International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075, Volume-8, Issue-7 May 2019

# Performance on Fraud Detection in Medical Claims of Healthcare Data

**P.Naga Jyothi, D Rajya lakshmi, K.V.S.N.Rama Rao**

***Abstract: Healthcare is more focused by most of the individuals. As the expenditure margin is beyond their limits too, the people do not care much about, apart from healthy livelihood. The interesting part of everyone's life is the government is providing policies and the individual is also holding private policies for their medical healthcare. At the same time, the fraud is growing faster in this scenario. Detection and prevention of fraudulent activities have been increasing with various sophisticated tools. But, still there are some lapses in analyzing and finding suspicious activities and mismanagement of system in medical insurance. In this paper, we described the survey of various technologies, methods applied to medical healthcare fraud detection of an individual, corporate hospitals and industries. The survey includes various characteristics of data, what are key steps for processing and analyzing the data for classification and finding of communities' methods for further fraud prevention and detection techniques. The majority of reviews the authors have concluded that the use of advanced machine learning techniques will improve the quality of healthcare systems. These algorithms can address some potential problems, comparisons and results were substantial with their limitations.***

***Keywords: medical healthcare, Machine learning, Fraud detection, Fraud Prevention Communities Suspicious activities, Classification.***

## I. INTRODUCTION

Every year the expenditure healthcare is being exceeded by many of the countries. Due to the extreme growth of market size and their influential factors this application domain requires a high-end data analytics mechanism. The significant problem of this wing is fraud, waste, abuse includes improper billing, repeated claims, uncovered services, drug abuses, counterfeit drugs, off-label marketing issues and many more. There is a series of technical challenges for data analytics. As there is massive storage of data over a period of time and from the representation point of view these all are many diverse datasets.

From the Traditional point of view, health fraud detection is much more dependent on domain expert, which formally predefined rules, but it, in turn, incurs factors of cost and time [1]. Therefore, to make the system more effective, many researchers have attempted to developan anti-fraud approach using data mining, Machine learning and many other sophisticated techniques [2].

**Revised Manuscript Received on May 07, 2019.**

**P.Naga Jyothi,** Research Scholar, Dept. of CSE, K L Educational Foundation, Guntur, India

**D Rajya lakshmi,** Professor, Dept. of Computer Science and Engineering, JNTUK UCEV, Narasaroa peta, India

**K.V.S.N.Rama Rao,** Professor, Dept. of CSE, K L Educational Foundation, Guntur, India

From the existing work, the present study incorporates the statistical methods for fraud detection of medical claims had been broadly categorized as Supervised and Unsupervised and hybrid approaches. Supervised methods is to be labeled by the experts and requires training dataset for the application domains whereas unsupervised methods purpose is to mine the outliers [2].The hybrid approach combines both features of prior approaches. To,attain objective for fraud detection and prevention involves the domain experts and statistical analysts for choosing the sample size, top priorities for the type of fraud, that leads to more financial lossas  data includes diverse information with certain constraints.

This paper aims to bring the background work of various fraudulent activities that occur in medical claim data. The objective of the study is to describe the different approaches for finding the behaviors and analyze the features using data mining algorithms. Our paper does not cover telecommunications, financial fraud, frauds in credit card, scientific fraud, computer intrusion and money laundering.

## II. CLASSIFICATION OF FRAUDS AND FRAUDULENT BEHAVIORS IN MEDICAL HEALTHCARE

According to US medical healthcare system, there are two main streams like and Medicaid and. Medicare services. Medicare is an insurance program controlled by the US government, it includes various, types of services like prescription drug coverage, hospital insurance, and medical insurance. Medicaid program which is run by state and it has its specific rules for the services [3] [4] [5].

In these two streams commonly there are three different parties includes a group of individuals which are generally involved in fraud in medical health care commission. 1. Service providers- who are generally doctors, hospitals, ambulance, companies and laboratories. These are various activities involves fraud Ex. Billing, unbundling, upcoding, falsify the patient's treatment, misrepresenting non covered treatment etc., 2. Insurance subscribers- which includes patients and patient's employers. The fraudulent activities carried in these ways like falsifying the records and illegal claims.3. Insurance carriers- who are mediators from government and private insurance companies will receive the premium amount from the insurance subscribers and pay on behalf of their subscribers to companies. Most of the ways they can involve in fraud falsifying the reimbursements, services and treatments. In addition to these, when a fraudulent activity carried by different parties it is termed as Conspiracy fraud. [1] [2] [6].


*Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication*

From the above classification, the majority of fraud is committed by service providers and these group of individuals is creating a loss to the health care system [7] [8].

## III. RELATED WORKS

Hossein Joudaki et.al (2015) describes different approaches to finding fraud and abuse data in healthcare data using data mining methods. His reviewed work from 21 papers, mentions the type of fraud detected and application of data mining method. Their work recommends seven common steps for finding fraud in health care claims [9].

The works of Karthiresanet.al (2015) has compared three reviews in healthcare claims for management of fraud and fraud detection. His work analyzed the classification variance, for abuse and not abuse data on employer dataset. The comparative study of accuracy is shown based on the techniques spectral analysis, Support vector machine and Multilayer Neural network and their merits and demerits are discussed [10].

Qi Liu et.al (2013) analyzes the features of data related health care using data mining and machine learning methods and relates three studies for both Medicaid and Medicare services. His work indicates three major phases like data preparation, preprocessing and analysis. The experimental analysis is done on Pneumonia, Rehabilitation procedure and Septicemiaon the claiming behavior for their principal diagnoses. The clustering method discriminates the distribution of claim amount for each disease and abstracts the fraudulent claim [1].

MelihKirlidoget.al (2012) discusses a few cases for finding fraudulent by analyzing the past insurance claiming behavior using data mining software in Turkish health insurance data. His work involves various data mining algorithms with various application examples to identify the fraud claims [12].

Electronic fraud detection techniques are discussed with a systematic survey of Peter Travaille et.al (2011). The results of fifteen review papersinclude the overview of fraud detection techniques and various features like industry, an objective the paper, type of fraud identified, techniques used, results and findings [13].

### Related works on Unsupervised Fraud detection

The work of CheNgufor and Janusz Wojtusiak (2013) proposed three on-line unsupervised learning algorithms Syn Two Moving,Two Moving and One Fixed based on the permutation test statistics. The detection of normal and abnormal payments of medical claims, where data changes over drifts of time. The algorithms were capable of handling univariate and multivariate data streams. The approach of the system is the prediction of medical claim with thenormal and abnormal amount and is carried by two main phases Model construction from historical data and payment prediction [14].

MingJian Tang et.al (2011) proposed a data-driven fraud detection system and an unsupervised and UNISIM, for identifying the non-complaint Medicare claimants in Australia. UNISIM system includes components like feature extraction, building classifier, construction of model and outlier detection. ApproxMap i.e. a density-based clustering algorithm for mining the prescription patterns and calculating the hierarchical edit distance between the prescriptions sequences of different customers. The Hidden Markov Model (HMM) is used to cluster the prescription sequences. The fraudulent consumers are identified by outlierness scores and theseare calculated using Local Correlation Integral (LOCI) i.e. based on the spatial distances [15].

Hossein Joudaki et.al (2016) proposed five-stepdata mining approach to find abuse and fraud behavior in physician’s drug prescription claims. In the process discriminant and cluster analysis for finding fraudulent behavior, he applied thirteen indicators to the sample data. The approach will assist the health insurance organization for streamline the auditing process. In specific to Low and middle-income countries it helps to suspect groups which are involving the abusive activities. The design for the methodology includes understanding the problem’s behavior, understand and preparation of data, indicators for creation and selection, clustering and discriminant analysis. The thirteen indicators were developed among them four for fraud detection and remaining eight for abusive behavior which is in detail discussed as two for the cost issues, four were related to patterns which are frequent visits, and seven are prescription related. The study helps to analyze the physicians instead of claims and objective to improve the quality of auditing practice [16].

Jiwon Seo et.al (2017) proposed a PageRank-based algorithm for identifying the frauds and anomalies in Medicare-B dataset. The Medicare-b dataset contains medical insurance claim information of public includes 162$ billion and 8880,000 providers and 10 million claims in the dataset. The author selected the personalized PageRank algorithm to identify anomalies based on the similarities between the providers and their prescriptions. High values of PageRank implies legitimate treatment by the provider with a different specialty. This algorithm uses a random walk to select a node and follows the similar nodes with similar prescriptions and converges the seed nodes to have high PageRank values. The algorithm repeats over a time and examines the providers with their PageRank values, if it is smaller, then it is likely to be fraudster or anomaly [17].

Shu-Li-Wang et.al (2017) develop an unsupervised approach to identify the health insurance fraud in dental services based on the score of truth worthiness on social networks. A retrospective study of claims of insurer based which tooth they operated by different provider by using rule-based mechanism. The author uses a measure of the truth worthiness score of the dentist, time gap between the first hand and second-hand treatment and cross dentists treatments to find the degree of truth worthiness. Here the fraud or suspicion is defined as claiming behavior for the treatment same tooth of the same patient and is provided by two different dentists. By examining the reliability of truth worthiness scores the claims are processed. The limitation of his work is the use of metric it returns a range numeric values as suspicious scores if it is higher than higher the fraudulent claims and the problem of imbalanced data i.e. the minority of fraudulent dentists against normal dentists so, classifying the trusted and untrusted is found to be difficult in data [18].

From the research work of Tiago Hillerman et.al (2017) proposes a model using different clustering algorithms for analysis of suspicious claims data provided by the healthcare providers. Anomaly detection is carried with the AHP investigation model it includes the first stage to build a model for analysis phase in a hierarchy structure and evaluates the claims processing workflow.The second stage is to select the criteria based on the historical provider behavior, patient characteristics and billing behavior. The third stage to find the suspicious claims it did based on sub-criteria, period, the growth rate of expenses, average visiting frequency and percentage of elderly patients etc. The building of multivariate clustering model by using k-means algorithms, hierarchical clustering and PAM algorithms to find the effectiveness of provider based on their chosen variables. The model simplifies the auditing process, evaluates the tangible and intangible aspects to suspect the entity and facilitates the communication between the analyst and decision makers. His work results cost-benefit approach and new control mechanism to avert the upcoming occurrence of fraud. The extension work is to link medical supplies and high-costtreatments [19].

Mohit Kumar et.al (2010) describes a system which generates explanations for the auditors to correct the claims and reduce the errors for processing the claims of US health insurers. His system includes works collection, feature construction, learning of model and scoring, item scoring, explanation generation, user feedback and user interface. The system carries model learning and selection with SVMs uses categorical data and for creation binary features help to finding the finest subset of data and it is useful to forecast rework for near future. The aggregation score is assigned to claim and is used to classify the unlabeled from the database. Finally, generating explanations decision tree is executed for validating or labeling the claim and produces predictions to the auditor. The inference score is added as a feature vector in classifying the claim. User feedback includes Claim feedback, Activelearning and feature feedback. Claim feedback which is produced to the auditor to rework and results in a binary judgment. Active learningimproves the efficiencyand it is a positive reinforcement to improve over a time. Standard metrics like accuracy and precision useful for comparing different models for exploring the better active learning approaches of the system. His work concludes it definitely reduces errors in processing claims using machine learning techniques and most applicable for large insurance companies in the US. Future implications of the researcher are to rework prediction using the classical ranking technique, handling of a multi-class classification, semi-supervised/Transductive learning and multi-instance learning as to overcome some of the limitations single class classification and space constraints [20].

**Related works on Supervised Fraud detection**

From the works of Richard A Bauder et.al (2016) gave a potential insight for finding the fraud, misuse in the billing procedures of medical insurance claims. The procedures used by the physicians and their claiming behavior is analyzed by multinomial Naïve Bayes algorithm and F-Score as a performance metric[47,48,49]. Provider type data is used for experimental purpose. The researcher had extended work fraud detection in procedures carried, charges and expenses from the providers claim using machine learning methods. Richard A Bauder et.al (2017) with four metrics for performance and class imbalance reduction for selection of samples. The identification of fraudulent providers is done by supervised, unsupervised and hybrid learners by using different learning algorithms. Training and evaluation of work experiment with the use of metrics like BACC (balanced accuracy), MCC (Matthew's correlation coefficient), F- measure and G-measure[47,49]. The work can be extended to improve performance by hyperparameter tuning, improvise the balance of samples to find fraud providers using unsupervised and supervised methods [4] [22].

To improve the eminence of healthcare administrators expenditure the paperworkother of Yang Xie et.al (2015), has developed better strategies for handling a huge scale of health systems in insurance claims. The method is predicting the fixed number of days of every customer based on their previous successive number of years. Hospitalization is a component for health expenditureand for some perspectives estimating the hospitalization days will enable to reduce the cost. The predictive model is constructed in three levelslike customer level information, hospital admission level include detailed information about the customer,provider,and claims, Hospital Procedure claim level describesitem type, accommodation, prosthetics etc. Hospital admission record may contain one or more procedure claim undergone by the customer. Regression trees used to train and computes average predictions of each individual and treebagger is used to estimate the prediction error. The performance metric of the model is indicated by RMSE for predicting accuracy [23].

Robson B et.al (2018) studies presented a constructive model for very large inference nets for the public health data analysis and applications in clinical and of medical claims. The proposed basic hyperbolic Dirac net approach for mathematical analysis for US healthcare. Dirac notation look like XML for probabilistic semantics. His study uses an odds HDN as decision theory, inference net with the use of maximum probability proportion and brute force algorithm with Bayes' rule. The methods and functionalities used for input flow are from SMASH2 (Direct Smash), Dirac Alert and BILL. The work defined wishlist and hitlist structure which are described by the end user which holds the positive and negative forms of attributes. The use of qualitative and quantitative relevant factors for prediction rule generation. Finally, the work proposes a specialized anomalous claims data detection model with use of hitlist and jackknifing to flag the data as normal and anomalous. The model produces weak and strong predictions with jackknife tests. The quality of prediction is assessed with decision signature and decision pattern.


Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

His study concludes fraud detection in health insurance claims like well-intentioned misdiagnosed cases and overuse of medical care resources.Usingthezeta function, HDN and QUEL philosophy his work explored freshmeans of progressions [24].

The Brazilian healthcare market which is proposed by Jackson Cunha et.al (2017) by classification algorithms with various class distributions. His work includes erstwhile analysis related to services by providers and agreeing them to detect fraud and abuse outlines. Measuring the loss of performance loss, prediction performance and recovery for performance are done by the different classifiers metrics with various algorithms Random Forest, C4.5, Ripper, Naïve Bayes and SVM. His work limits to use of three datasets for claim authorization and limited training data sets so insufficient training for learning and if training data set is too large then impractically for the computational process. The enhancement of his work suggests using optimum use of training data set for treatment and learning algorithms in order to overcome the negative learning effects [25].

**Table. 1 Summary on studies related to healthcare fraud in medical claims**

| S.No. | Reference Number | Journal/ Conference | Citations | Data mining Approach | Model proposed | Type of Fraud | Limitations | Future scope |
|---|---|---|---|---|---|---|---|---|
| 1. | [46] | Statistics for Hospital Epidemiology(2004) | 256 | Survey | SEMMA (Sample, Explore, Manipulate, Model, and Assess) model automated surveillance systems. | | | Statistics methods |
| 2 | [45] | IEEE trans,(2005) | 16 | Unsupervised | Multivariate probability density function (pdf) through a hidden sequence of states generated by a Gaussian mixture. | Patient's Medical claims | The patient's profile uses benefits payment as the only feature. So it is One-dimensional case | Up gradation of medical services in health insurance scheme |
| 3. | [13] | 7th Conference of AMCIS(2011) | 21 | Survey | The work describes various types of frauds and fraud medical schemes, application of classification techniques and finds the effectiveness of usage of fraud detection. Discussed strengths and weakness of systems. | | | To discuss the effective impact of current methodologies and tools to detect and prevent fraud in CMS. |
| 4 | [44] | Journal of Computer Science(2006) | 228 | Supervised | DSS ANN early diagnosis of diabetic nephropathy disease. | | | |
| 5 | [43] | Conference on Data Mining(2006) | 85 | Supervised | Multilayer perceptron neural networks (MLP), divide-and-conquer approach. | Patient's Medical claims | | Improved model to estimate costs and savings |
| 6 | [2] | Springer ,Health care Manage and science(2007) | 106 | Survey | Methods to detect fraud , application of statistical methods | | | |
| 7 | [17] | Article(2009) | 55 | Survey | Monitor for evidence of fraud, waste and abuse claims processing | | | Introduce Anti-fraud effort |
| 8. | [41] | KDD(2010) | 43 | Unsupervised | generates explanations for the auditors to correct the claims and reduce the errors for processing the claims | providers claims | | |
| 9 | [40] | Journal of Health care Information Management(2011) | 563 | Survey | Treatment effectiveness, healthcare management for chronic diseases, highlight inappropriate prescriptions , Decision tree | Patient's Medical claims | Costly time consuming, missing corrupted, inconsistent data | Integrate data with text mining |
| 10. | [39] | Proceedings Australia DM'11,(2011) | 15 | Unsupervised | Data driven fraud detection system UNISIM | prescription patterns | | Complex real life interactions |
| 11. | [12] | Elsevier,2012 | 42 | Supervised | The fraudulent activity is observed by analyzing the past insurance claims i.e. by likelihood or probability of each record to be | Insurance Experts fraud patterns | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 12. | [38] | IEEE(2015) | 11 | Supervised | Based on the behavior of overall expenditure of medical for lifestyle-related diseases | Claim amount prediction by medical provider | To find the expenditure correlation between the consultation rate of the cerebral infraction of high blood pressure syndrome. | Enhancement of exploratory visual analysis function in the future. |
| 13. | [37] | Elsevier(2012) | 47 | Supervised | Novel distance based, fraudulent risk of prescriptions. | prescription patterns | | |
| 14. | [28] | IEEE computer society, Article(2013) | 88 | Supervised | It takes measure related to span of stay in the hospital and admission the claims are predicted using. I+Plus. | Provider and Patient claims | Using traditional transaction processing systems it is not able to recover hidden cost. | The claims are checked over time and abstracts the complex relationships. |
| 15. | [36] | Int. journal of Bio science and Biotechnology(2013) | 242 | Survey | Highlights advantages and disadvantages of association, regression, classification etc. based on single disease. | | | |
| 16. | [14] | Computer science,(2013) | 8 | Unsupervised | Drift learning for noisy data, Syn Two Moving, Two Moving and One fixed algorithms for claims processing /classification /regression model. | | Synthetic data sets | Use of adaptive window sizes, optimizing of false positive and true positives rates |
| 17. | [35] | Chemical Engineering Transactions(2013) | 29 | Supervised | Bayesian co-clustering and statistical methods | Provider and beneficiaries fraud | | Integration of information systems and that combine different sources ,dynamic monitoring |
| 18. | [34] | Springer, Health Information Science and Systems(2014) | 1090 | Survey | sophisticated technologies and make informed decisions | | | Privacy, safeguarding security, establishing standards, and governance. |
| 19. | [9] | Global Journal of health science(2015) | 52 | Survey | Proposed a seven step model to find the fraudulent and abusive behavior and pitfalls for various DM models (Supervised, unsupervised and hybrid data mining). | Focuses on subset of claims not on all claims. | Set up for low availability computerized resources. | |
| 20. | [23] | IEEE JBHM, 2015 | 15 | Supervised | Use of Regression decision tree algorithm. | Provider claims | Missing data could reduce predictive power | Requires improvement of prediction accuracy |
| 21. | [33] | Elsevier(2016) | 12 | Supervised | Finds the uncertainty and variability of disease, age groups and other information with use of bagged DT and MCC(Matthews correlation coefficient) | Patient's Medical claims | | Probable predictive power embedded was not investigated within the timing of claims |
| 22. | [32] | Elsevier(2016) | 5 | Supervised | The patient suffering with allogeneic HCT and chemotherapy and the investigation of claims data related to the costs and service utilization. | Administrative claims | | To build model for that is economic for analysis of HCT and related treatments for hematological malignancies. |
| 23. | [31] | Elsevier(2016) | 7 | Supervised | Classification for assessing the health insurance claim based on product value indicators and service quality indicators | Patient's Claims reimbursement | | Outlines the contributions, or cost-sharing methods such as co-payments |


Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 24. | [4] | IEEE Conference,(2016) | 7 | Supervised | The use of multinomial Naïve Bayes algorithm for calculating recall ,precision, F-score with 5-fold cross-validation. | Physician claims | | To generalize the model which works for physicians to find anomaly detection methods. |
| 25. | [16] | IJHPM,(2016) | 13 | Unsupervised | Application of discriminant analysis and identify the suspect physicians, and assess the validity of the clustering approach. | Physician claims | | |
| 26. | [29] | Science direct Journal of Formosan Medical Association,(2016) | 4 | Supervised | Studies the relation and distinguishes, the services of Medical between subgroups of the elderly population.. | Elderly people medical claims | | |
| 27. | [19] | Elsevier,(2017) | 10 | Unsupervised | Analytic Hierarchy Process model (AHP) multi measures for identifying the suspect entities for subsequent auditing. | Provider claims | | Expanding of claims linked to high-cost treatments and medical supplies with the possible use of text mining techniques. |
| 28. | [11] | IEEE(2017) | | Unsupervised | PageRank-based algorithm to recognize abnormalities/frauds in healthcare data. | Provider's Claim | | |
| 29. | [25] | IEEE(2017) | | Supervised | Claim authorization process and prediction by various classification algorithms | Providers and patient's claims | limited to smaller datasets | Larger datasets with effective machine learning algorithms |
| 30. | [22] | IEEE(2017) | 3 | Hybrid/Mixed | Discusses how to employs supervised, unsupervised and hybrid learners for training and evaluation using f-measure and g-measure. | providers claims | Difficulty in finding fraud in specialized provider type having general specialties. | Model performance improvisation by hyper parameter tuning. |
| 31. | [18] | Elsevier(2017) | 2 | Unsupervised | A retrospective study of claims of insurer based which tooth they operated by different provider by using rule based mechanism. | Patient's Medical claims | It uses a metric it returns a range numeric values as suspicious scores. If it is higher than higher the fraudulent claims and problem of imbalanced data i.e. the minority of fraudulent dentists against normal dentists so, classifying the trusted and untrusted is found to be difficult in data. | |
| 32. | [30] | Elsevier,(2018) | | Supervised | The model of HTM(Hierarchical topic model), Tag-HTM for hierarchical categorization of diseases | Prescription patterns and costs | | Limited to some features.(Excludes demographic features) |
| 33. | [21] | Elsevier,(2018) | | Supervised /Unsupervised | Aggregate data analysis, group wise feature selection method that addresses | | | |
| 34. | [24] | Elsevier, Computers in biology and medicine(2018) | | Supervised /Unsupervised | Basic hyperbolic Dirac net approach for mathematical analysis for US healthcare. Use of qualitative and quantitative relevant factors for prediction rule generation. Model produces weak and strong | specialized anomalous claims | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | predictions with jackknife tests. | | | |

## IV. CONCLUSION

This part of the survey demonstrates various perspectives of data mining approaches for finding fraudulent activities. As there is massive data in health care systems, where it involves millions of people having different attributes for each tuple and data with various characteristics. Forming the communities based on various data mining models. To attain objective for fraud detection and prevention involves the domain experts and statistical analysts for choosing the sample size, top priorities for the type of fraud, which leads to more financial loss as the data includes diverse information with certain constraints. As the data related to medical health care for fraud detection mainly exists at the insurance carriers which includes government and private health companies and so, collecting data is a tricky task. The raw data is mostly from the insurance claims data of physicians, data of prescription given by physician and bills and transactions involves insurance subscribers and service provider. The characteristics of data describe the behaviors of both parties which involves fraud where each claim have their distinctive identifiers for insurance subscriber and service provider. With the help these identifiers. It is possible to get behaviors over a period of time.

Most research works focused on the kind of data mining approach and which are suited to carry the fraud and abuse detection in healthcare data. Whereas, fraud detection is the major task towards the claiming behavior, payments by insurers, unduly increasing the cost of treatments and improper procedures etc., while the major part of the works concludes inclusion of IT-enabled auditing tools and sophisticated machine learning techniques will challenge the potential problems in finding the abuse and fraud behaviors.

This paper recommends five general steps which are required to mine the healthcare claims and to detect fraud and abuse. From the reviews of various researchers, our proposed workis defined as follows:

1) Initial preprocessing stage

2) Analysis stage- To build a model in a hierarchy structure and evaluates the claims processing workflow.The basic experimental model which evaluates the dataset and Identifies the important attribute of data which discriminates the behavior of the whole data set[19][20].

3) Development stage-Application of data mining algorithms and use of statistical measures to extract new features and which assists to mine abnormal or abusive behavior[21][22].

4) Identifying the unusual data with extracted features, investigating and label them as outliers[14][15].

5) We recommend our model will help in finding the fraudulent claiming behaviorsfor particular procedures which are suited to any size datasets where applicable to medium and high-income insurance companies.

Our paper concludes, the performance of technique be determined on the type of dataset and experimental model. As, on no single data mining approach which gives dependable outcomes for all types of healthcare data. For actual use of data mining in healthcare organization there is a need to incorporate the security to data among different parties. As there vast improvement to access data over internet. There is a need to have standard approach to organize, access and to secure healthcare data.

## REFERECES


1. MiklosVasarhelyi, Qi Liu, “A survey and a clustering model incorporating Geo-location information”, 29WCARS, Nov 2013.
2. Jing Li, JianjunShi Jionghua Jin and Kuei-Ying Huang, “A survey on statistical methods for health care fraud detection”, Springer Science, May 2009.
3. D.Thornton, P.Schoutsen J.vanHillegersberg and R.M.Mueller, “A Multidimensional Data Model and Analysis Techniques for Fraud Detection, Procedia Technology”, vol. 9, pp. 1252-1264, 2013.
4. R. A. Bauder, A. Richter , M. Herland and T. M. Khoshgoftaar, “Predicting Medical Provider Specialties to Detect Anomalous Insurance Claims”, 2016.
5. L. K. Branting, T. Champney and F. Reeder, J. Gold, “ Graph Analytics for Healthcare Fraud Risk Estimation”, pp. 845-851, ICTAI, 2016.
6. Aarti M. Karandikar Shivani S. Waghade, “A Comprehensive Study of Healthcare Fraud Detection based on Machine Learning”, Vol.13(6), pp.4175- 4178, 2018.
7. Ebru Aydogan, Duman Seref, Sagıroglu, “Health Care Fraud detection Methods and New Approaches” UBMK’17, pp. 839-844, 2017.
8. “The Problem of Health Care Fraud: A serious and costly reality for all Americans”, Report of National Health Care AntiFraud Association (NHCAA). 2005.
9. Joudaki, H., Geraili, B., Minaei-Bidgoli, B., Mahmoodi, M., Arab, M and Nasiri, M, Rashidian A., “A review of literature: Using data mining to detect health care fraud and abuse”. Global journal of health science, 7(1), 2015.
10. Karthiresane T, S Gunasekaran and Fasela V S, “Health Insurance Claim fraud detection: A survey”, IJLTET, Vol.5(2), pp.43-51, 2015.
11. Seo, Jiwon, and Ofer Mendelevitch. “Identifying frauds and anomalies in Medicare-B dataset”. Engineering in Medicine and Biology Society (EMBC), IEEE, 2017.
12. Melih Kirlidog and Cuneytasuk, “A fraud detection approach with data mining in health insurance”, Elsevier, vol.62, pp.89-994, 2012.
13. Travaille, Dallas Thornton, Roland M Muller and Jos Van Hillegersberg, “Electronic fraud detection in U.S”., Proc.7th Americans conference on information systems, pp.1-10, 2011.
14. CheNgufor and JanuszWojtusiak, “Unsupervised labeling of data for supervised learning and its application to medical claims prediction”, Computer science, vol.14 (2), pp.191-214, 2013.
15. Tang, MingJian, U. Mendis, B. Wayne Murray, Sumudu, D, Yingsong Hu, and Alison Sutinen. "Unsupervised fraud detection in Medicare Australia." Conference proceedings -Vol.121, pp.103-110, 2011.
16. Joudaki, Hossein, Bijan Geraili, Behrouz Minaei-Bidgoli, Arash Rashidian, Mahmood Mahmoodi, Mohammad Arab and Mahdi Nasiri,. “Improving fraud and abuse detection in general physician claims: a data mining study”. IJHPM, Vol. 5(3), pp.165.-172, 2016.
17. JiwonSeo and OferMendelevitch, “Identifying frauds and anomalies in the medicare-B dataset”. IEEE, pp.3664-3667, 2017.
18. Shu-Li-Wang, Mei-Fang Wu,FanWu, Hao-Ting Pai, Chen_LinLi, “The evaluation of trustworthiness to identify health insurance fraud in dentistry”, Elsevier, vol 75, pp:40-50,2017.
19. Tiago Hillerman, Joao Carlos F.Souza,Ana Carla B.Reis,Rommel N Carvalho, “Applying Clustering and ahp methods for evaluating suspect healthcare claims”, Elsevier, vol.19,pp.97-111,2017.
20. Mohit Kumar et.al. “Data mining to predict errors in health insurance claims processing”, KDD’10, ACM, pp.1-9,2010.
21. Kang.S and Song.J, “ Feature selection for continuous aggregate response and its application to auto insurance data”, Expert Systems with Applications, 93, pp.104-117,2018.



Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

22. Richard A Bauder and T. M. Khoshgoftaar, “Medicare fraud detection using machine learning methods”, IEEE-conference, pp. 858-865, 2017.
23. Yang Xie,Gunter Schreier,David C W.Chang,Sandra Neubauer,YingLiu,Stephen Nigel H Lovell and j Redmond, “Predicting Days in Hospital Using Health Insurance Claims”, IEEE, JBHI, Vol.19(4), pp.1224-1233, 2015.
24. S Boray and Robson B, “Studies in the extensively automatic construction of large odds-based inference networks from structured data”, Vol.95, pp.147-166, 2018.
25. Jackson Cunha, Pedro Santos Neto, Ricardo Lira Rabelo and Andre Macedo Santana, “ Investigating the effects of class imbalance in learning the claim authorization process in the Brazilian health care market”, IEEE, pp.3265-3271, 2017.
26. Viveros MS, Rothman MJ, Nearhos JP, “Applying data mining techniques to a health insurance information system”. Conference-proceedings, 286-294, 1996.
27. G. H. Sekharan and P. Dora, “Healthcare Insurance Fraud Detection Leveraging Big Data Analytics”, IJSR, vol. 4, pp. 2073-2076, 2015.
28. B. Arunasalam and U. Srinivasan, “Leveraging big data analytics to reduce healthcare costs”, vol. 15, pp. 21-28, 2013.
29. Hsu, Yi-Ping Hsu and Wen-Chin, “ Patterns of outpatient care utilization by seniors under the National Health Insurance in Taiwan”, pp. 325-334, 2016.
30. Shin, Su-Jin, Il-Chul Moon, Minki Kim, Sungrae Park, and Je-Yong Oh, “Hierarchical Prescription Pattern Analysis with Symptom Labels”.2015 IEEE, pp.178-187, 2015.
31. Nsiah-Boateng, Eric, et al. “Value and service quality assessment of the National Health Insurance Scheme in Ghana-evidence from Ashiedu Keteke District”, Vol.10, pp.7-13, 2016.
32. Preussler, Jaime M., et al, “Administrative claims data for economic analyses in hematopoietic cell transplantation: challenges and opportunities” , Vol.10, pp.1738-1746, 2016.
33. Xie.Y, Schreier, M, Liu, Neubauer, S.G, Hoy, Lovell N. H and Chang, D. C, “Analyzing health insurance claims on different timescales to predict days in hospital”, JBI, 60, pp.187-196, 2016.
34. Raghupathi, Viju Raghupathi and Wullianallur, “Big data analytics in healthcare-promise and potential”. Health information science, 2014.
35. Ekina, Tahir, et al. “Application of bayesian methods in detection of healthcare fraud” Transaction 33, 2013.
36. Tomar, Sonali Agarwal and Divya, “A survey on Data Mining approaches for healthcare", International Journal of Bio-Science and Bio-Technology 5.5, pp. 241-266, 2013.
37. Aral, Karca Duru, et al. "A prescription fraud detection model", Computer methods and programs in biomedicine, 106 ( 2012), 37-46.
38. Toyoda, Shuichi, and Noboru Niki. “Visualization-based medical expenditure analysis support system”. EMBC- International Conference of the IEEE, 2015.
39. Tang, MingJian, et al. “Unsupervised fraud detection in Medicare Australia” Proceedings in Data Mining Conference, Vol.121, 2011.
40. Gerald Tan, Koh and Hian Chye,. “Data mining applications in healthcare”. JHIM Journal19.2 pp.65-72, 2011.
41. Mohit, Kumar, Zhu-Song Mei. and Rayid Ghani, “Data mining to predict and prevent errors in health insurance claims processing”. Proceedings with conference. ACM, 2010.
42. Morris, L. “Combating fraud in health care: an essential component of any cost containment strategy”. Health Affairs, pp.1351-1356, 2009.
43. Ortega, P.A., Ruz, G.A., and Figueroa, C.J. “A Medical Claim Fraud/Abuse Detection System based on Data Mining” Case study, DMIN, 6, pp.26-29, 2006.
44. Wasan, S.K., and Kaur, H. “Empirical study on applications of data mining techniques in healthcare”, Journal of Computer science, pp.194-200, 2006.
45. Tsoi, A.C., Hagenbuchner, M. and., Zhang, S ,“Pattern discovery on australian medical claims data-a systematic approach”, IEEE Transactions, (10), pp.1420-1435, 2005.
46. Obenshain, M.K., “Application of data mining techniques to healthcare data”. Hospital Epidemiology”, 25(8), pp.690-695, 2004.
47. S.Sivakumar, S R Nayak, Ashok Kumar, S. Vidyanandini, G Palai, “An empirical study of supervised learning methods for Breast Cancer Diseases", International Journal for Light and Electron Optics, 175, pp.105-114, 2018.
48. P.Ganesan, S.Sivakumar, and S.Sundar, "A Comparative Study on MMDBM Classifier Incorporating Various Sorting Procedure," IJST, Vol 8(9), 868–874, 2015.
49. P.Ganesan, S.Sivakumar, and S.Sundar, "An Experimental Analysis of Classification Mining Algorithm For Coronary Artery Disease", IJAER, Vol 10(6), pp. 14467-14477, 2015.