



IJIRCCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 4, April 2022

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.165



9940 572 462



6381 907 438



ijircce@gmail.com



www.ijircce.com

Enriching Fraud Claims Detection in Health Sector using Machine Learning

Rutuja Dhumal, Harsha Baviskar, Sayali Bhawalkar, Jui Yadav, Prof. Geetanjali Yadav

Dept. of Computer, Keystone School of Engineering, Pune, Maharashtra, India

ABSTRACT: Health Insurance is mandatory in most of the developed nations such as Europe, the United States of America, etc. India also has facilities for a Health Insurance amenity for its inhabitants covering their medical costs. The insurance works in this way, when an insured person goes to a doctor, he/she can avail the medical facilities free-of-cost, they do not have to pay for any utilities that have been used at the doctor. The doctor then contacts the insurance company with the particulars of your visit and is then reimbursed by the insurance company. There is a loophole in this technique as the doctor can provide fraudulent invoices to the company to extract more money through unethical means. There are a few techniques that have been utilized to reduce the occurrence of this fraud but with low and varying precision. Therefore, this paper proposes a new methodology based on Machine Learning algorithms such as Artificial intelligence and K-Nearest Neighbours in addition to the Logistic Regression and Entropy analysis.

KEYWORDS: K-Nearest Neighbor, Entropy Estimation, ANN, Correlation Estimation.

I. INTRODUCTION

Medical care is an important part among majority people's daily existence, and it must therefore remain inexpensive. The necessity for treatment is especially critical for the ageing demographic. They want more attention and, as a result, adequate healthcare coverage for a number of treatment medications and procedures. The proportion of senior people has gradually increased significantly in subsequent times, compared to a far smaller growth in the percentage of people under 65. As a result, the aged, as well as their friends and relatives, place a greater emphasis on maintaining and improving their wellbeing. This growing healthcare demand requires money, which is generally covered by healthcare coverage.

Undoubtedly, these programs must be cost-effective for the broader public, yet programs expenses, together with the older population, are continuing to rise, putting people and communities in economic jeopardy. Authorities and economic organizations are increasing their investments. Despite with all these significance and economic implications, attempts to reduce fraud, waste, and abuse are critical to lowering expenses.

The concentration of the study is on the Health insurance scheme fraud because of the importance of geriatric treatment. As medical costs and expectancy intervals continue to go up, individuals and organizations that help support medical appointments and procedures are under increased financial strain. As a result, the medical platform's purpose should have been to deliver adequate and required healthcare to as many people as appropriate at a reasonable cost to both patients and health care providers. Medicare is one such health system. It is a government-run programs that offers monetary support to pensioners as well as other designated demographics. Each authorized operation is assigned a particular process, comparable to certain other health coverage systems such as Medicaid.

A physician, in essence, undertakes a sequence of operations and then files a claim to Healthcare insurance for reimbursement. This issue is growing increasingly prevalent all across the world these days. In particular, in underdeveloped countries in which the governments has only just recently begun to provide treatment to the general public. A correct allocation of assets to all essential regions is necessary. The identification of illicit activity is critical, and an FWA identification system is required to address this issue.

Throughout most nations, healthcare system seems to have become a substantial source of economic spending. Because of the massive amounts of money associated in the healthcare industry, it has now become a target for fraud. Health care fraud costs the United States huge amounts of money each year. Health-care fraud could take place in a variety of human jobs. Individualized health fraud perpetrators can be easily distinguished. In essence, these individuals are manipulated and subjected to medical interventions and operations that are either unneeded or inappropriate. Alternatively, their medical files may have been hacked, or their insurance credentials may have been exploited to fabricate claims.

Fraud is a serious crime that costs both individuals and organizations a lot of money. As a result, fraud identification is required to reduce costs and improve the integrity of healthcare system. In underdeveloped nations, the administration has lately launched an initiative to cover each family's medical costs. Medical fraud must be recognized and eliminated on the scene, which necessitates the development of fraud identification technologies. The major aim is to deter people from engaging in fraudulent actions such as exploiting medical care system funds to meet unrelated personal requirements. Universal basic healthcare plans will spur rapid improvement in the implementation of and accessibility to high medical services for the world's poorest people. There would be a requirement for algorithms to recognize any fraudulent claims or activities when using these universal medical treatment programs.

Detecting health care related fraud is a difficult and time-consuming process. In the past, health insurance companies conducted thorough examinations and employed set procedures to detect fraud. Because of the expanded size of records, traditional approaches may be unable to discover a large number of instances for two key considerations. Fundamentally, it is impossible to detect all Medicaid fraud by individually analyzing large datasets. Following then, new varieties of Medicaid fraud having emerged on a constant basis. The latest emerging fraud techniques are undetectable by structured query language approaches predicated on a set of regulations.

Similar situations need the use of more complex structured approaches and strategies capable of identifying fraud events in large datasets. Due to the obvious magnitude of the medical coverage organization and the large sums of money associated, Medicaid type of fraud is a lucrative fraudulent field. An intentional act of deception or fabrication by an organization or group with both the knowledge that the misrepresentation may result in an unlawful compensation to that person or group of people is known as Medicaid fraudulent activity. As a consequence, effective fraud early detection is critical for improving the standards of medical insurance coverage while lowering costs.

This research paper dedicates section 2 for analysis of past work as literature survey, section 3 deeply elaborates the proposed technique and whereas section 4 evaluates the performance of the system and finally section 5 concludes the paper with traces of future enhancement.

II. LITERATURE SURVEY

A three-layer MLP feed-forward neural network was presented by H. Peng [1] for detecting healthcare fraud. Meanwhile, the author used the over-fitting and under-fitting problems to build the pharmacopeia spectrum analytic hierarchy tree and the neural network's clustering technique. The comparative outcome proves that the presented technique is superior to other unsupervised clustering methods. Furthermore, the authors developed a technique for calculating influence factors, which may be utilized to estimate the impact that each input has on the decision of whether or not a medical item is a fraud.

G. Saldamli et al. created a conceptual model in which false information is deleted largely by following a set of regulations called HIPAA Privacy standards defined by the US Department of Health and Human Services to securely keep each individual's health-related information in the country [2]. Furthermore, insurance fraud may be avoided by utilizing Blockchain technology to permanently register all policyholder transactions. The authors have limited access to the data to specific parties by utilizing security measures available on the blockchain. By eliminating the requirement for a third party, blockchain ensures that each block of data is visible to all parties engaged in the chain. Companies collaborating to exchange data on Blockchain would result in less policy manipulation. Using Blockchain technology, the authors have combined and connected information about the claimant from several health insurance providers, which is used as a reference to identify fraud.

I. Matloob et al. [3] present a new fraud detection algorithm for detecting fraudulent claims in government-sponsored medical assistance programs. The authors used a private hospital's employee's five-year insurance claim data to validate their technique. The proposed technique is being considered as a pilot module for the government-level project indicated above. The research makes a significant addition in the system the authors create a collection of sequences for each specialty after examining five years of transactional data. For fraud detection, the suggested system employs sequence mining and sequence prediction. Sequence mining is done by constructing a sequence rule engine for each specialty that is based on a collection of common sequences and the probability of rare sequences. The process also employs the Bayes Theorem, which is applied to infrequent sequences and the probability of their occurrence is calculated.

R. A. Bauder examines performance outcomes and statistical significance of several machine learning algorithms for detecting fraudulent Medicare providers. The author employs four distinct performance measures and two different sample strategies. Each of the 10 models is trained and evaluated using Medicare PUF data from 2015 and fraud labels from the LEIE database [4]. The experimental findings reveal a significant performance disparity between the sampling strategies. In comparison to oversampling, the 80-20 sampling strategy improves learner performance. Oversampling results in low performance across the board for all students.

M. R. Sumalatha [5] utilized predictive analytics to provide a unique way to detect the likelihood of a medical claim being false. To achieve maximum accuracy, a prediction model is built using logistic regression and multi-criteria decision analysis. The characteristics of these models are based on the National Insurance Institute of India's trigger-based scoring methodology. This medical fraud detection technology might be used in medical and insurance applications in the future. The established model may also be employed in sensitive applications that deal with fraudulent behaviors, such as detecting auto insurance fraud and credit card fraud.

The efficacy of two commonly used classification algorithms for finding outliers in healthcare datasets was investigated by R. Thomas [6]. Based on these two models, a hybrid outlier detection model is also presented. The performance of these classifiers was examined and compared to the performance of the proposed hybrid model for locating outliers in the UCI machine learning repository's breast cancer dataset and cardiocography dataset. The F1-score was employed as an assessment criterion for the performance comparison, and the suggested hybrid model outperformed the other two models. For any unbalanced dataset, this model may be utilized as a binary classification model. The model's complexity for high-dimensional large data remains a barrier, which may be mitigated by using feature bagging and suitable subsampling.

M. Herland et al. [7] study to create a system that can successfully discover physicians who work beyond the norm in their profession by utilizing anomaly detection. To enhance fraud detection skills, the authors continue their past studies and expand on this basic model. In this research, they evaluated and verified their original model against known fraudulent physicians, resulting in the successful labeling of 12 of the 18 (67%) physicians. The introduced model hypothesis is that if a physician does not submit procedures in the same way as their colleagues, which is considered odd, that physician may be engaging in the fraudulent or wasteful activity.

A. Gangopadhyay created algorithms aimed towards a certain form of scam. That is the questionable provider groups that either share or refer patients to one another. These communities are generally tiny and have only internal links with no external connections. Although the links between these communities appear to be suspect, the authors cannot be certain if they are engaging in fraudulent activity [8]. They must also examine other aspects, such as missing data and so on. These communities can be added to a watch list to be investigated and reviewed further. The additional evaluation will help avoid payments to certain categories of providers or patients being made in error.

A. Verma et al. suggested a healthcare fraud prevention methodology that is both cost-efficient and effective. Because detecting misconduct in health care is such a difficult undertaking, effective strategies are required to identify misconduct in this sector. The authors divide deceptive practices into two groups: period-based claim abnormalities and disease-based claim abnormalities [9]. Period-based claim abnormalities are explored using a statistical prediction model, which aids in the discovery of anomalies and frauds, and then clustering is utilized to ease the fraud detection technique. By uncovering association rule mining and recognizing common patterns, disease-based anomalies may be detected.

C. Sun et al. focus on collaborative fraud detection and present an abnormal group-based joint fraud detection method [10]. The presented strategy can overcome the difficulty of distinguishing suspicious joint fraudsters from those who have a high degree of resemblance due to periodicity. As a result, the proposed approach can guarantee a high level of precision. Furthermore, the authors propose a two-step H-graph-based MCE to minimize the computation time.

Extensive tests on a medical insurance dataset reveal that the presented strategy exceeds previous methods in precision by more than 20%.

I. Matloob suggested a framework that is made up of a system of three operational components: patients, providers, and services. Each component has a knowledge base that was generated through data learning. Patterns are discovered from transactional data in the first stage following data cleansing and reduction [11]. Each transaction's time traces have been created based on each dimension of the OLAP cube. This visualization will help in the identification of abnormal behavior and the validation of discovered abnormal transactions will be conducted using the current knowledge base and learned patterns.

H. Cui et al. created a healthcare fraud detection approach that depends on doctor trustworthiness, called GM-FP, to detect fake entries in healthcare records. This technique produced a reasonable treatment model for a specific disease by combining a graph-based mining algorithm with a frequent pattern mining algorithm. The anomalous records were identified using GM-FP by assessing the similarity of each unknown record to the rational model [12]. Furthermore, the authors addressed the issue of duplicating prescription behavior in the treatment sequence, which is important in determining a doctor's credibility. To show the effectiveness of the introduced strategy, the authors undertook a large experiment using medical insurance claim data.

R. Bauder investigates the effects of classification on the identification of fraud in the Medicare dataset using LEIE fraud labels using big data. To do this, the author first analyzes the Medicare dataset by aggregating data for each provider across all medical operations and mapping the associated LEIE database fraud categories. To address the issue of class imbalance, the author uses the random undersampling approach to generate seven class distributions or ratios, and develop Random Forest models for each distribution [13]. He employ 5-fold cross-validation performed 10 times for each model to decrease bias, and evaluate fraud detection performance using the Area Under the Receiver Operator Characteristic Curve. The findings show that class distributions with a 90:10 ratio generate the greatest overall outcomes.

III. PROPOSED SYSTEM

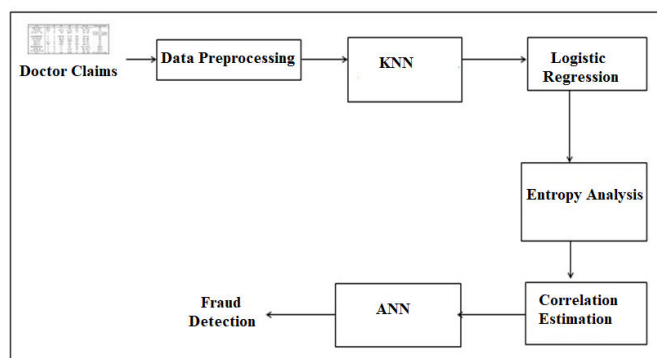


Figure 1: Overview of Health care Fraud Detection System

The proposed system for detection of Healthcare Fraud is elaborated below.

Step 1: Data Collection – The presented technique has been deployed for use by two individuals or professionals, such as the Health Council, that establishes the rules of the disease such as dietary suggestions, procedures, service cost, Referring Doctor Speciality, Time between Recalls, Number of Recalls, etc.

Another entity that can have access to this system is the staff from the Health insurance company, these people need the access as these are the people in charge of evaluating the workbook that has all the claims that need to be processed. The workbook is read by the system based on the Java interface with the help of a JXL API, and it is stored in a list of double dimensions.

Step 2: Pre-processing – Due to the fact that this is a Machine Learning process, the algorithms are unsupervised and have to be trained prior to its deployment. The system utilizes the data from past successful claims to learn and understand the protocols. The integer form is derived from the protocols and added to a double dimensional list. The

appropriate attributes are then selected from these protocols as the pre-processed list in the form of a double dimensional list.

Step 3: KNN – The pre-processed double dimensional list generated in the earlier step is utilized. The pre-processed list is then subjected to the calculation of mean along with the values given in the row of the list. the means of all the rows are then added to the end of the row. Equation 1 depicts the process accurately.

$$\mu = \frac{(\sum_{i=1}^n xi)}{n} \text{-----}(1)$$

Where,

xi=Each Attributes

n = number of attributes i.e. 8

To calculate the K Nearest Neighbours, the Euclidean distance between each of the rows of the pre-processed list needs to be evaluated. The resultant value is known as the row distance which is used to calculate the maximum and the minimum distance present in the pre-processed list. These two values of Maximum and minimum distance and then used to assign upper and lower boundary limits for the clusters by utilizing the algorithm 1.

Algorithm 1: Cluster Boundary formation

```
// Input: MinD, MaxD,K
[ MinD: Minimum Distance, MaxD: Maximum Distance, K: Number of Clusters]
// Output: BSET [Boundary List]
Function: boundaryFormation(MinD, MaxD,K)
Step 0: Start
Step 1: DIST= (MaxD, MinD) / K
Step 2: for j=0 to K
Step 3: R1= MinD
Step 4: R2= R1+DIST
Step 5: TSET=∅
Step 6: TSET[0]=R1, TSET[1]= R2
Step 7:ADD TSET TO BSET
Step 8: R1= R2
Step 9: End for
Step 10: return BSET
Step 11: Stop
```

Step 4: Regression Analysis and Entropy Estimation – The output of step 1 called the input list is generated in this step for the purpose of claims. The attributes mentioned in Equation 1 are mapped onto the double dimensional list with the values in each of their rows. Every value from the row is then correlated to the corresponding mean in the list to generate the attribute mean factor. The values of the absolute difference between the mean factors less than 0.1 are selected.

Each of these rows with an absolute difference less than 0.1 is then subjected to the entropy estimation, as the whole cluster is evaluated, according to the Shannon Information Gain theory as given in equation 2.

$$E = -\frac{X}{Z} \log \frac{X}{Z} - \frac{Y}{Z} \log \frac{Y}{Z} \text{-----}(2)$$

Where

X= Row Count

Z= Total number of rows of a cluster

Y= Z-X

E = Entropy Gain factor

Step 5: Correlation Estimation – As the last step binds the claim index with a respective cluster, this step utilizes those input claim indices to extract the corresponding disease and a correlation list is generated with both of those values in a double dimension list.

Step 6: ANN and Fraud Claim Estimation –The correlation list obtained in the previous step is then utilized further here to evaluate the cluster rows by assuming the attribute means from Equation 1. This value is then amalgamated with the neuron cluster row to extract the resultant claim index and its attribute mean factor greater than 0.4.

The correlation list count can be utilized to add or subtract to form the resultant values that indicate the level of fraud that has been committed by the Doctor. This information is then displayed to both the entities mentioned in step 1.

IV. RESULT AND DISCUSSIONS

The proposed Technique for the detection of Health insurance fraud committed by the Doctor has been programmed in Java on an IDE called NetBeans 8.0. This methodology has been implemented on a Machine running Windows with 6 GB physical memory and an Intel Core i5 as the Central Processing Unit. The database activities were handled by the MySQL Relational Database.

To measure the effectiveness of the model proposed system uses the Root mean square error(RMSE). Which is eventually used to measure the error rate between the two outcomes. Here the outcome is the prediction of the fraud Claims. This can be measured with the below mentioned equation.

$$RMSE_{fo} = \left[\sum_{i=1}^N (z_{fi} - z_{oi})^2 / N \right]^{1/2} \quad \text{---(3)}$$

Where

\sum - Summation

$(Z_{fi} - Z_{oi})^2$ - Differences Squared for the prediction of Fraud Claims

N - Number of experiments

Input Claims	No of Actual Fraud Claims	No of Predicted Fraud Claims	MSE
10	8	6	4
20	13	11	4
30	22	18	16
40	29	24	25
50	32	29	9

Table 1: Mean Square Error Reading

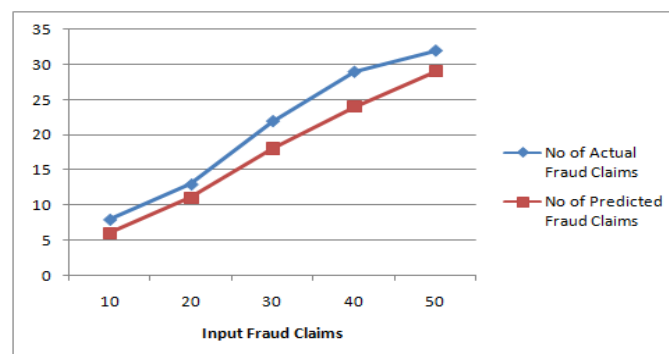


Figure 2: MSE in between No of Actual Fraud Claims V/s No of Predicted Fraud Claims

Mean Square Error (MSE) reading is tabulated in the table 1. And the average of the MSE is calculated as 11.6 and the square root of this is RMSE i.e. around 3.405. The yielded RMSE of our system is very low and it is better in the first trail of our experiment comparing with the traditional system. The plot in Figure 2 indicates the less distance between the actual and obtained results, that is eventually a good sign.

V. CONCLUSION AND FUTURE SCOPE

This Paper has effectively utilized Machine Learning Tools such as K Nearest Neighbour Clustering algorithm with Artificial neural network to estimate the level of Health Insurance fraud committed by a doctor. The clusters generated are experimented extensively to evaluate the insurance claims for an indication of fraud committed. The possibility of fraud is estimated by the use of various techniques such as Entropy Analysis, Logistic Regression, and correlation. The Artificial Neural Network increases the accuracy of the prediction by a large margin.

The direction for future research in this field will be enhanced to work more efficiently on the complicated attributes in a real-time implementation in a medical field. There can also be a provision for displaying the credentials of errant doctors by publicly humiliating them by displaying their actions on the internet.

REFERENCES

- [1] H. Peng and M. You, "The Health Care Fraud Detection Using the Pharmacopoeia Spectrum Tree and Neural Network Analytic Contribution Hierarchy Process," 2016 IEEE Trustcom/BigDataSE/ISPA, 2016, pp. 2006-2011, DOI: 10.1109/TrustCom.2016.0306.
- [2] G. Saldamli, V. Reddy, K. S. Bojja, M. K. Gururaja, Y. Doddaveerappa, and L. Tawalbeh, "Health Care Insurance Fraud Detection Using Blockchain," 2020 Seventh International Conference on Software Defined Systems (SDS), 2020, pp. 145-152, DOI: 10.1109/SDS49854.2020.9143900.
- [3] I. Matloob, S. A. Khan and H. U. Rahman, "Sequence Mining and Prediction-Based Healthcare Fraud Detection Methodology," in IEEE Access, vol. 8, pp. 143256-143273, 2020, DOI: 10.1109/ACCESS.2020.3013962.
- [4] R. A. Bauder and T. M. Khoshgoftaar, "Medicare Fraud Detection Using Machine Learning Methods," 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), 2017, pp. 858-865, DOI: 10.1109/ICMLA.2017.00-48.
- [5] M. R. Sumalatha and M. Prabha, "Mediclaime Fraud Detection and Management Using Predictive Analytics," 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), 2019, pp. 517-522, DOI: 10.1109/ICCIKE47802.2019.9004241.
- [6] R. Thomas and J. E. Judith, "Hybrid Outlier Detection in Healthcare Datasets using DNN and One Class-SVM," 2020 4th International Conference on Electronics, Communication, and Aerospace Technology (ICECA), 2020, pp. 1293-1298, DOI: 10.1109/ICECA49313.2020.9297401.
- [7] M. Herland, R. A. Bauder, and T. M. Khoshgoftaar, "Medical Provider Specialty Predictions for the Detection of Anomalous Medicare Insurance Claims," 2017 IEEE International Conference on Information Reuse and Integration (IRI), 2017, pp. 579-588, DOI: 10.1109/IRI.2017.29.
- [8] A. Gangopadhyay and S. Chen, "Health Care Fraud Detection with Community Detection Algorithms," 2016 IEEE International Conference on Smart Computing (SMARTCOMP), 2016, pp. 1-5, DOI: 10.1109/SMARTCOMP.2016.7501694.
- [9] A. Verma, A. Taneja, and A. Arora, "Fraud detection and frequent pattern matching in insurance claims using data mining techniques," 2017 Tenth International Conference on Contemporary Computing (IC3), 2017, pp. 1-7, DOI: 10.1109/IC3.2017.8284299.
- [10] C. Sun, Z. Yan, Q. Li, Y. Zheng, X. Lu and L. Cui, "Abnormal Group-Based Joint Medical Fraud Detection," in IEEE Access, vol. 7, pp. 13589-13596, 2019, DOI: 10.1109/ACCESS.2018.2887119.
- [11] H. Cui, Q. Li, H. Li, and Z. Yan, "Healthcare Fraud Detection Based on Trustworthiness of Doctors," 2016 IEEE Trustcom/BigDataSE/ISPA, 2016, pp. 74-81, DOI: 10.1109/TrustCom.2016.0048.
- [12] I. Matloob and S. Khan, "A Framework for Fraud Detection in Government Supported National Healthcare Programs," 2019 11th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), 2019, pp. 1-7, DOI: 10.1109/ECAI46879.2019.9042126.
- [12] R. Bauder and T. Khoshgoftaar, "Medicare Fraud Detection Using Random Forest with Class Imbalanced Big Data," 2018 IEEE International Conference on Information Reuse and Integration (IRI), 2018, pp. 80-87, DOI: 10.1109/IRI.2018.00019.



INNO  SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

 **doi**[®]
cross **ref**

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  ijircce@gmail.com



www.ijircce.com

Scan to save the contact details