International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT228664

Fraud Detection in Medical Insurance Claim System using Machine Learning : A Review

Paresh Gohil [*1], Dr. Sheshang Degadwala[2], Dhairya Vyas[3]

[*1]Computer Engineering Department, Sigma Institute of Engineering, Vadodara, Gujarat, India
[2]Computer Engineering Department, Sigma Institute of Engineering, Vadodara, Gujarat, India
[3]Managing Director, Shree Drashti Infotech LLP, Vadodara, Gujarat, India

Article Info

Publication Issue :
Volume 8, Issue 6
November-December-2022

Page Number : 417-427

Article History
Accepted: 20 Nov 2022
Published: 05 Dec 2022

ABSTRACT

Since the beginning of the insurance industry, there has been the problem of fraudulent insurance claims. These are a broad variety of illegal activities, the most of which are never uncovered while costing the insurance industry billions of dollars annually. It is estimated that India's insurance industry is suffering losses of around 600–Rs. 600 million each year because of India's growing economy, more awareness, and strengthened distribution networks. 800 crores in losses sustained yearly due to bogus claims. India comes up at number 10 for gross premiums collected by life insurance companies and number 15 for the total amount earned by non-life insurance companies. As a result of this, we are presenting a framework for the selection of features to be used in machine learning, which will enable the robust categorization of insurance claims. It will demonstrate how these technologies might be used to the development of a system that can prevent certain kinds of fraud in the field of healthcare. Several different studies have been carried out to demonstrate that the established approach may effectively identify instances of healthcare fraud. As a result, it may be useful in the prevention of false claims and gives greater insight into how to enhance patient management and treatment methods.

Keywords: Medical Insurance Claim, Support Vector Machine, K-nearest Neighbor, Random Forest, Decision Tree, Navier Bayes.

I. INTRODUCTION

Since the beginning of the insurance industry, dishonest practises have plagued policyholders. These are a broad variety of illegal activities, the most of which are never uncovered while costing the insurance industry billions of dollars annually. It is anticipated that by the year 2020, the insurance market in India would be worth a total of $280 billion USD. This development will be fuelled by greater economic activity, heightened consumer knowledge, and improved distribution channels. India comes up at number 10 for gross premiums collected by life insurance companies and number 15 for the total amount earned by non-life insurance companies. As a result of this, we have decided to implement a

framework based on blockchain technology that will enable safe transactions and the transmission of data between the many interacting agents in the insurance network.

In a perfect world, a representative from an insurance company would be able to evaluate each claim and determine whether it is legitimate. Having said that, this procedure is not only time demanding but also expensive. It is just impossible to find and pay for the specialised labour needed to evaluate each of the thousands of claims that are made each day, thus this task cannot be completed. An automated approach has proven to be the most effective technique so far. Nevertheless, the tools that were accessible in the past would have only enabled basic analysis with a restricted degree of precision. Even after the detection of a claim that may have been the result of fraudulent activity, an insurance agent would still need to do more research.

Benefits:

1. Fraud will be more precisely recognised in all claims that are suspected of being fraudulent.
2. The processing of data takes place in very rapid successions.
3. The system can show where links may exist between a variety of parameters, even if those relationships aren't visible to the human eye.
4. The ongoing modification of these kinds of schemes, together with the use of different approaches to data analysis, will make it possible to anticipate the finding of new fraud schemes.

II. LITERATURE STUDY

No	Title	Publication-Year	Methods use	Advantages	Limitation
1	A Secure AI-Driven Architecture for Automated Insurance Systems: Fraud Detection and Risk Measurement [1]	AIDS Research and Therapy-2020	IEEE Acess-2020	XGboost achieves high performance gains compared to other existing learning algorithms. For instance, it reaches 7% higher accuracy compared to decision tree models when detecting fraudulent claims.	Focus on enhancing the proposed architecture and implementing AI solutions that are tailored to other insurance services.

2	Machine Learning adoption in Blockchain-based Smart Applications: The challenges, and a way forward [2]	IEEE Access-2017	Machine Learning (ML) methods like Support Vector Machine (SVM), Clustering, Bagging, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) in Blockchain technology (BT) based Smart Applications.	The combination of ML and BT can provide highly precise results and more resilient against attacks.	Emphasized on several research prospects like infrastructure availability, quantum resilience and privacy issues that can serve as future research direction in the field.
3	Prediction of Claims in Export Credit Finance: A Comparison of Four Machine Learning Techniques [3]	IEEE Access-2017	Machine Learning (ML) methods like Support Vector Machine (SVM), Clustering, Bagging, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) in Blockchain technology (BT) based Smart	The combination of ML and BT can provide highly precise results and more resilient against attacks.	Emphasized on several research prospects like infrastructure availability, quantum resilience and privacy issues that can serve as future research direction in the field.

			Applications.		
4	Healthcare Insurance Frauds: Taxonomy and Blockchain-Based Detection Framework (Block-HI) [4]	IEEE-2021	Blockchain Technology	Blockchain based health insurance fraud detection framework for automated fraud detection and potential of identifying different fraud scenarios compared to manual process.	1.Compare the performance of Block-HI with current manual process and developing a framework of interoperability among data claims used by different healthcare insurance branches
5	Sequence Mining and Prediction-Based Healthcare Fraud Detection Methodology [5]	IEEE Access-2020	Sequence Mining and Sequence Prediction (Patient time series traces)	Average accuracy upto 85% in detecting fraud.	The data set used to validate the proposed methodology was difficult to obtain as it contains private and confidential information. The data set was in raw form and to handle the missing and redundant information was also time consuming.
6	Predicting Fraudulent Claims in Automobile Insurance [6]	ICICCT 2018/IEEE Xplore Complaint 2018	Data Mining and Machine Learning (ML) methods like Naïve Bayes, J48 and Random Forest.	Random Forest outperforms the remaining two algorithms over Insurance claim dataset and Naïve Bayes performs well in Premium dataset under all three test options.	In upcoming work, important relations between Insurance claim and premium dataset will be discovered. Classification algorithms will be customized to optimize results over real datasets
7	Blockchain-Powered Parallel Healthcare	IEEE 2018	Artificial System + computational	Blockchain-powered PHS based on the ACP	To build a consortium blockchain that contains patients,

	Systems Based on the ACP Approach [7]		experiments + parallel execution (ACP) approach in Blockchain-powered Parallel Healthcare Systems (PHS).	approach uses artificial system modeling to simulate and represent the actual healthcare scenarios.	hospitals, health bureau and so on with the purpose of enabling the PHS more integrity, scalability and security. Further improve the blockchain powered PHS and make it available for more disease treatments scenarios
8	A secure Healthcare System Design Framework using Blockchain Technology [8]	ICACT-2019	Internet of Things (IOT) and the Blockchain Technology based Secure Healthcare System Design Framework.	The proposed system implements the notion of fetching data remotely from the patient's wearable devices and Bio Sensors with the concept of blockchain which is a consortium of multiple stakeholders such as hospitals, doctors, etc.	The data generated from the wearable devices and bio sensors need to be provided timely and accurate and always governed in proper way.
9	Untangling Blockchain: A Data Processing View of Blockchain Systems [9]	IEEE-2017	Ethereum, Parity and Hyperledger Fabric in Blockchain technology (BT).	Benchmarking framework, BLOCKBENCH which is designed to evaluate the performance of blockchains as data processing platforms.	Four potential research directions inspired by database design principles, for improving blockchain performance.

10	To Blockchain or Not to Blockchain: That Is the Question [10]	IEEE-2018	Blockchain technology (BT) In Insurance Sector.	Using blochchain and smart contracts for lowering the operating costs, improving customer experience, and increasing transparency in a new emerging market for a company.	Specific Domain evaluation and analysis could easily be extended to other scenarios sharing comparable use cases, thus helping professionals make decisions in different contexts and sectors.
11	A Blockchain Framework for Insurance Processes [11]	IEEE-2018	Application of Blockchain framework systems like Smart contracts and Hyperledger Fabric in Insurance processes.	Blockchain-based framework for implementing insurance transaction processes as smart contracts to automate and speed up business processes, to make fraud detection easier using decentralized digital repository, to reduce administrative and operational costs, to enable regulators and auditors to detect suspicious patterns and market behaviours.	The database can be encrypted with fine grained access control. Further, it can extend even to the transaction level, to enable separate set of endorsing peers for each encryption.
12	Detecting Insurance Claims Fraud Using Machine	IEEE Access-2017	Machine Learning (ML) methods like Decision	Decision Tree and Random Forest have better performance than	In future can be work with more algorithms and finally calculate which provide more

	Learning Techniques [12]		Tree (DT), Random Forest (RF) and Naïve Bayes (NB) in detecting fraud claims in auto/ vehicle insurance.	Naïve Bayes for detecting fraud claims in auto/vehicle insurance.	accuracy, precision, and recall.
13	Integrating Blockchain for Data Sharing and Collaboration in Mobile Healthcare Applications [13]	IEEE-2017	Blockchain technology (BT) based system on Hyper ledger Fabric for personal health data sharing and collaboration using mobile user controlled.	The system can handle a large dataset at low latency which indicates the scalability and efficiency of data process. By adopting Merkle tree method to batch data, implement an algorithm with computation complexity of O(log2n) when data records are collected at a high frequency.	In future it can be explore, how to combine both personal health data and medical data together and cover a broader scenario.
14	An Efficient Authentication Scheme for Blockchain-Based Electronic Health Records [14]	IEEE-2019	Multiple Authorities – Identity Based Signature (MA-IBS) for blockchain based Electronic Health Records (EHR).	The proposed authentication scheme for blockchain-based EHRs has lower computation and communication costs including better resistance to collision attack compared to the	In future there may be application of other algorithms and compare the performance.

				only two existing authentication schemes for blockchain-based EHRs.	

III. METHODOLOGY

A. Datasets

Healthcare fraud is considered a challenge for many societies. Health care funding that could be spent on medicine, care for the elderly, or emergency room visits is instead lost to fraudulent activities by materialistic practitioners or patients. With rising healthcare costs, healthcare fraud is a major contributor to these increasing healthcare costs.

Detailed Data File:

https://www.kaggle.com/datasets/tamilsel/healthcare-providers-data.

B. Noise Removal Techniques

Abuse of shortenings, data transfer mistakes, duplicate entries, and missing values account for most of the noise that requires normal noise reduction procedures to rectify. In this context, one important area of study is the de-duplication problem, which involves the detection and removal of duplicate entries from a database. Databases provide a problem for investigation because they may include both precise and hazy versions of a given record. At this point, the numerical dataset has been acquired and any unnecessary information has been removed.

(1) Shortenings [1, 2]: Data is not valuable in a array. Data is only valuable once information, insight or in other words knowledge is extracted from it and is used to make decisions. In data shorting data is divided into equal size raw and column.

(2) Missing esteems [1, 4]: Missing value imputation is crucial in machine learning when data is small, and all available data must be used. It affects model classification performance. As our dataset is tiny and all characteristics except output have missing values, we must impute them. In this study, we employed mean and mode missing value imputation. Mean was used to infer numerical characteristics and mode for categorical.

(3) Copy records [1,3]: Create a single statement to get rid of all the extra copies at once. You need to choose which duplicates to retain before you may delete them.

(4) Data passage botches [2]: In this part data is process in badly manner so the data passage botches algorithm will remove bugs data.

C. Attribute Selection

Attribute selection is a process of reducing the dimension of a dataset by eliminating the attributes of less importance.

(1) Best initial search traversal method [2, 6]: The best initial search traversal method prioritises the most promising nodes to visit first. This is done by letting an evaluation function choose the path taken.

(2) Wrapper subset attribute evaluator [4, 6]: Methods for encapsulating the selection of feature subsets When there are fewer variables, performance improves. Manual processes may begin with a whole collection of attributes, then iteratively eliminate the least useful ones until just the optimal attribute remains. The "wrapper" approach employs a cross-validation loop around a classifier, using the classifier to do an exhaustive search of the attribute space in order to identify the optimal subset of attributes. Forward, backward, and even directed searches are possible from any subset.

D. Machine Learning

(1) SVM [1,4,6]: For this classification task, supervised machine learning techniques such as the support vector machine (SVM) are often utilised. To perform classification and other tasks, such as outlier identification, SVM creates a hyperplane or group of hyperplanes in a high dimensional space. The hyperplane with the greatest distance to the closest training-data point of any class provides the best classification.

(2) NB [9,10,15]: The conditional probabilities are the foundation of a Naive Bayes classification method. Naive Bayes employs the Bayes formula, which determines the likelihood of an event by summing the occurrences of each value and each value combination in the past. The INB model is simple to construct and effective for massive data.

(3) k-NN [3,5]: An example of a supervised machine learning method is the K-nearest neighbour (KNN) algorithm. It's applicable to both regression and classification issues in predictive modelling. The voting system of an object's closest neighbours is used to determine its classification. The parameter "k" in k-NN is left up to the researcher. It's a lazy learning algorithm that doesn't rely on any parameters.

(4) Decision Tree [2,7]: With a decision tree, data is organised in a tree structure, and then branches out to be categorised. All those forks signify different possibilities. The choices and their potential effects and benefits are shown in a tree-like form. Furthermore, it may be integrated with several other algorithms.

(5) RF [11,12,14]: The Random Forest method constructs an ensemble decision tree and is used in supervised machine learning. It may be put to work in both classification and regression analysis. Simple random forest constructs many different decision trees and then merges them into a single, correct conclusion. In RF, the greater the number of trees used, the more precise the final outcome will be.

IV.Comparative Analysis

TABLE I

COMPARATIVE ANALYSIS

Method	Advantage	Limitation
SVM [1,2,15]	It works relatively well once there's a transparent margin of separation between classes. it is simpler in high dimensional spaces. it is effective in cases wherever the number of dimensions is bigger than the number of samples. it is comparatively memory efficient.	SVM rule isn't appropriate for giant data sets. SVM doesn't perform alright once the information set has additional noise
Rf [3,14]:	It reduces overfitting in call trees and helps to enhance the accuracy. It's versatile to each classification and regression problems. It works well with both categorical and continuous values. It automates missing values present within the data.	The main limitation of random forest is that an outsized range of trees will make the algorithmic rule too slow and ineffective for period predictions. In general, these algorithms are quick to train, however quite slow to form predictions once they're trained.

KNN [2,6,8]	New knowledge will be additional seamlessly which cannot impact the accuracy of the algorithm. KNN is incredibly simple to implement. There are solely 2 parameters needed to implement KNN i.e., the worth of K and the distance perform.	With massive data, the prediction stage may be slow. Sensitive to the size of the info and impertinent features. Need high memory to store all the coaching data. In case it stores all the training, it can be computationally expensive.
DT [10,12]	Their outputs are simple to read and interpret without applied Statistical knowledge. simple to prepare. Less data cleansing required.	They are unstable, that means that a little amendment within the knowledge will result in a large change in the structure of the optimal decision tree. typically, comparatively inaccurate.
Logistic Regression [1,2-6]	Logistic regression is easier to implement, easier to interpret, and very efficient to train. It is very quick to classify unknown records. it Does not make any assumptions about class distributions in the feature space.	The main limitation of logistic regression is the assumption of linearity between the dependent variable and the independent variables.
Naive Bayes [6-12,14]	It is straightforward and simple to implement. It doesn't need the maximum amount coaching information. It handles each continuous and separate data. it's extremely scalable with the number of predictors and data points.	Drawback of Naive Bayes is the belief of independent predictors. In actual life, it's far nearly not possible that we get a fixed of predictors which can be independent.

V. CONCLUSION

Healthcare insurance fraud detection is a major concern for the healthcare industry. Tens of billions of dollars are lost every year due to healthcare insurance frauds. Some frauds are at the risk of patient health. To the best of our knowledge, I have study ML work in the literature for fraud detection framework based on fixed fraud scenarios.

In future work, ML and blockchain-based framework for insurance system fraud detection, based on our taxonomy of fraud scenarios.

VI.REFERENCES

[1] N. Dhieb, H. Ghazzai, H. Besbes, and Y. Massoud, "A Secure AI-Driven Architecture for Automated Insurance Systems: Fraud Detection and Risk Measurement," IEEE Access, vol. 8, pp. 58546–58558, 2020, doi: 10.1109/ACCESS.2020.2983300.

[2] S. Tanwar, Q. Bhatia, P. Patel, A. Kumari, P. K. Singh, and W. C. Hong, "Machine Learning

Adoption in Blockchain-Based Smart Applications: The Challenges, and a Way Forward," IEEE Access, vol. 8, pp. 474–448, 2020, doi: 10.1109/ACCESS.2019.2961372.

[3] M. Bärtl and S. Krummaker, "Prediction of claims in export credit finance: a comparison of four machine learning techniques," Risks, vol. 8, no. 1, 2020, doi: 10.3390/risks8010022.

[4] L. Ismail and S. Zeadally, "Healthcare Insurance Frauds: Taxonomy and Blockchain-Based Detection Framework (Block-HI)," IT Prof., vol. 23, no. 4, pp. 36–43, 2021, doi: 10.1109/MITP.2021.3071534.

[5] I. Matloob, S. A. Khan, and H. U. Rahman, "Sequence Mining and Prediction-Based Healthcare Fraud Detection Methodology," IEEE Access, vol. 8, pp. 143256–143273, 2020, doi: 10.1109/ACCESS.2020.3013962.

[6] G. Kowshalya and M. Nandhini, "Predicting Fraudulent Claims in Automobile Insurance," Proc. Int. Conf. Inven. Commun. Comput. Technol. ICICCT 2018, no. Icicct, pp. 1338–1343, 2018, doi: 10.1109/ICICCT.2018.8473034.

[7] S. Wang et al., "Blockchain-Powered Parallel Healthcare Systems Based on the ACP Approach," IEEE Trans. Comput. Soc. Syst., vol. 5, no. 4, pp. 942–950, 2018, doi: 10.1109/TCSS.2018.2865526.

[8] S. Chakraborty, S. Aich, and H. C. Kim, "A Secure Healthcare System Design Framework using Blockchain Technology," Int. Conf. Adv. Commun. Technol. ICACT, vol. 2019-February, pp. 260–264, 2019, doi: 10.23919/ICACT.2019.8701983.

[9] T. T. A. Dinh, R. Liu, M. Zhang, G. Chen, B. C. Ooi, and J. Wang, "Untangling Blockchain: A Data Processing View of Blockchain Systems," IEEE Trans. Knowl. Data Eng., vol. 30, no. 7, pp. 1366–1385, 2018, doi: 10.1109/TKDE.2017.2781227.

[10] W. Kozlow, M. J. Demeure, L. M. Welniak, and J. L. Shaker, "Acute extracapsular parathyroid hemorrhage: Case report and review of the literature," Endocr. Pract., vol. 7, no. 1, pp. 32–36, 2001, doi: 10.4158/ep.7.1.32.

[11] M. Raikwar, S. Mazumdar, S. Ruj, S. Sen Gupta, A. Chattopadhyay, and K. Lam, "2018 9th IFIP International Conference on New Technologies, Mobility and Security, NTMS 2018 - Proceedings," 2018 9th IFIP Int. Conf. New Technol. Mobil. Secur. NTMS 2018 - Proc., vol. 2018-January, 2018.

[12] R. Roy and K. T. George, "Detecting insurance claims fraud using machine learning techniques," Proc. IEEE Int. Conf. Circuit, Power Comput. Technol. ICCPCT 2017, 2017, doi: 10.1109/ICCPCT.2017.8074258.

[13] X. Liang, J. Zhao, S. Shetty, J. Liu, and D. Li, "Integrating blockchain for data sharing and collaboration in mobile healthcare applications," IEEE Int. Symp. Pers. Indoor Mob. Radio Commun. PIMRC, vol. 2017-October, pp. 1–5, 2018, doi: 10.1109/PIMRC.2017.8292361.

[14] F. Tang, S. Ma, Y. Xiang, and C. Lin, "An Efficient Authentication Scheme for Blockchain-Based Electronic Health Records," IEEE Access, vol. 7, pp. 41678–41689, 2019, doi: 10.1109/ACCESS.2019.2904300.

Cite this article as :

Paresh Gohil, Dr. Sheshang Degadwala, Dhairya Vyas, "Fraud Detection in Medical Insurance Claim System using Machine Learning : A Review", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 6, pp. 417-427, November-December 2022. Available at doi : https://doi.org/10.32628/CSEIT228664
Journal URL : https://ijsrcseit.com/CSEIT228664