risks

MDPI

Article

Fraud Detection in Healthcare Insurance Claims Using Machine Learning

Eman Nabrawi [1,2] and Abdullah Alanazi [1,2,*]

[1] Health Informatics Department, King Saud Ibn Abdulaziz University for Health Sciences, P.O. Box 3660, Riyadh 11481, Saudi Arabia
[2] King Abdullah International Medical Research Center, Riyadh 14611, Saudi Arabia
* Correspondence: abdullahgcc@gmail.com or anaziabdul@ksau-hs.edu.sa; Tel.: +966-12195453

Abstract: Healthcare fraud is intentionally submitting false claims or producing misinterpretation of facts to obtain entitlement payments. Thus, it wastes healthcare financial resources and increases healthcare costs. Subsequently, fraud poses a substantial financial challenge. Therefore, supervised machine and deep learning analytics such as random forest, logistic regression, and artificial neural networks are successfully used to detect healthcare insurance fraud. This study aims to develop a health model that automatically detects fraud from health insurance claims in Saudi Arabia. The model indicates the greatest contributing factor to fraud with optimal accuracy. The labeled imbalanced dataset used three supervised deep and machine learning methods. The dataset was obtained from three healthcare providers in Saudi Arabia. The applied models were random forest, logistic regression, and artificial neural networks. The SMOT technique was used to balance the dataset. Boruta object feature selection was applied to exclude insignificant features. Validation metrics were accuracy, precision, recall, specificity, F1 score, and area under the curve (AUC). Random forest classifiers indicated policy type, education, and age as the most significant features with an accuracy of 98.21%, 98.08% precision, 100% recall, an F1 score of 99.03%, specificity of 80%, and an AUC of 90.00%. Logistic regression resulted in an accuracy of 80.36%, 97.62% precision, 80.39% recall, an F1 score of 88.17%, specificity of 80%, and an AUC of 80.20%. ANN revealed an accuracy of 94.64%, 98.00% precision, 96.08% recall, an F1 score of 97.03%, a specificity of 80%, and an AUC of 88.04%. This predictive analytics study applied three successful models, each of which yielded acceptable accuracy and validation metrics; however, further research on a larger dataset is advised.

Keywords: fraud; insurance claims; artificial neural networks (ANN); logistic regression (LR); random forest (RF); Saudi Arabia

Citation: Nabrawi, Eman, and Abdullah Alanazi. 2023. Fraud Detection in Healthcare Insurance Claims Using Machine Learning. *Risks* 11: 160. https://doi.org/10.3390/risks11090160

Academic Editor: Mogens Steffensen

Received: 15 June 2023
Revised: 4 August 2023
Accepted: 29 August 2023
Published: 5 September 2023

1. Introduction

Fraudulent use of health insurance strains available funds and raises the cost of healthcare (Chen et al. 2020). According to the NHACC, it is defined as the "deception or intentional misrepresentation that the person or entity makes knowing that the misrepresentation could result in an unauthorized benefit for the person, entity, or another part" (NHCAA 2018). Healthcare insurance fraud poses a substantial financial challenge, as the US Department of Justice and the Department of Health and human services has reported USD 2.6 lost in 2019 for fraud recoveries (Mackey et al. 2020). Manufacturers, hospitals, pharmacies, healthcare providers, distributors, and payers—the parties most impacted by healthcare fraud—are just a few of the sectors of the healthcare system that are affected (Mackey et al. 2020). Healthcare fraud includes billing for services undelivered or unnecessary, misrepresentation, or willful omission that are essential in determining benefits to be paid, rebilling, readmission, upcoding, unbundling, kickbacks practicing, and unjustified distribution of healthcare services and medications (Nicholas et al. 2020). Therefore, the effects of deception go beyond the monetary harm done to equity and include

injury and the potential jeopardization of patients' safety. For instance, the research found that up to 46% of healthcare practitioners who are not covered by insurance programs endanger patients' safety (Nicholas et al. 2020).

Saudi Arabia's health system is complex in structure. However, in line with the Saudi Vision 2030, the system will undergo significant changes from a national health model to a more liberated model enabling the private sector to engage in a dynamic, active service transitional model with an enhanced quality of service (Alharbi 2018). The formation of the Cooperative Council for Health Insurance (CCHI) in 2005 to supervise the private sector's participation and adherence to regulations, according to scholars, revealed the existence of 26 insurance agencies, more than 5202 healthcare 'services providers, and more than 9 million consumers listed in the CCHI (Alharbi 2018). The CCHI has issued a policy to govern private health insurance conduct dedicated to fraud, waste, and abuse. The policy mandates keeping a detailed record of health claims fraud that contains the type of fraud, how it was committed, weakness in the procedure, a fraudster with their history, and fraud monitoring audit. This record must be available upon authorities' request (CCHI 2021). Moreover, The National Platform for Health and Insurance Exchange Services (NPHIES) in the Council of Health Insurance (CHI) has launched its first phase by completing 8.4 processes within 140 health entities.

NPHIES' current linkage rate is 65% among entities, and its main goal, besides interoperability, is to enhance the efficiency and quality of healthcare services with reduced cost and time. According to Alshaqi, the platform aims to use innovative technology to strengthen digital insurance transformation (SPA 2022). In 2021, the Saudi Arabian health insurance industry had USD 6.5 billion worth. By 2027, the market is anticipated to have grown by a compound annual growth rate (CAGR) of 6.4%, reaching USD 9.4 billion (SAMA 2021). According to Alonazi, there are more than 196 reported fraud cases with approximately 15% claim rejection rates. However, dental and obstetrics–gynecology services have the most reported fraud cases (Alonazi 2020). Saudi Arabia is prepared to utilize machine learning to spot and foretell fraud in health claims. Healthcare insurance fraud detection methods encompass using paper or electronic claims documents to compare and divide data into certain levels to segregate classes. According to scholars, there are various fraud detection methods; for instance, one depends on decoding the data into five stages. The first three stages only investigate claims, providers, and services, while the fourth stage estimates the risk size. The last step includes a decision-making mechanism to declare or not to report the existence of fraud (Thaifur et al. 2021). Therefore, manual fraud detection is inefficient time-consuming, and exhausting work. Alternatively, harnessing information technology advancements such as data mining and machine learning techniques poses an intuitive automatic way in which models can detect and predict fraud. Specifically, data mining techniques are categorized as supervised, semi-supervised, and unsupervised learning. For example, Medicare and Medicaid Services (CMS) centers used unsupervised rule-based data mining to detect anomalies within healthcare claims. CMS data were also used to determine fraud risk by applying supervised machine learning analytics such as decision tree models (Waghade and Karandikar 2018).

Moreover, Suri and Jose have explored fraud detection classification techniques and applied logistic regression (LR) with an accuracy of 92% and random forest (RF) with an accuracy of 88% (Suri and Jose 2019). Additionally, Shipe et al. indicated that the clinical usage of predictive models using LR and retrospective cohort study designs effectively used LR to predict outcomes (Shipe et al. 2019).

To further comprehend the capacity of LR and RF in addressing healthcare fraud detection, it is paramount to investigate the underlying statistical technique for LR and RF. On the one hand, LR is used to comprehend complicated events where the researcher aims to determine predicting factors of dichotomous binary dependent variables. The predicting independent variable can be an interval, a ratio, a nominal, or an ordinal variable (Connelly 2020). On the other hand, RF is also a supervised machine learning technique used primarily for classification and possibly regression. Thus, it builds decision

trees and forecasts the best outcome or solution based on voting (Kumar et al. 2021). However, researchers Bauder and Khoshgoftaar indicated the lack of efficiency when using an unbalanced dataset with an RF when it was used to detect Medicare fraud claims (Bauder and Khoshgoftaar 2018). Sumalatha and Prabha have also used LR to detect healthcare fraud (Sumalatha and Prabha 2019). The proposed model effectively detected fraud from claims with 83.35% accuracy. Furthermore, Patel and his colleagues conducted a study to detect fraud via pattern matching (Patel et al. 2019). The researchers used rule-based mining techniques on actual insurance claims. The analysis was categorized based on anomalies within period claims and models based on disease. Their proposed design has proven efficient in detecting fraud within health insurance claims. Nonetheless, Mayaki and Riveill emphasized the significance of balancing the data and how unbalanced data drastically affect the model accuracy (Mayaki and Riveill 2022). Although the researchers created a model to predict and detect fraud from CMS unbalanced data, the model was biased towards the interest class. However, they have proven that artificial neural networks (ANN) can substantially detect fraud. Moreover, the study revealed that incorporating multiple inputs in neural networks improves outcomes (Mayaki and Riveill 2022). The study by Varmedja and his colleagues included various models to detect fraud; however, the RF algorithm performed the best in accuracy, recall, and precision (Varmedja et al. 2019).

Furthermore, Severino and Peng used various machine learning models to detect fraud in Brazilian real-world insurance claims. Nine models were compared, and the RF algorithm performed the best in accuracy. This model is easier to generalize and apply to future operational risk management tools because it is based on actual data. Furthermore, the model can be used as a probabilistic guide in predicting whether a policy will result in fraud (Severino and Peng 2021). Shimitha and Ilango developed a real-time ANN model that detects fraudulent health insurance claims. ANN is a deep learning technique in which neurons behave similarly to biological neurons. It is composed of several layers of interconnected nodes. The input layer comes first, followed by the hidden classifying layer, and finally by the output layer (Shamitha and Ilango 2020). Within their study aimed at detecting credit card fraud using ANN, Asha and Suresh achieved approximately 100% model accuracy in their supervised ANN model (Asha and Kumar 2021).

This study aims to shed light on the importance of using ML in detecting fraud in health claims. The surging cost of healthcare services may affect providers and payers regarding the affordability of services and financial sustainability. Therefore, there is a real need to employ ML to flag fraudulent cases among the thousands of claims submitted daily. Therefore, this study aims to build a machine learning model that detects fraud based on labeled data of fraud or no fraud. The study objectives are to create supervised machine and deep learning models that analyze labeled data (with fraud or no fraud) to detect healthcare insurance fraud. As a result, future risky potential fraud cases will be identified before or during the fraud incident, allowing for better system surveillance and control. However, our study excludes the following activities in health reimbursement: error resulting through unintentional actions; abuse, through manipulating a rule or guideline or taking advantage in the absence of management; and corruption, resulting through abusing power with the involvement of a third party.

2. Materials and Methods

This study project applied a retrospective cohort study design. The machine learning models used a retrospective dataset collected and verified by Saudi CHI and the insurance payers to predict future outcomes of the probability of fraud occurrence within a specific context within healthcare. The model should highlight areas for improvement to avoid future fraud incidents. The sample size of the collected dataset was obtained from three healthcare providers in Saudi Arabia from January 2022 to May 2022. The dataset included anonymous features for all identifying instances and providers. It did not contain a variable for the commercial insurance companies' payers' names; only the policy types were included.

For data preprocessing and balancing, most cases in the dataset are labeled as fraud; thus, the dataset is highly imbalanced and requires intervention. Therefore, the synthetic minority oversampling technique (SMOTE) was applied. SMOTE is a valuable oversampling algorithm that can help make datasets more representative of their population and enhance the accuracy of classification models and their capability to avoid overfitting the training dataset (Fotouhi et al. 2019). SMOTE is particularly beneficial in fraud detection and medical diagnosis, where accuracy is critical. As a result, the sample size became 396 post-SMOTE applications.

To assess the reliability of the models, the following assessments were used: confusion matrices, which indicate the model's accuracy by statistically distinguishing actual positive and negative instances from false positive and negative classifications. Accuracy is another critical metric that indicates the correctly identified type of the model performed. The precision value shows the number of correct positive predictions to all positive predictions by the model. At the same time, the recall indicates the number of accurate positive predictions to overall positive predictions by distinguishing between classes. Receiver operating characteristic (ROC) is a statistical calculation used for determining the quality of differentiating between two groups in a binary classifier. ROC uses a recall and false positives to indicate models' performance. AUC is used to quantify a model's ability to discriminate between positive and negative outcomes and provides an overall assessment of future model performance. Therefore, the higher the AUC, the better the classifying model, an AUC less than 0.5 indicates suboptimal model performance. In some cases, the accuracy outcome may be a misleading optimist. Therefore, having an F1 score as a safeguard is useful for balancing precision and recall. F1 measures accuracy by considering false positive and false negative predictions. (F1 score = (2 (Precision $\times$ Recall))/(Precision + Recall)).

2.1. Statistical Analysis

Machine learning (ML): In recent years, ML has become an increasingly important tool for health services researchers. The researchers demonstrated how ML solves various health services research problems, such as predicting disease progression, analyzing trends in patient outcomes, and identifying subgroups of patients most likely to benefit from intervention health (Doupe et al. 2019). In comparison, deep learning (DL) is a revolutionary technology that uses neural networks capable of recognizing complicated patterns and processing them quickly and accurately. Deep learning algorithms can perform complex tasks, ranging from image recognition and decision making to natural language processing (Asha and Kumar 2021).

2.2. Statistical Analytical Models

Model One: Random forest (RF) or random decision forests is an ensemble learning technique for classification, regression, and other tasks. This technique works by creating several decision trees at training time and outputting the class, that is, the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests are thriving in combating decision trees' tendency to overfit their training set. One of the main advantages of random forests is the immediate connection between the ensemble of trees and the output it can generate. RF introduces a supplementary layer of randomness to the process to yield even more accurate and precise predictions, which has been demonstrated to be beneficial in Rahman et al.'s study on liver disease prediction using supervised machine learning algorithms (Rahman et al. 2019). Subsequently, the number of trees indicated by the max_depth variable in this research project was set to five, as illustrated in Figure 1: RF.

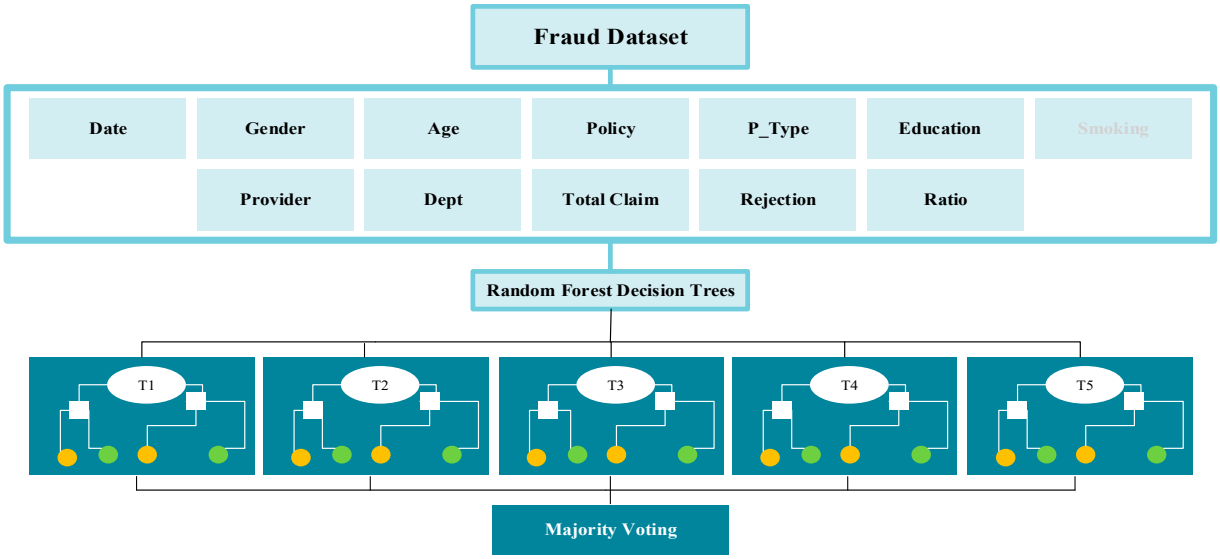

Figure 1. Random forest (RF).

Model Two: Logistic regression (LR) is one of the most important and widely used predictive modeling types used to analyze a dataset and identify relationships between a categorical dependent variable and one or more independent variables. It is employed in various disciplines, including marketing, risk assessment, and medical research, and it can be utilized to make conclusions from multiple datasets. LR is a robust tool used in healthcare fraud detection, allowing organizations to identify, investigate, and prosecute fraudulent activities. It is a predictive modeling technique that estimates the probability of an individual or event belonging to a particular group (Connelly 2020). LR can predict whether a healthcare claim is valid or fraudulent. By utilizing a combination of input variables, such as patient demographics, medical codes, and provider information, logistic regression can accurately identify fraudulent healthcare claims and provide organizations with an effective means to reduce fraudulent activities. Subsequently, the logistic regression model fitting takes the training features dataset alongside the corresponding label dataset to conduct the regression analysis, as illustrated in Figure 2: LR.

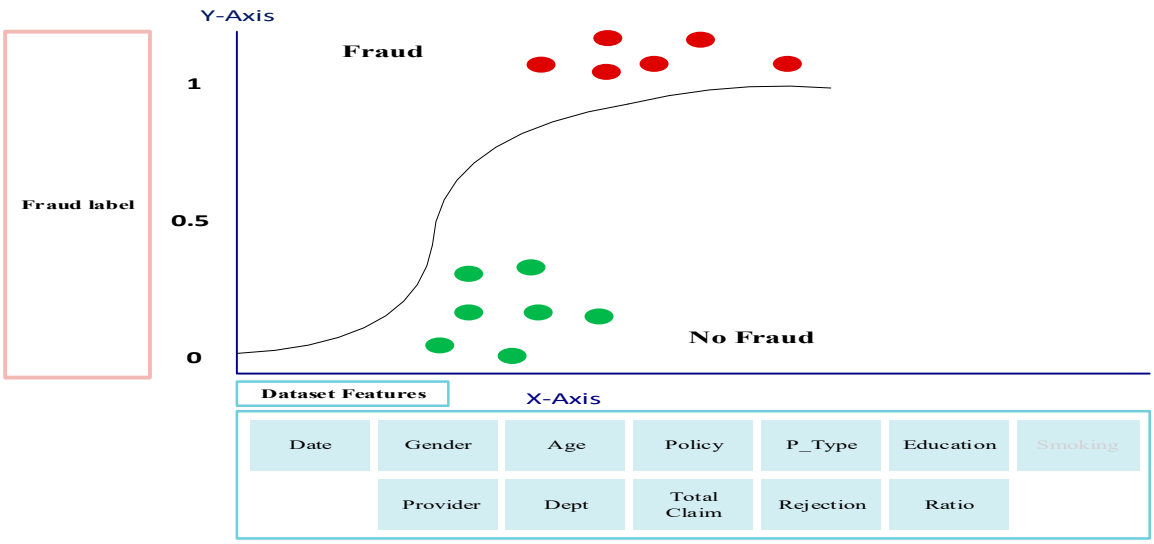

Figure 2. Logistic Regression (LR).

Model Three: Deep learning via artificial neural networks (ANNs) simulates the behavior of biological neurons in human brains. It is a highly efficient computing model that

employs interconnected nodes to detect patterns related to the given inputs. Furthermore, the ANNs use multilayer perceptron (MLP) to distinguish, process, and analyze datasets in conducts that typical systems cannot replicate. ANN is a deep learning method that consists of multiple layers of connected nodes. The first layer is the input layer, where we feed our variables within the dataset; then, the hidden classifying layer that works to solve the problem at hand; and finally, the output layer, which indicates the occurrence of the measured outcome and which in our research project is fraud or no fraud (Shamitha and Ilango 2020). The construction of this ANN model encompasses eleven input nodes as the first layer to match the number of desired features indicated in Figure 3: ANN, using the rectifier activation function. Subsequently, the first fully connected hidden layer encompasses six hidden nodes, using the rectifier activation function, followed by a fully connected node as an output layer, using the sigmoid activation function.

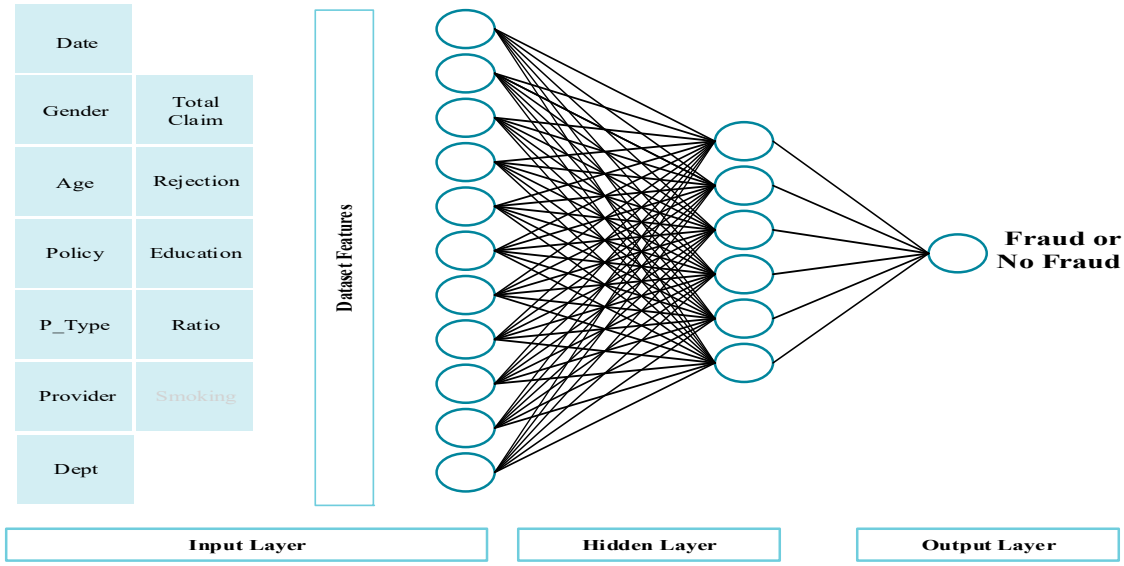

Figure 3. Artificial Neural Networks (ANNs).

Further, the ANN model uses an Adam optimizer for backward weight propagation and a binary cross-entropy for the output compilation (Ahmed and Brifcani 2019). The model fitting takes the feature training set, labels training set, batch size, and epochs. The batch is a variable that sends a selected number of records across the network for training purposes. Further, the batch size is an arbitrary value adjusted frequently until the model converges appropriately. Therefore, the final batch size selected was 16 patches per epoch (Kaur et al. 2020) (see Figure 3). Moreover, the epochs variable addresses how often the selected batches run through the network. Thus, adjusting the model weight through the backpropagation information of the loss function calculates the variation between the true value of the label and the predicted value from the first epoch training and is adjusted for the following epoch. Moreover, the number of epochs is also an arbitrary value and is subjected to trial. In this model, the number of epochs that yielded an optimum result was 50 (Karmiani et al. 2019). We used a trial error approach to determine the correct hyperparameters of the models. The best performance accuracy was achieved with the default parameters.

2.3. Data Science Platform

Anaconda allows different environments of Python to be managed. To be specific, we used scikit-learn for the model implementation and evaluation.

3. Results

Data were collected from three anonymized tertiary healthcare providers in Saudi Arabia. For accurate processing, the dataset was balanced to equalize the fraud and nonfraud cases using the SMOTE technique. Therefore, the sample size was (n = 396). The dataset has 64% positive fraud cases and 36% negative cases. Most healthcare consumer cases were female (93%). The relation between fraud and gender is unreliable due to the percentage of females within the dataset. The consumer age in the dataset ranged from 10 to 98 years old. A total of 22% of cases were under 21 years old, 38% were from 21–40, 27% were from 41–61, 12% were from 61–80, and 1% for older. The age group from 21 to 40 committed more fraud cases. The level of education for the consumers' education is Ph.D. (8%), master's (8%), bachelor's (31%), and high school or lower (53%). Concerning fraud, the less educated group contributed more fraud, according to our dataset, while Ph.D. consumers were the least likely to commit fraud. The insurance policy includes individual policies by the consumer, services policies by employers, and other categories. Concerning fraud, most fraud cases were among employer-type policies. The frequency of the VIP class was 11%, A class was 25%, B class was 26%, and C class was 37%. Most fraud cases were among the VIP class, followed by the C category. The dataset was collected from various hospital departments, and most of the claims were from OBGYNs. However, no conclusion was inferred concerning hospital departments.

The training and validation dataset was split into an 80:20 ratio. However, the imbalance was assured in both 80 (64% positive, and 36% negative cases) and 20 (64% positive, and 34% negative cases).

Random forest (RF) evaluation metrics were promising. Accuracy was 98.21%, which is explained by high true positive and negative cases. However, the model obtained 0% false negatives and 20% false positives. According to the model, policy type was the most crucial feature leading to fraud, followed by the education and age of the claimants. See Table 1.

Table 1. RF features significance.

Feature	Importance
Gender	0.627%
Age	6.25%
Policy Type	20.90%
Education	9.26%
Department	3.02%

Logistic regression (LR) evaluation metrics could have been more promising than RF. However, the measures still indicate vital evaluation metrics. Accuracy was 80.36%, which is explained by high true positive cases of 80% and true negative cases of 80%. Finally, artificial neural networks (ANN) evaluation metrics were robust; accuracy was 94.64%, which is explained by high true positive cases of 98% and true negative cases of 80%.

Regarding feature engineering, we considered temporal feature, as it is important for this binary outcome of fraud or no fraud. The time of day of the transaction and the sequence of the transaction (first, second, third by same person) both added to the complexity of the model without any improvement in its performance. Table 2 describes the performance of different models.

Table 2. Performance of different models.

Model Metrix	Accuracy	AUC	Recall	Precision	Specificity	F1 Score
RF	98.21%	90.00%	100.00%	98.08%	80.00%	99.03%
LR	80.36%	80.20%	80.39%	97.62%	80.00%	88.17%
ANN	94.64%	88.04%	96.08%	98.00%	80.00%	97.03%

4. Discussion

Identifying the most significant feature that leads to fraud will enhance strategic planning and future risk assessment. Therefore, it will eventually be possible to prevent healthcare fraud. However, the highly skilled professionals and specialized services encountered in healthcare and the lack of oversight over the individuals performing the work all contribute to healthcare's uniqueness. Furthermore, the emotional context, particularly in crisis-type healthcare decision making, makes it more challenging to detect fraudulent acts (Byrd et al. 2013). Before examining the study findings, it is worth noting that health insurance fraud is commonly referred to as claimant deception without further investigation into possible root causes and context. Treating the abovementioned issue without addressing the underlying causes may deteriorate the societal problem (Lesch and Baker 2013).

To assess our model performance compared to models in other studies, Sumalatha and Prabha's LR model accuracy was 83.35%, which follows our findings within the same model of 80.36% (Sumalatha and Prabha 2019). Suri and Jose's model explored different fraud detection classification techniques and found the accuracy of LR to be 92%, with an RF accuracy of 88% and a recall of 0.54 (Suri and Jose 2019). However, their study was applied to an open dataset retrieved from Kaggle. Both Varmedja et al. and Severino and Peng's studies concluded that the RF model was the finest performing one based on accuracy (Severino and Peng 2021, Shamitha and Ilango 2020). Therefore, our model supports their conclusion with the highest accuracy score of 98.12%, leading to better generalization and application of future operational risk management tools. Shamitha and Ilango achieved an 85.3% ANN model accuracy, a precision of 97%, and a recall of 73% when detecting fraudulent healthcare insurance claims (Shamitha and Ilango 2020). In comparison, our model optimally achieved 94.64% accuracy with 98% precision and 96% recall.

Among the key findings of our study, the most important contributing factors to health insurance fraud are policy type, education, and claimant age. Like our findings, Villegas-Ortega et al. identified macroenvironmental factors (culture, regulations); mesoenvironmental factors (provider characteristics, management policy, reputation); microenvironmental factors (sex, race, insurance condition, language, treatments, future risk of disease); and other factors that influence health insurance fraud (Villegas-Ortega et al. 2021). Regarding policy type, claimants with VIP (comprehensive-coverage policy) were associated with more fraudulent acts. This could be attributed to what Zhou et al. called in their paper "Claimant negotiating power" (Zhou et al. 2017). As the negotiating power increases, the estimated risk or loss of the claimant is lessened. Further, these claimants tend to be experts in procedures and the legal system. Furthermore, these claimants are usually well-versed in legal procedures. Moreover, some providers may be more lenient in providing extra services for the comprehensive-coverage policy, either because they anticipate that it will cover all services or because of the claimants' social status.

Concerning education, the study found that most of the claimants who participated in the fraudulent activities held a bachelor's degree or lower. This finding is like the study of Timofeyev and his colleagues, as they concluded that individuals with a college degree have a higher tendency to participate in fraudulent acts (Timofeyev and Busalaeva 2019). Age is another important variable in the model. The claimants who were in the age category of 20–45 years old tended to have a high probability of conducting fraud. Age was a determinant factor in fraudulent automobile payments (Doerpinghaus et al. 2008). At the same time, age was positively correlated to fraudulent activities in China, as older claimants tend to experience greater financial risk (Shamitha and Ilango 2020). Similar to our study, Timofeyev et al. discovered that 34 year olds (ranging from 23 to 45) participate in 42% of health insurance fraudulent activities (Timofeyev and Busalaeva 2019).

The increased use of technology in healthcare has facilitated medical identity theft (MIT) fraud, which is defined as the fraudulent theft of health information and personally identifiable information to obtain medical goods and services. The individuals in the Saudi community are inclined to assist others, particularly family members, relatives, and friends. This support could be in the form of sharing medical records to gain access to

health services and resources; this may be more obvious in governmental and not-for-profit hospitals, as it is anticipated that the free national healthcare system covers this. In addition to the harm of this practice to the patient, it encompasses financial loss, especially with the national move towards value-based healthcare and the corporatization of governmental hospitals (Zarour et al. 2021).

This research should form the basis for future fraud detection investigations in Saudi Arabia. It focuses on the most significant factors that contribute to healthcare fraud in the given context. This study provides insights on the health fraudulent acts in Saudi. Using machine learning would help to automate the fraud detection process with minimal human intervention. It would reduce the subjectivity that may be encountered if the process is handled manually by a human. Hence, it enhances the accuracy of fraud detection as well as saves time. However, the quality of prediction is solely dependent on the data acquisition. The model is as good as the used data. Additionally, the model needs to be retrained continuously and updated for any concept drift that may be encountered. Healthcare fraud auditing systems protect payers in the following ways: recognize inconsistencies and rule-breaking behavior; regularly mine data for new fraudulent patterns; and develop rules to detect and prevent improper payments.

Nevertheless, our study has several limitations, including a small sample size, a limited number of features or variables, and the departments represented in the dataset. As mentioned earlier, only 196 fraud cases are reported annually with approximately 15% claim rejection rates (Alonazi 2020). Thus, building a reliable and valid fraud detection model emphasizes the importance of having more data; thus, the larger the size, the more robust the model. Furthermore, it is recommended that all available features be used to cover all aspects of fraud not covered in this research project (Fotouhi et al. 2019). The effect of gender, department, and provider features in fraud detection, for example, may require additional research because females predominated in the study dataset; additionally, the dataset did not cover all departments and was limited to three healthcare providers. Also, it is recommended to try various cross-validation strategies, including nested cross-validation CV, which involves a double CV loop. For each outer CV fold, the grid-search inner-CV can be conducted on the training set, followed by assessing the outer CV test set. As the dataset suffers from imbalance and accuracy is the best measurement, we used other measures like recall, precision, and specificity. However, future studies can use the measurement index of intervention rate (Xie et al. 2022). We acknowledge the possibility of variables overlapping in the model, which is another limitation in our study. Finally, it is worth mentioning that the aim of this study is to assess the relationship between variable and outcome at a single point in time (cross-sectional study) However, if the proposed model is set in operation, it would be required to retrain the model periodically and, consequently, update coefficients to address any concept drift that may be encountered over time.

Healthcare fraud detection is fertile soil for research and, with continuously generated healthcare data, it is necessary for regulatory bodies (CHI), integration entities (NPHIES), healthcare providers, and insurance payers to collaborate to enable the building of more ML fraud detection models. Therefore, future research should encourage investigators to obtain a large representative dataset that contains various healthcare providers; departments; and a variety of ages, genders, education, and races, to create better fraud detection models.

5. Conclusions

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 logistics regression, random forest, and artificial neural networks 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.

Author Contributions: Conceptualization, E.N.; validation, A.A.; investigation, E.N.; data curation, E.N.; writing—original draft preparation, E.N.; writing—review and editing, A.A.; supervision, A.A.; project administration, A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was approved by the King Abdullah International Medical Research Center Review Board (SP22R/182/07).

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

Ahmed, Omar, and Adnan Brifcani. 2019. Gene Expression Classification Based on Deep Learning. Paper presented at the 4th Scientific International Conference Najaf (SICN), Al-Najef, Iraq, April 29–30; pp. 145–49.

Alharbi, Mohammad F. 2018. An analysis of the Saudi healthcare system's readiness to change in the context of the Saudi National Healthcare Plan in Vision 2030. *International Journal of Health Sciences* 12: 83–87. Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5969787/ (accessed on 26 March 2022).

Alonazi, Wadi B. 2020. Fraud and Abuse in the Saudi Healthcare System: A Triangulation Analysis. *Inquiry* 57: 1–8. [CrossRef]

Asha, R. B., and K. R. Suresh Kumar. 2021. Credit card fraud detection using artificial neural network. *Global Transitions Proceedings* 2: 35–41. [CrossRef]

Bauder, Richard A., and Taghi Khoshgoftaar. 2018. Medicare fraud detection using random forest with class imbalanced big data. Paper presented at the 2018 IEEE 19th International Conference on Information Reuse and Integration (IRI), Salt Lake City, UT, USA, July 6–9; pp. 80–87.

Byrd, James, Paige Powell, and Douglas Smith. 2013. Health care fraud: An introduction to a major cost issue. *Journal of Accounting, Ethics and Public Policy* 14: 521–39.

CCHI. 2021. Private Health Insurance Sector Conduct Policy (Fraud, Waste, and Abuse). Available online: https://www.cchi.gov.sa/en/AboutCCHI/Rules/document/PrivateHealthInsuranceSectorConductPolicy.pdf (accessed on 28 March 2022).

Chen, Zhen Xing, Lindsey Hohmann, Bidur Banjara, Yi Zhao, Kavon Diggs, and Salisa C. Westrick. 2020. Recommendations to protect patients and health care practices from medicare and medicaid fraud. *Journal of the American Pharmacists Association* 60: e60–e65. [CrossRef] [PubMed]

Connelly, Lynne. 2020. Logistic Regression. *Medsurg Nursing* 29: 731–35.

Doerpinghaus, Helen I., Joan T. Schmit, and Jason Jia-Hsing Yeh. 2008. Age and gender effects on auto liability insurance payouts. *Journal of Risk and Insurance* 75: 527–50.

Doupe, Patrick, James Faghmous, and Sanjay Basu. 2019. Machine Learning for Health Services Researchers. *Value in Heal* 22: 808–15. [CrossRef]

Fotouhi, Sara, Shahrokh Asadi, and Michael W. Kattan. 2019. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *Journal of Biomedical Informatics* 90: 103089. [CrossRef]

Karmiani, Divit, Ruman Kazi, Ameya Nambisan, Aastha Shah, and Vijaya Kamble. 2019. Comparison of Predictive Algorithms: Backpropagation, SVM, LSTM and Kalman Filter for Stock Market. Paper presented at the 2019 Amity International Conference on Artificial Intelligence (AICAI), Dubai, United Arab Emirates, February 4–6; pp. 228–34.

Kaur, Sukhpal, Himanshu Aggarwal, and Rinkle Rani. 2020. Hyper-parameter optimization of deep learning model for prediction of Parkinson's disease. *Machine Vision and Applications* 31: 1–15. [CrossRef]

Kumar, Yogesh, Sameeka Saini, and Ritu Payal. 2021. Comparative Analysis for Fraud Detection Using Logistic Regression, Random Forest, and Support Vector Machine. *SSRN Electronic Journal* 7: 726–31.

Lesch, William C., and Brent R. Baker. 2013. Balancing the Insurance Equation: Understanding the Climate for Managing Consumer Insurance Fraud and Abuse. *Journal of Insurance Issues, Western Risk and Insurance Association* 36: 82–120.

Mackey, Tim Ken, Ken Miyachi, Danny Fung, Samson Qian, and James Short. 2020. Combating health care fraud and abuse: Conceptualization and prototyping study of a blockchain antifraud framework. *Journal of Medical Internet Research* 22: e18623.

Mayaki, Mansour Zoubeirou A., and Michel Riveill. 2022. Multiple Inputs Neural Networks for Medicare fraud Detection. *arXiv* arXiv:2203.05842. [CrossRef]

NHCAA. 2018. The Problem of Health Care Fraud: A Serious and Costly Reality for All Americans. National Health Care Anti-Fraud Association (NHCAA). Available online: http://www.nhcaa.org/resources/health-care-anti-fraud-resources/the-challenge-of-health-care-fraud/ (accessed on 8 October 2021).

Nicholas, Lauren Hersch, Caroline Hanson, Jodi B. Segal, and Matthew D. Eisenberg. 2020. Association between Treatment by Fraud and Abuse Perpetrators and Health Outcomes among Medicare Beneficiaries. *JAMA Internal Medicine* 180: 62–69.

Patel, Pinak, Siddharth Mal, and Yash Mhaske. 2019. A Survey Paper on Fraud Detection and Frequent Pattern Matching in Insurance Claims using Data Mining Techniques. *International Research Journal of Engineering and Technology* 6: 591–94.

Rahman, A. Sazzadur, F. M. Javed Mehedi Shamrat, Zarrin Tasnim, Joy Roy, and Syed Akhter Hossain. 2019. A comparative study on liver disease prediction using supervised machine learning algorithms. *International Journal of Scientific & Technology Research* 8: 419–22.

SAMA. 2021. Saudi Insurance Market Report. Available online: https://www.sama.gov.sa/en-US/Insurance/Publications/Insurance_Market_Report_2021_English.pdf (accessed on 5 February 2023).

Severino, Matheus Kempa, and Yaohao Peng. 2021. Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata. *Machine Learning with Applications* 5: 100074. [CrossRef]

Shamitha, S. Kotekani, and Velchamy Ilango. 2020. A time-efficient model for detecting fraudulent health insurance claims using Artificial neural networks. Paper presented at the 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, India, July 3–4.

Shipe, Maren E., Stephen A. Deppen, Farhood Farjah, and Eric L. Grogan. 2019. Developing prediction models for clinical use using logistic regression: An overview. *Journal of Thoracic Disease* 11 S4: S574–84.

SPA. 2022. Council of Health Insurance Completes First Phase of Linking to NPHIES Platform. Saudi Press Agency. Available online: https://www.spa.gov.sa/viewfullstory.php?lang=en&newsid=2317570 (accessed on 23 March 2022).

Sumalatha, M. R., and M. Prabha. 2019. Mediclaim Fraud Detection and Management Using Predictive Analytics. Paper presented at the 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), Dubai, United Arab Emirates, December 11–12; pp. 517–22.

Suri, Sheffali, and Deepa V. Jose. 2019. Effective Fraud Detection in Healthcare Domain using Popular Classification Modeling Techniques. *International Journal of Innovative Technology and Exploring Engineering* 8: 579–83.

Thaifur, Andi Yaumil Bay R., M. Alimin Maidin, Andi Indahwaty Sidin, and Amran Razak. 2021. How to detect healthcare fraud? "A systematic review". *Gaceta Sanitaria* 35: S441–49. [CrossRef] [PubMed]

Timofeyev, Yuriy, and Tatiana Busalaeva. 2019. Current Trends in Insurance Fraud in Russia: Evidence from a Survey of Industry Experts. *Security Journal* 34: 1–25. [CrossRef]

Varmedja, Dejan, Mirjana Karanovic, Srdjan Sladojevic, Marko Arsenovic, and Andras Anderla. 2019. Credit Card Fraud Detection—Machine Learning Methods. Paper presented at the 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH), East Sarajevo, Bosnia and Herzegovina, March 20–21; pp. 1–5.

Villegas-Ortega, José, Luciana Bellido-Boza, and David Mauricio. 2021. Fourteen years of manifestations and factors of health insurance fraud, 2006–2020: A scoping review. *Health & Justice* 9: 1–23.

Waghade, Shivani S., and Aarti M. Karandikar. 2018. A comprehensive study of healthcare fraud detection based on machine learning. *International Journal of Applied Engineering Research* 13: 4175–78. Available online: https://www.ripublication.com/ijaer18/ijaerv13n6_140.pdf (accessed on 22 April 2023).

Xie, Yu, Guanjun Liu, Chungang Yan, Changjun Jiang, Mengchu Zhou, and Maozhen Li. 2022. Learning transactional behavioral representations for credit card fraud detection. *IEEE Transactions on Neural Networks and Learning Systems*, 1–14. [CrossRef]

Zarour, Mohammad, Mamdouh Alenezi, Md Tarique Jamal Ansari, Abhishek Kumar Pandey, Masood Ahmad, Alka Agrawal, Rajeev Kumar, and Raees Ahmad Khan. 2021. Ensuring data integrity of healthcare information in the era of digital health. *Healthcare Technology Letters* 8: 66–77.

Zhou, Jiantao, Shanshan Wang, Jianbo Zhou, and Yanli Xu. 2017. Measurement of the Severity of Opportunistic Fraud in Injury Insurance: Evidence from China. *Emerging Markets Finance and Trade* 53: 387–99.