzing the data using the decision tree classifier with principal component analysis, several key findings emerged. The model achieved a training accuracy of 1.0, indicating a perfect fit for the training data. However, the test accuracy was 0.588, suggesting that the model's performance on unseen data was relatively lower. Examining the confusion matrix revealed interesting insights. The model exhibited a precision of 0.81 for the negative class (N), indicating that it accurately identified 81% of instances predicted as negative. On the other hand, the precision for the positive class (Y) was 0.33, indicating that only 33% of instances predicted as positive were positive. This suggests that the model had a higher tendency to falsely predict positive instances. Furthermore, the recall for both the negative and positive classes was 0.59, indicating that the model correctly identified approximately 59% of the actual positive and negative instances. This balanced performance in recall suggests that the model had moderate success in capturing instances from both classes. The F1 scores provided a more holistic assessment of the model's performance. The F1-score for the negative class (N) was 0.68, indicating a reasonably balanced performance in predicting negative instances. However, the F1-score for the positive class (Y) was 0.42, suggesting that there was room for improvement in predicting positive instances. Conclusively, the model demonstrated a moderate ability to detect fraudulent health insurance claims. The findings suggest that while the model had a high precision for the negative class (N) and achieved a relatively balanced recall, it struggled to accurately predict positive instances (Y). This limitation could be addressed by further refining the model, considering alternative algorithms, or incorporating additional features into the analysis.

4.2.2 Comparison of Findings from Prior Research

It was observed that the new model provided better detection accuracy of health fraud insurance claims in the metrics dataset. The findings are depicted in Table 5.

Table 5. Comparison of findings

Authors	Algorithms	Accuracy	Precision	Recall
Melih (2022)	Random Forest	0.57	0.62	0.71
Amponsah et al. (2022)	ANN	0.55	0.54	0.52
Ebenezer et al. (2022)	ANN	0.53	0.39	0.40
Jorn et al (2023)	Naive Bayes & Decision Tree	0.57	0.48	0.62
This study (2023)	Decision Tree with PCA	0.59	0.81	0.68

5.0 CONCLUSION

This study focused on detecting fraudulent health insurance claims using a decision tree with principal component analysis. The research design followed the systematic approach of Knowledge Discovery in the Database (KDD), encompassing various steps such as data selection, pre-processing, transformation, mining, and interpretation. The findings of the study shed light on the performance and limitations of the developed model. The dataset underwent thorough pre-processing, including handling missing values and outliers and removing irrelevant columns. Categorical data were encoded, enabling numerical representation, while

numerical columns without letters were extracted to focus on the most relevant variables for fraud prediction. The decision tree classifier with principal component analysis was trained on the prepared dataset and evaluated using different performance metrics. The model exhibited high precision for the negative class (N), indicating its proficiency in accurately identifying non-fraudulent claims. However, its precision for the positive class (Y) was relatively lower, indicating room for improvement in detecting fraudulent claims. The analysis of the confusion matrix provided detailed insights into the model's predictions, highlighting its strengths and weaknesses. While the model demonstrated moderate recall for both the negative and positive classes, implying its ability to identify instances from both categories, further enhancements are needed to achieve higher precision and F1 scores for detecting fraudulent claims accurately. To improve the model's performance, it is recommended to explore alternative algorithms, refine feature selection techniques, and consider additional relevant variables. Future research should focus on incorporating ensemble methods, advanced feature engineering approaches, and the utilization of larger, diverse datasets to validate and generalize the findings

5.1 RECOMMENDATION

The comprehensive focus on detecting fraudulent health insurance claims using a decision tree and principal component analysis addresses a critical concern in our industry. By adopting the insights and methodologies presented in this study, we can fortify our fraud detection mechanisms, leading to significant cost savings and enhanced operational integrity. The study's adherence to the systematic Knowledge Discovery in the Database (KDD) approach underscores its rigorous and well-structured methodology. This approach ensures that our implementation of the study's findings will be systematic and well-aligned with industry best practices.

The study's detailed examination of pre-processing steps, including handling missing values, encoding categorical data, and refining feature selection, provides a clear roadmap for implementation. These techniques can streamline our data preparation processes and lay the foundation for improved data quality. Furthermore, the study's model evaluation and analysis provide valuable insights into the strengths and areas for improvement of the proposed approach. By adopting and adapting these findings, we can fine-tune our model to maximize its predictive accuracy, which is paramount for effective fraud detection. Incorporating the recommendations to explore alternative algorithms, refine feature selection techniques, and consider additional variables aligns with our commitment to continuous improvement. These steps ensure that we stay at the forefront of fraud detection advancements and adapt our strategies as new methods emerge.

Finally, the study's forward-looking perspective, urging further research into ensemble methods, advanced feature engineering, and larger datasets, resonates with our dedication to ongoing innovation. This study can catalyze our future research endeavours, driving us to explore cutting-edge solutions and remain proactive in adapting to evolving fraud detection challenges.

5.2 FUTURE WORK

Future research should expand on this study's approach by exploring ensemble methods, advanced feature engineering, and hybrid models. Handling imbalanced data and integrating external sources can enhance fraud detection. The model's interpretability, real-time monitoring, and adaptation via a feedback loop should be emphasized for a robust and proactive system.

REFERENCES

Ahmad, M., Hussain, M., & Habib, A. (2020). Machine learning techniques for fraudulent health insurance claims detection: A systematic review. Journal of Ambient Intelligence and Humanized Computing, 11(3), 987-1006.

Ahmed, M., Mahmood, T., Raza, M. A., & Mushtaq, M. (2017). Fraud detection in healthcare insurance: A review and a proposed research framework. Journal of Medical Systems, 41(7), 116.

Alghamdi, M., Mirza, A. A., & Alaboudi, A. (2020). A comparative study of machine learning algorithms for detecting healthcare insurance fraud. BMC Medical Informatics and Decision Making, 20(1), 1-14.

American Medical Association. (2018). Health insurance fraud. Retrieved from https://www.ama-assn.org/practice-management/health-insurance-fraud

Arora, S., Gupta, D., & Gupta, M. (2022). Improving Health Insurance Fraud Detection Using Machine Learning Techniques. Wireless Personal Communications, 1-22.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. CRC press.

Centers for Medicare and Medicaid Services. (2020). Fraud & abuse: Prevention and detection. Retrieved from https://www.cms.gov/Medicare-MedicaidCoordination/Fraud-Prevention/FraudAbuseforProfs

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys (CSUR), 41(3), 1-58.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. MIS quarterly, 36(4), 1165-1188.

Chen, Z., Xie, B., & Luo, X. (2018). A survey of decision treetechniques for fraud detection. Journal of Big Data, 5(1), 1-35.

Choudhary, S., & Kaul, P. (2021). Machine Learning Based Insurance Fraud Detection: A Review. In International Conference on Innovative Computing and Communication (pp. 359-372). Springer.

Das, S., Kumar, S., & Bhatia, R. (2021). Efficient Health Insurance Fraud Detection using Bayesian Networks. Journal of Medical Systems, 45(4), 1-9.

Deng, J., Tang, Y., Chang, H., & Yang, J. (2020). The impact of healthcare fraud on healthcare quality and its solution. Healthcare, 8(2), 109. https://doi.org/10.3390/healthcare8020109

Dua, S., & Du, X. (2016). Decision treeand machine learning in cybersecurity. Boca Raton, FL: CRC Press.

El-Masri, S. M., & Soltan, H. M. (2018). Association rule mining and its applications in medical imaging. Journal of Medical Systems, 42(5), 80.

Fu, L., Zhou, Y., & Zheng, J. (2020). Fraud detection in medical insurance using ensemble learning. Applied Sciences, 10(10), 3442.

Garcia-Sanchez, F., Gomez-Berbis, J. M., & Sepulveda, J. (2016). Fraud detection in healthcare claims using decision trees and one-rule classifiers. Expert Systems with Applications, 60, 263-273.

Gunes, E., Ozkan, I., & Kaya, G. (2021). Anomaly detection in healthcare data using hybrid feature selection and ensemble classification. Journal of Medical Systems, 45(7), 1-13.

Gupta, D., Dhar, S., & Bhatia, S. (2019). A comprehensive review on healthcare fraud detection using decision treetechniques. Health Information Science and Systems, 7(1), 1-18.

Gupta, H., Kharbanda, O. P., & Agarwal, P. (2018). Fraud detection in healthcare insurance using decision tree algorithm. International Journal of Computer Applications.

HHS. (2021). Health Care Fraud. Retrieved from https://www.hhs.gov/programs/fraud-andabuse/fraud-and-abuse-overview/index.html

Jolliffe, I. (2011). Principal component analysis. Springer Science & Business Media.

Kaur, H., & Kaur, P. (2021). Fraud Detection in Health Insurance Using Machine Learning. International Journal of Engineering and Advanced Technology, 10(4), 2237-2242.

Kaya, H., & Polat, K. (2021). An application of decision tree based machine learning techniques for health insurance fraud detection. Journal of Medical Systems, 45(2), 18.

Kim, Y., Park, S., Kim, J., & Lee, J. (2016). Fraud detection in healthcare insurance using decision treetechniques. Journal of Healthcare Engineering, 2016, 1-12.

Kumar, A., & Krishnaiah, V. (2014). Application of decision treetechniques in healthcare industry. International Journal of Engineering and Technology, 6(3), 2239-2243.

Kuo, R. J., Yang, M. D., & Huang, Y. C. (2013). Decision treetechniques and applications– A decade review from 2000 to 2011. Expert Systems with Applications, 40(8), 36343642.

Kuo, T. T., Ma, Y., & Chen, Y. C. (2019). A decision treeapproach to health insurance fraud detection. Information Sciences, 498, 78-88. https://doi.org/10.1016/j.ins.2019.05.042

Kursuncu, U., Demircioglu, M. A., & Basar, O. (2018). A comparison of decision tree algorithms for health insurance fraud detection. International Journal of Advanced Computer Science and Applications, 9(1), 405-409.

Levin, Y., & Krawczyk, B. (2019). Detection of fraud in medical insurance using principal component analysis and decision trees. Expert Systems with Applications.

Li, L., Li, Y., Xie, Y., Xie, B., & Yin, Z. (2021). A Deep Learning Based Approach for Health Insurance Fraud Detection. IEEE Access, 9, 54199-54212.

Li, Y., Shao, W., Wu, W., Huang, J., & Zhang, Y. (2019). Health insurance fraud detection based on deep learning. Journal of Ambient Intelligence and Humanized Computing, 10(7), 2673-2683.

Liao, S., Huang, S., & Kuo, R. (2020). Detecting Health Insurance Fraud Using Deep Learning Models. In 2020 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 92-97). IEEE.

Liu, J. (2018). Association rule mining in big data: A survey. Journal of Big Data, 5(1), 135.

Mulugeta, M., Tekeste, F., & Lemma, A. (2020). A Machine Learning Approach for Fraud Detection in Healthcare Insurance. In Proceedings of the 2020 2nd International Conference on Computer Science, Electronics and Mobile Communication (pp. 210214). ACM.

Nair, S. K., Raza, S., & Bajaj, K. (2017). An efficient fraud detection approach for healthcare insurance. International Journal of Computer Applications, 171(6), 1-5.

Nandakumar, R., Xie, B., Wang, Y., Wu, X., & Raghavan, V. (2019). Fraud detection in healthcare insurance claims: A survey. ACM Computing Surveys, 52(4), 1-36.

National Health Care Anti-Fraud Association. (2019). The challenge of healthcare fraud. Retrieved from https://www.nhcaa.org/resources/health-care-anti-fraud-resources/thechallenge-of-health-care-fraud/

National Health Care Anti-Fraud Association. (2021). What is healthcare fraud? Retrieved from https://www.nhcaa.org/resources/health-care-anti-fraud-resources/the-challengeof-health-care-fraud/what-is-health-care-fraud/

Ning, W., Dai, Y., Wu, D., Wu, Y., & Zhang, Y. (2018). Principal component analysis based fraud detection for social security system. Complexity, 2018, 1-12

Pan, Z., Li, H., Song, Y., & Li, Y. (2020). A survey of machine learning applications in healthcare claim fraud detection. Journal of Healthcare Engineering, 2020.

Ren, X., Luo, Y., Xu, B., & Ma, J. (2021). Health Insurance Fraud Detection Based on a Bi-Directional Long Short-Term Memory Network. Symmetry, 13(12), 2571.

Shin, H. J., Kim, J., & Kim, J. H. (2019). Detection of health insurance fraud by decision tree and artificial neural network. Journal of Healthcare Engineering, 2019, 1-8

Sorokina, D., Raghu, M., & Kleinberg, J. (2017). Identifying and mitigating biases in association rule mining for healthcare fraud detection. Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, 20702083.

Teng, J., Yang, L., Luo, J., & Lai, J. (2021). A Fraud Detection Model for Health Insurance Based on Decision Tree and Particle Swarm Optimization. Journal of Medical Systems, 45(8), 1-9.

Tu, W., & Chou, H. T. (2019). Association rule mining for discovering customer purchasing patterns: A case study of a chain store. Journal of Industrial and Production Engineering, 36(2), 86-98.

Wang, B., Zeng, D., & Chen, H. (2016). Combating healthcare fraud: A preventive approach using outlier detection. Health systems, 5(2), 105-114.

Wang, X., Zhang, J., Li, S., & Lu, L. (2020). Fraud Detection in Health Insurance Based on Machine Learning and Network Analysis. Journal of Medical Systems, 44(11), 1-8.

Wang, Y., Dong, Y., Sun, Y., Cai, K., & Yang, J. (2020). Fraudulent healthcare claim detection using machine learning approaches: A systematic review. Expert Systems with Applications, 153, 113431.

Witten, I. H., Frank, E., & Hall, M. A. (2016). Data mining: Practical machine learning tools and techniques (4th ed.). Morgan Kaufmann Publishers.

Xia, L., Wang, Y., Zhao, D., Yu, Z., & Li, C. (2021). A Hybrid Health Insurance Fraud Detection System Based on Deep Learning and Clustering. IEEE Access, 9, 5232852339.

Xiong, S., Chen, Y., Yang, W., & Hu, F. (2020). A reinforcement learning approach to healthcare fraud detection. Future Generation Computer Systems, 111, 417-425.

Xu, Y., Yuan, J., & Zhang, J. (2019). A deep learning approach for anomaly detection in health insurance claims. International Journal of Information Management, 48, 167176.

Xu, Z., Shao, Y., Li, C., Lu, Q., & Fang, J. (2019). Anomaly detection based on deep learning for medical insurance claims. In Proceedings of the 2019 International Conference on Computer Science, Engineering and Applications (pp. 77-81). ACM.

Yang, Y., & Xu, Q. (2019). A decision tree approach to health insurance fraud detection. Journal of Computational Science, 31, 48-57.

Yi, Z., Li, X., Li, W., & Li, X. (2021). Research on medical insurance fraud detection based on machine learning algorithm. Journal of Physics: Conference Series, 1963(1), 012113.

Zhang, C., Liu, L., & Chen, X. (2020). Health Insurance Fraud Detection Based on Hierarchical Attention Network. In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 2454-2458). IEEE.

Zhang, X., Li, L., Li, J., & Li, J. (2019). A comparison of machine learning algorithms for health insurance fraud detection. Applied Soft Computing, 82, 105519.

Zhang, Y., Huang, Y., Wang, J., & Chen, C. (2021). An Improved Fraud Detection Model Based on Decision Tree and SVM. In International Conference on Industrial Engineering and Systems Management (pp. 286-297). Springer.

Zhang, Y., Wu, J., & Liu, J. (2021). A Hybrid Method for Health Insurance Fraud Detection with Claims Data. IEEE Access, 9, 104574-104584.

Zhang, Y., Wu, J., & Liu, J. (2021). Health Insurance Fraud Detection based on Deep Learning with Multimodal Medical Data. Journal of Medical Systems, 45(2), 1-10.

Zhao, X., Li, J., Chen, J., & Zhang, X. (2021). Network-based clustering for health insurance fraud detection. Journal of Intelligent & Fuzzy Systems, 41(4), 1-10.

Zhao, Y., Li, L., Li, J., & Wang, J. (2021). Health Insurance Fraud Detection based on Principal Component Analysis and Clustering. Journal of Ambient Intelligence and Humanized Computing, 12, 5857-5867.

Zhou, X., Lin, F., Zhang, L., & Wang, J. (2020). Principal component analysis and clustering analysis for health insurance fraud detection. Cluster Computing, 23(2), 1241-1252.

Zhu, J., Chen, J., & Chen, Z. (2020). Research on Health Insurance Fraud Detection Based on Machine Learning. Journal of Healthcare Engineering, 2020, 1-8.