applied sciences

MDPI

Article

Medical Health Benefit Management System for Real-Time Notification of Fraud Using Historical Medical Records

Irum Matloob [1,*], Shoab Khan [1], Habib ur Rahman [2] and Farhan Hussain [1]

[1] Department of Computer and Software Engineering, College of Electrical and Mechanical Engineering (CEME), National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan; shoabak@ceme.nust.pk (S.K.); farhan.hussain@ceme.nust.pk (F.H.)

[2] Shifa International Hospital, Islamabad 44000, Pakistan; hrfacc@gmail.com

* Correspondence: irum.matloob@ceme.nust.edu.pk or irum.matloob@ceme.nust.pk

Received: 26 June 2020; Accepted: 22 July 2020; Published: 27 July 2020

Abstract: This paper presents a novel framework for fraud detection in healthcare systems which self-learns from the historical medical data. Historical medical records are required for training and testing of machine learning models. The main problem being faced by both private and government health supported schemes is a rapid rise in the amount of claims by beneficiaries mostly based on fraudulent billing. Detection of fraudulent transactions in healthcare systems is a strenuous task due to intricate relationships among dynamic elements including doctors, patients, service. In light of aforementioned challenges in health support programs, there is a need to develop intelligent fraud detection models for tracing the loopholes in procedures which may lead to successful reimbursement of fraudulent medical bills. In order to address the issue of fraud in healthcare programs our solution proposes a framework based on three entities (patient, doctor, service). Firstly, the framework computes association scores for three elements of the healthcare ecosystem namely patients, doctors or services. The framework filters out identified cases using association scores. The Confidence values, after G-means clustering of transactional data, are computed for each service in each specialty. Rules are generated based on the confidence values of services for each specialty. Then, an evaluation of identified cases is done using rule engine. The framework classifies cases into fraudulent activities based on the similarity bit's value. The validation of framework is performed on local hospital employees transactional data which includes many reported cases of fraudulent activities in addition to some introduced anomalies.

Keywords: anomaly; association rules; association score; clustering; fraud; outlier

1. Introduction

'Fraud' and 'abuse', these two phrases are generally used to identify the major medical reimbursement issues that defeat the ultimate objective of a valid claim. We divide the healthcare frauds into two major categories, service_availing patterns and service_providing patterns. Any fraud can occur, either in the service_providing patterns or in service_availing patterns. Figure 1 explains these two categories of the healthcare frauds. The service_availing patterns capture all the services availed by the patients, duplication of either services (actually not availed) or claims against those services. In simple words, a misrepresentation of the services (or products) for which, the bills are generated but actually not availed. For example, an insurance claim provided by the patient can be inconsistent with his age or gender. There is a possibility that one patient is availing the same service again and again or he/she is availing the service less frequently. In such a case, the frequency of the visits of patients to the hospitals or doctors is either quite high or low. The service_providing patterns

refer to the misrepresentation of facts by the doctors, pharmacies or hospitals. There is a possibility that these service_providers generate duplicate bills for the same provided service. The doctors or hospitals can prescribe unnecessary treatments to the patients; the pharmacies can charge patients twice for the same medicine whereas the doctors can prescribe or perform unnecessary procedures and the providers may allow the medical card's misutilization.

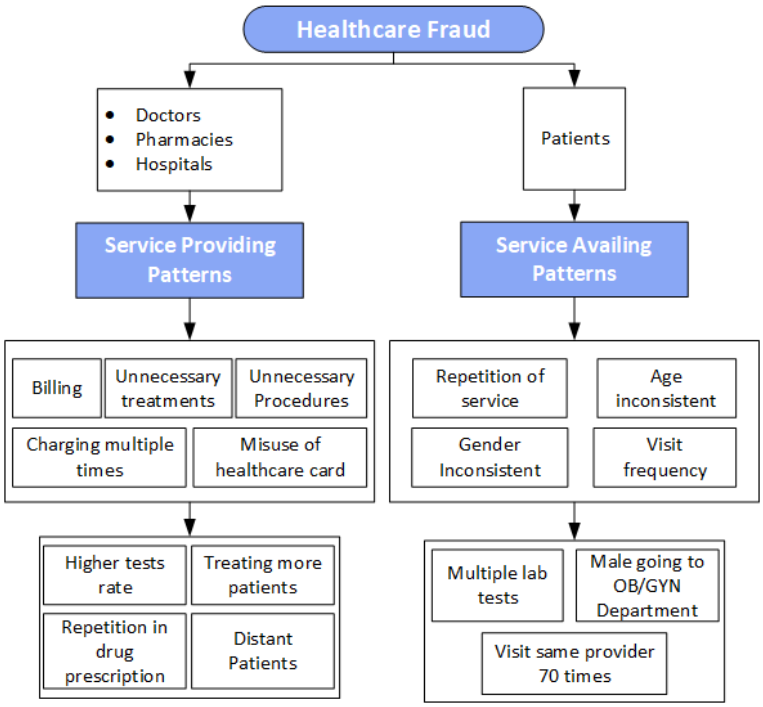


Figure 1. The Figure 1 explains two categories of healthcare frauds: Frauds in Service availing patterns(Patients) and Frauds in Service providing patterns(Doctors, pharmacies, hospitals).

Though many companies normally maintain their 'Special Investigation Departments' to control all the frauds and abuses in the re-imbursement of the medical bills but this is not enough to fulfil the purpose. Such departments get the guidelines from multiple sources and apply 'Conventional Surveillance Techniques' [1]. Whenever these departments detect any fraudulent payments, they proceed for the recovery of funds and then try to introduce the controls to avoid a future occurrence of such misrepresented billings. Once any claimed case is identified as a fraud/abuse, it can be recognized as an identified pattern. Such identified patterns are then utilized to make the adjustments in the billing policies of the existing system in order to prevent the reoccurrence of fraudulent activities. This type of approach commonly known as 'pay and chase', is not an efficient manner of detecting a fraud as it only generates an extra expense [2]. It is of partial use against the healthcare fraud cases because there are high degrees of variations in the clinical practices and billing patterns due to the complex healthcare services. For example a variation can be noticed in the doctors fee structures despite the fact that they are working in the same specialized departments. Many studies have demonstrated variations as high as 400 percent in the frequency of the major procedures among different doctors of the same hospital. There are four categories of the claim analysis. The first one is claim-centric which identifies whether the provided services are according to a patient's age, gender and diagnosis. The second is member-centric which identifies whether the provided services are according to the specialty of the doctor. The third one is provider-centric which identifies whether the claimed services are provided by the specific hospital and the last category is the 'network analysis' which is based on the combination of the member-centric and provider-centric analysis [3]. Our research is focused precisely on claim-centric and member-centric. In recent past, several studies proposed techniques to develop fraud detection systems. Many of these studies used

payment-based analysis to detect frauds. They use one of the healthcare elements to identify any one type of healthcare fraud. To best of our knowledge no one considered delivered or availed service as separate element. Whereas our proposed framework detects fraudulent activities, using all the three main elements namely doctor, patient and service. The important part of the framework is the rule engine, which process over five years original transactional data of employees of a local hospital, for generating rules. Moreover, a self-learned fraud detection system detects patient, doctor and service level frauds.

The fraud-related claims in healthcare are the sources of burden and inconvenience to the overall society. A fraud in healthcare, affects both, the public as well as private sector employers in the form of high-cost over-runs. There are many victims of healthcare frauds who are exploited by the unnecessary treatments. In some cases the patient's data is compromised to generate any fraudulent claims. It will be a meaningless statement if we say that the healthcare fraud is not a crime or there are no victims of this fraud. Accordingly, a detection of fraud in healthcare is a hot topic of research, nowadays. There is a need to cut down this increasing cost as the victims of the healthcare fraud are none other than a common man. In most of the countries including Pakistan, the government has just initiated medical support programs through several national-level initiatives. One of these initiatives is the establishment of Prime Minister Task Force on IT and Telecom in 2018 to lay down the foundation of the data standards and annotations for incorporating the improved plans in healthcare service delivery to the common man. Our work is part of this program, proposing a framework that can be adopted for this national initiative. The major concern is to reduce/prevent the chance of fraudulent activities in such programs. This can only be achieved by implementing different adaptive-modelling techniques for detecting fraud through the healthcare data. For this we have utilized last five years insurance claim data of employees of one of the largest and well-equipped hospital of Pakistan. They provided us sufficient details for supporting this National level objective. According to the provided statistics each day thousands of patients visit this hospital and it has 62 different specialties.

Research Contributions

In recent years, the focus is more on fraud detection in the healthcare as the people in well-developed countries think that the fraud increases an overall expenditure and makes the health insurance problematic for the genuine people. In most of the developing countries, the government has started medical support programs and if such programs face any victimization of the healthcare fraud then there will be no support for genuine patients. This paper presents a novel framework for the fraud detection in healthcare; which considers all three main elements, namely, Patient (service-consumer), Doctors (service_providers) and Services (lab tests and treatments). Our proposed framework provides following significant contributions required in any health care fraud detection scheme.

- The framework provides a self-learned knowledge base system, on the original five years transactional data of a local hospital.
- The novel concept of generating association scores between doctors, patients and services is introduced. The association scores are computed based on frequency of visits between the above mentioned elements and used these association scores to detect anomalies.
- Another novel idea of generating confidence values of all services in each specialty of a local hospital is introduced. As per domain knowledge only *Cardiologist* can recommend *ECG* whereas in real life even an *ENT* specialist can also recommend it, framework computes confidence value of service named *ECG*, have in *ENT* specialty. Similarly, even a *peadiatrician* can recommend *kidney ultrasound* and framework computes confidence value of the service named "kidney ultrasound" in peads specialty. There are many other examples of this. Based on these confidence values, rule engine is generated.
- Another contribution is that this work is part of the national medical support program. We consider a private hospital as our pilot project because in our country due to lack of resources, the electronic health records are not well maintained in the public sector hospitals.

Whereas private sector hospitals are using the automated and autonomous Electronic Health Record Systems and the availability of the patient's data from the private sector is also better. For this reason, we consider the transactional data of a private sector hospital. The public sector programs normally run parallel to a private sector but this research is representing the private sector in the National Health Programs.

2. Related Work

A drastic rise in the healthcare expenditures for the treatment of patients, has led to an introduction of fraud controlling techniques in the hospitals so as to ensure the delivery of more efficient and high quality healthcare services. In many developing countries, there is no such substantial system developed yet, for handling insurance or fraud issues in healthcare. To get better understanding of the fraud detection in healthcare, there is a need to conduct a detailed literature survey of the existing frameworks, techniques and approaches. After conducting a detailed review of the related literature, we get to know that many authors have proposed solutions for the fraud detection in healthcare and have also applied the data-mining approaches and machine-learning algorithms. Large numbers of fraud detection researches are successfully conducted around the globe from local to national levels to control healthcare frauds. The researches vary in data sets, type of healthcare frauds and analysis scale and techniques. The research by [4] is based on a detailed survey of the statistical approaches and these approaches are still being applied for the identification and classification of fraud in healthcare. Another research study, related to the application of supervised and unsupervised learning techniques for healthcare fraud, is conducted [5]. Fraud detection systems are implemented in many other industries and detailed survey is performed by [6]. Ortega [7] designed a system which applied multi-layer perceptron neural networks on the data of Chilean private health insurance company to detect the fraudulent activities; the detection rate of this system is 75 frauds per month. Another framework is proposed [8], which introduced an adaptable model using clinical ways for an automatic fraud detection. The framework for the fraud detection using unsupervised learning for a detection of the outliers in Medicaid insurance-claimed data is proposed [9]. Qi Liu [10] considered a clustering model which is based on the geographical location of Medicaid service_providers and clients to identify fraudulent claims. Moreover, a detection of fraud without considering the roles of the providers and clients is proposed [11]. It is the machine learning based system which involves the hierarchical processing along with assigning the weight to actors, the expectation maximization clustering technique is applied to find out the related groups of the actors. Thorton [12] applied the multidimensional data models and approaches for the prediction of fraudulent claims in Medicaid and the proposed system detected fraud cases. Many recent studies have utilized Public Use Files (PUF) data from CMS for detecting any fraudulent activities using the data mining techniques [13–19]. All the researches have focused on 'PART-B' of this above-mentioned data (PUF). Many statistical techniques are also used to generate decision rules and k-means clustering is applied on a time series-based insurance claim data for the identification of anomalies and outliers. These disease-based outliers are used to detect the fraud related activities [20].

1. Most of the researches related to an anomaly detection in the healthcare, have considered the clinical processes for a particular disease and utilized prior knowledge and applied the unsupervised models [21–23].
2. Many researchers have focused on the statistical financial data and performed analytics using a variety of tools. Fuzzy and Neurofuzzy analysis is performed in the multiple researches for extracting interesting patterns [24–26].
3. Many frameworks for the fraud detection are proposed and the focus of the authors is on the correlation of the medicines, diseases and patients. Frauds are detected by assigning weights to highly correlated data. Many authors have utilized the concept of 'graph theory' for connecting the patients, diseases and medicines. Most of the times, the studies are supplemented with the prior knowledge of the medicines that were being used for the various diseases and they

established a correlation between the reference set (the original knowledge) and the candidate set (the extracted knowledge) [27,28].

4. To identify the joint fraudsters, the clustering technique is applied. The similarity adjacency graph is used along with group mining for distinguishing the normal behaviour from abnormal behaviours [29]. The treatment model for different diseases, that of, assessing the doctors' trustworthiness, which is one of the critical metrics for detecting fraud at the provider level, is introduced [30]. The association rule mining is a very important technique which generates rules for the frequently occurring items. This technique is being utilized in many previous researches for generating rules from the domain knowledge provided by the domain's experts. This technique generates the rules out of which some are significant and some are insignificant [31]. There are two most useful parameters to analyze the strength of the association rules namely: confidence and support [32–34]. The characteristics like uniqueness, understandability, applicability and reliability for assessing the generated rules are discussed [35]. The classification of the insurance claims are performed by using the Genetic Support Vector Machine [36]. Table 1 is providing detailed comparison between the exisiting systems and the proposed framework.

All aforementioned researches focus on disease correlations and medications whereas our proposed framework generates associations between the doctors, patients and services. Most of the existing studies use the domain knowledge to make the knowledge base but our system learns knowledge from the five years transactional data of a local hospital using the machine learning techniques. Based on this knowledge base, we classify the fraud cases. Most of the previous researches are based on the financial analysis for the detection of fraudulent activities but our research identifies anomalies using the association scores and performs the rule-engine analysis for the identification of the fraud cases. We analysed some of the most relevant researches with respect to data mining techniques used to detect types of frauds in Table 1. The comparative analysis highlights that most of the researches lack inclusion of all three key elements (doctors, patients, service) during analytical processing of data. The payment based analysis is utilized to detect patient level frauds and medication/disease associations are analyzed for detecting doctor level frauds. The most critical element missing in all these recent researches are services, which are either provided or availed.

Table 1. Comparison with Existing Literature.

Frameworks and References	Data Mining Approach	Type of Detected Fraud	Applied Data Mining Technique (s)
GSVMs [36]	hybrid	classifying insurance claims	Genetic support vector machines
Medical provider specialty predictions for anomaly detection [14]	Supervised	Physician related frauds	Multinomial Naïve Bayes
Fraud detection and frequent pattern matching [20]	Unsupervised	Disease based anomalies/frauds or period based claim related frauds	K means clustering
Fraud detection using outlier predictor in health insurance data [27]	Hybrid	Disease, medication related frauds	Discrimination rule based outlier analysis using clustering and graph theory
Healthcare fraud detection based on trustworthiness of doctors [30]	Hybrid	Provider (doctor) related fraud	Graph based mining Frequent mining algorithms
Fraudulent claims detection from expected payment deviations [17]	Supervised	Medicine payment related frauds	Regression models used

Table 1. *Cont.*

Frameworks and References	Data Mining Approach	Type of Detected Fraud	Applied Data Mining Technique (s)
Predicting medical provider specialties to detect anomalous insurance claim [15]	Supervised	Fraudulent payments detected in dermatology and optometry	Bayesian inference, using probabilistic programming
Medical school training relate to practice evidence from big data [13]	Unsupervised	Unsupervised Dental service provider related frauds	Fisher–Yates distribution analysis K-means clustering Gcross algorithm
Interactive machine-l earning-based electronic fraud and abuse detection system [11]	Interactive machine learning	Prescription based abnormal behaviour	Pair wise comparison expectation maximization (EM)
Outlier-based Health Insurance Fraud Detection [12]	Unsupervised	Dental provider related frauds	Multi-dimensional data models Multivariate Clustering
A Survey: Healthcare fraud detection [10]	Hybrid	Rehabilitation, Septicaemia Pneumonia, payment related fraud detection	Geo-location Cluster analysis
Knowledge discovery from massive healthcare claims data [19]	Hybrid	Providers Related Frauds Social network	Social network analysis methods
Predicting healthcare fraud in Medicaid [9]	Hybrid	Patient related frauds Physician related frauds	Data models for patient claim and physicians.

3. Material and Methods

3.1. Dataset Details

The analysis is conducted on five years [2013, 2014, 2015,2016, 2017, 2018, 2019] insurance claim transactional data of a local hospital. These are hospital employees who are availing insurance policies provided by hospital management. Based on the designation, insurance policies are allocated to each employee. The size of transactional data is shown in Table 2. The initial framework is proposed in [37].

Table 2. Attributes in Dataset.

Attributes	Value
Unique number of serviceIds	1206
Unique number of Doctors	486
Unique number of specailityId	62
Total number of transactions	441,506

The set of attributes which are providing details about the availed and provided services are shown in Table 3.

The framework involves an implementation of the three phases for detecting fraudulent activities:

- Association scores generation and threshold application
- Rule generation engine
- Similarity Function

We have implemented the fraud detection system by incorporating the above-mentioned three phases. Detecting a fraud from the healthcare data is actually an identification of outliers from such records. In the first phase, we identified the "outliers" and "need to be investigated" cases. In the second phase, we implemented rule engine for further analyzing the identified cases from the first phase. In the third phase, we checked each current transaction against the generated rules. The proposed framework is depicted in Figure 2, the association between the doctors, patients and services are computed and whenever a case of fraud is identified, the rating score of that element,

gets reduced. Based on the number of visits, the association scores are found and these are giving an in-depth understanding of the behaviour of each element.

Table 3. Each transaction's Attributes in data set.

Attributes	Data Types
MRNO	Varchar (255)
Gender	Char
Date of birth	Date
Employee id	Varchar (255)
Department name	Char
Relation	Char
Serviceid	Varchar (255)
Service description	Char
Doctor	Varchar
Specialty	Varchar
Amount	Money
Category	Char

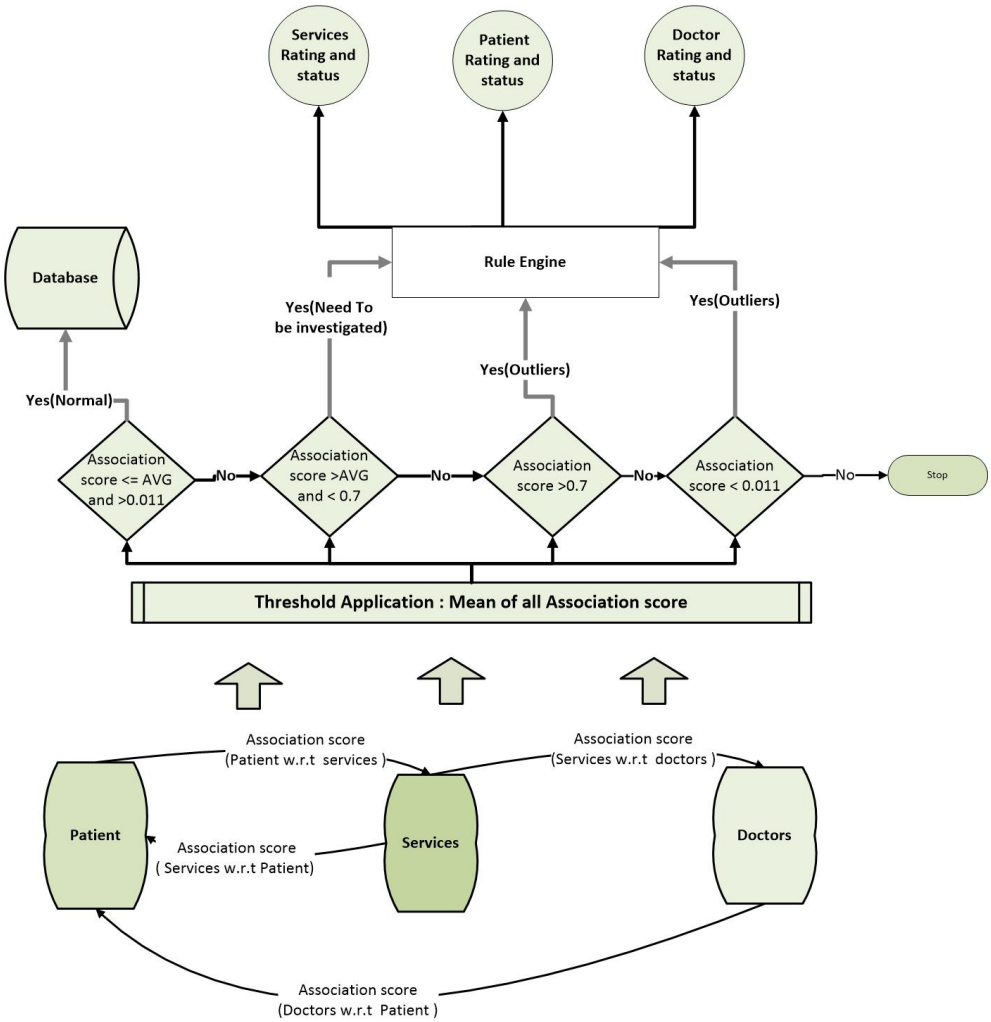


Figure 2. This figure depicts functionality of proposed fraud detection model. First association scores are computed among services, doctors and patients then based on association scores cases are forwarded to Rule engine for further processing.

3.2. Association Scores Computation and Threshold Application

The three main elements of the proposed framework are patients, providers (doctors, pharmacy, hospitals) and services. These three elements are actually associated with each other. There is a need to find out the association score of each element with another element. The association scores are computed based on the frequency of visits or frequency of the prescriptions. If a patient visits frequently to avail a specific service (e.g., X-rays, ECGs). In this case, a patient is prescribed X-rays again and again from one doctor. This is considered as outlier. We compute association scores based on the frequency of the patient visits to the providers and services. The purpose of this step is to forward only those patient records to rule engine, which are identified as the "outliers" or "need to be investigated". We computed the association scores by using Equations (1)–(4).

- Doctor (Association score) Y is computed by denoting i as number of times patient P_k checked by doctor D_j and D_n is representing total number of patients checked by doctor D_j. As shown in Equation (1)

$$Y = (i/D_n) \tag{1}$$

- Patient with services (Association score) is also computed by denoting m as number of patients availed service S_h and S_x is representing number of times patient P_k availed S_h service. (for all patients)

$$S_p = (m/S_x) \tag{2}$$

- Service with Doctor (Association score) is also computed by denoting T as number of times doctor D_j prescribed service S_h and S_n is representing number of times all doctors prescribed service S_h (for all services)

$$S_d = (T/S_n) \tag{3}$$

- Patient (Association score) is computed by denoting G as number of times doctor D_j examined patient P_k and P_n is representing total number of patient P_k visits.

$$F = (G/P_n) \tag{4}$$

The association scores are between 0 and 1. After the computation of the association scores, we calculate threshold by computing an average of all the association scores for each provider, service or patient. All those transactions which are less than average but greater than the minimum threshold and equal to the average, are considered as the normal cases whereas all the association scores which are greater than the average but less than the maximum threshold, are considered as the "need to be investigated". The minimum threshold value and maximum threshold value is set up to identify the outliers. The minimum threshold indicates that anything that has happened just once is an anomaly. It means that if any patient, visits a provider only once that can be an anomaly (or any doctor prescribing any service just once to only one patient). Thus, we have kept the minimum threshold as 0.011. Similarly, we have chosen the maximum threshold by considering the fact that if a patient is visiting the same doctor and out of a total of his 100 visits, he visits the same doctor more than 70 times, there could be an anomaly. That is why, we have kept the association scores greater than 0.7, as the maximum threshold. All those association scores which are less than the minimum threshold and greater than the maximum threshold, are identified as the outliers. The flowchart of this phase is shown in Figure 3. Patient association scores are denoted as F, doctor's association scores are denoted as Y and association scores of services with respect to doctors or patients are denoted as S_p and S_d respectively. We set threshold for all association scores as discussed above. Figure 3 explains the flow of the first phase of proposed framework. A hash algorithm is applied for the de-identification of patient records. The variables Y, S_p, S_d and F holds association scores of patient with respect to doctors, service with respect to doctor, service with respect to patient and doctor with respect to patient. The threshold is computed separately for each type of association scores. The variable Z is

representing function or container which holds values for all four types of association scores after computation of threshold values. We apply four checks on Y, S_p, S_d and F separately. Based on these check 'outlier' and 'need to be investigated' cases are identified. The Rating score is initially set as 100 for each element of the framework and after first phase rating score is updated based on the occurrence of identified cases. Each time identified case is found, rating of that particular element is decremented as shown in Figure 3. The cases of the "need to be investigated" and "outliers" are analyzed in the second phase.

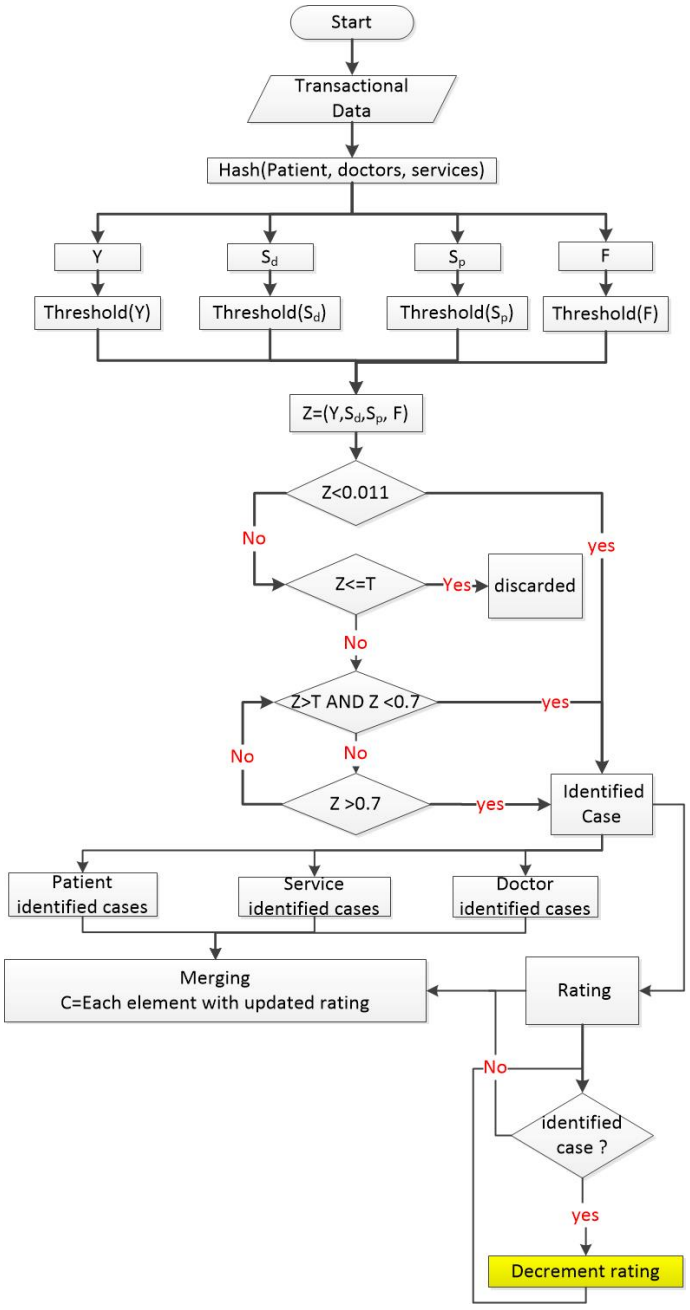


Figure 3. This figure depicts functionality of first phase. Association scores are computed among three elements.

3.3. Rule Engine Generation

The second phase of the proposed framework generates rules for each specialty of the local hospital. It is already mentioned that the proposed framework is validated on an original data of local

hospital. There are 62 specialties in this hospital. Following are the two main tasks which are executed under this phase:

- We perform hashing on the patients data by assigning separate identification numbers to every service, every doctor and specialty.
- Clustering the transactional data and generated association rules.

During cluster analysis we found outliers within different clusters. For this purpose, after applying the clustering to transactional data we applied concepts of support and confidence to these generated clusters. We applied three different clustering algorithms on this transactional data: Gmeans, Xmeans and Fuzzy Cmeans. G-Means clustering algorithm, is an extension of KMeans. The G-means algorithm is density based clustering; it tries to find a subset of data that fits a Gaussian distribution. G-means executes k-means, increments value of variable k hierarchically until the data assigned to each centroid are Gaussian. It is identified by research that Gmeans is improved form of clustering which has provided an intrusion detection with the high Detection and the low False Positive Rate. This technique can approximate number of the clusters in the considered data and initialize the centroids which results in fast convergence of algorithm [38]. The X-means [39] executes K-means multiple times and during each run, it takes local decisions whether to create a subset of current centroid or not and this splitting decision is taken by the computation of the Bayesian Information Criterion (BIC) [40]. We have compared the generated clusters of all three algorithms in Table 4. We took one cluster and computed Mean of that cluster. Centriods are generated by Fuzzy C-means, G-means and X-means. Pick the centriod generated by each algorithm, which is closest to the computed Mean. Actual center is the computed mean of selected cluster. Computed center is the centriod computed by the algorithms. Difference is the subtraction of actual center from algorithm computed centriod. Based on our analysis, it is found that the G-means clustering is more efficient as compared to the other two clustering techniques for this transactional data.

Table 4. Comparison of Clustering Algorithms.

Actual Center		Technique	Computed Center		Difference	
590	800.745	Fuzzy Cmeans	586.9281	977.215	3.071856	−176.47
590	800.745	Xmeans	476.5837	613.5539	113.4163	−26.6258
590	800.745	Gmeans	586.7819	968.0782	3.218119	−167.333

3.4. Rule Engine Algorithm

Following steps are used to generate rule engine

3.4.1. Step 1

Perform de-identification of patient records.

- Each patient assigned $patient_n$ unique number
- Each doctor/specialization assigned $doctor_n$ unique identifier
- Each service assigned $service_n$ unique identifier

3.4.2. Step 2

Grouping of patient records based on the specaility_id from where they availed service. Guassian based clustering is used for the identification of clusters as shown in Figure 4.

$$Support(S_h) = Count(S_h)/cluster_n \tag{5}$$

$$Confidence(S_h \cap D_j) = Support(S_h \cap D_j)/Support(S_h) \tag{6}$$

where $cluster_n$ is the total number of elements in clusters. Transaction c_n is representing transactions of the patients P_k who are identified as two separate cases namely "Need to be Investigated" and "outliers". All transactions which are identified with these labels are transferred for further analysis, to the rule engine. We computed confidence value for each service within clusters. We apply threshold on confidence values for all members within clusters and all members whose confidence values are on boundaries are identified as anomaly. The flowchart for second phase is shown in Figure 4 which shows how clusters are processed to generate rules. We find support count of each specialty D_j in all clusters and then find support count of each service S_h for this specialty D_j.

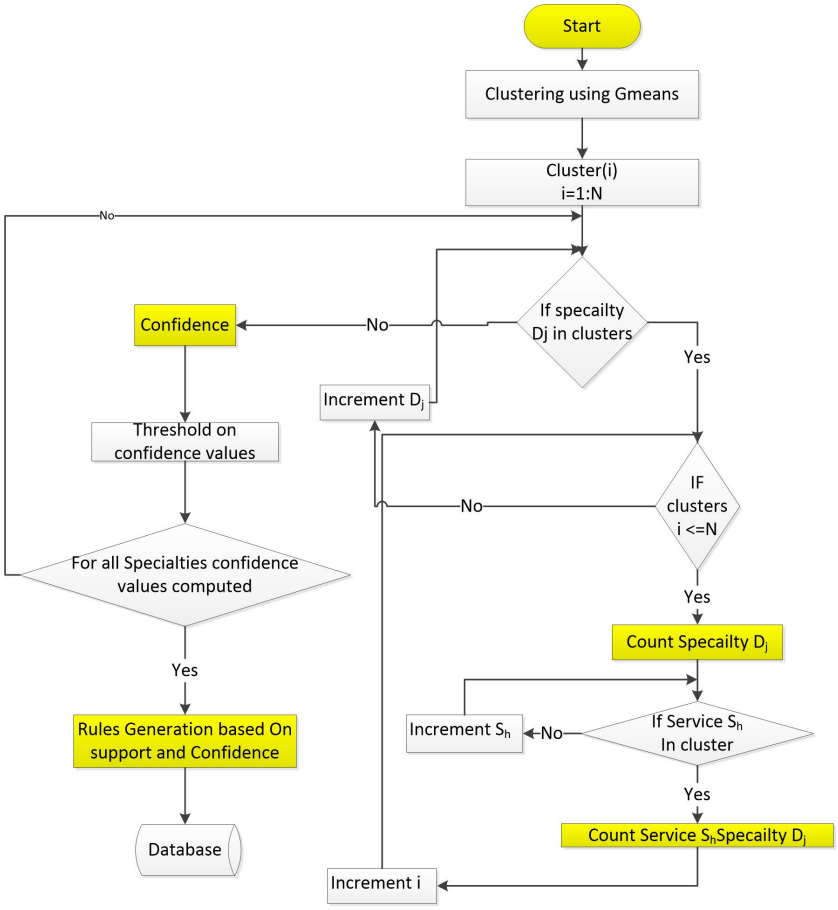


Figure 4. The rule engine is computing confidence values for all services in all specialties.

Finally, these support counts are used for computing confidence values. The last condition is for checking whether confidence values are computed for all specialties or not. Based on the computed confidence values, rules are generated which are stored in database for the third phase. Figure 5 describes the complete fraud detection system. In Figure 5, there are three main elements, and each element is receiving transactional data from different hospital servers. Each element (Patient, provider, service) has its own storage. Association scores are computed between each pair namely service with respect to doctor, service with respect to patient, patient with respect to doctor, and doctor with respect to patient. Once the association scores are computed and thresholds are applied, we get set of identified cases. Transactions are identified in two cases "outliers" or "need to be investigated". The rating of each element (Patient, provider, service) whose transactions are found to be suspicious will be decremented. These cases are used as an input to the rule engine. The Rule engine further analyzes the transactions and if these cases are detected as fraud then rating score of involved element, will remain same otherwise rating score will be updated. Basically set of rules are generated for each specialty_id. Whenever any patient visits the hospital for availing the particular set of services, system first checks

which specialty_id patient visits, and then evaluates according to the rules already computed for each specialty_id. The third phase of the proposed framework is shown in Figure 6, Similarity function is used for computation of similarity between current transaction c and generated rule R. Similarity bit is denoted by a and Similarity Function denoted by H

$$Similarity\ function\ H = R \cap c \tag{7}$$

Similarity bit is a equal to 1, if after the similarity computation the size of the input transaction c is equal to size of similarity function H, and if after similarity computation the size of the input transaction c is not equal to size of similarity function H then similarity bit will be equal to zero. If the similarity bit is not 1, then transaction will be marked as a fraud. Otherwise it will be marked as normal.

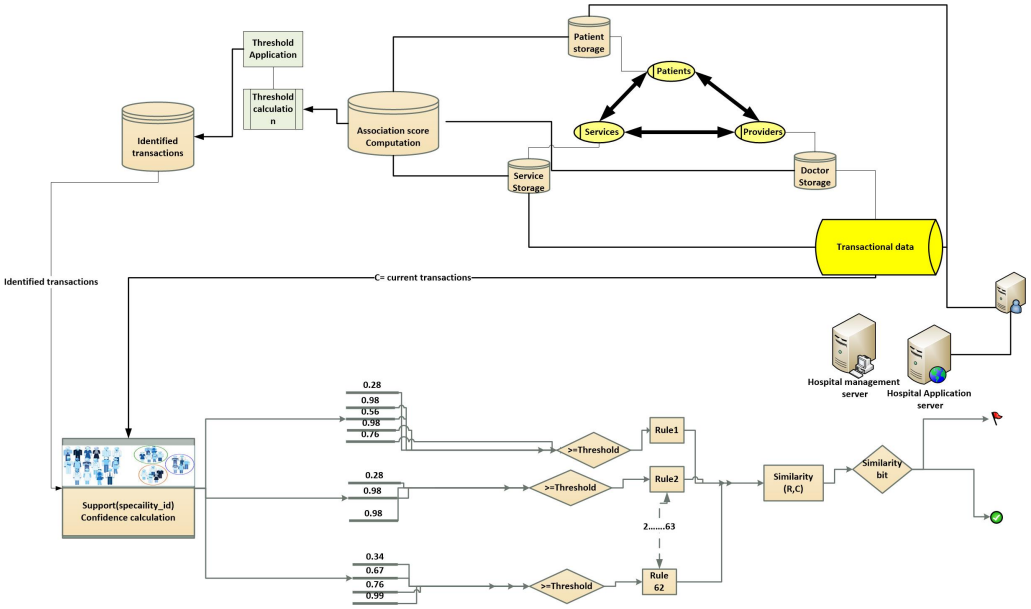


Figure 5. Detailed visualization of fraud detection system.

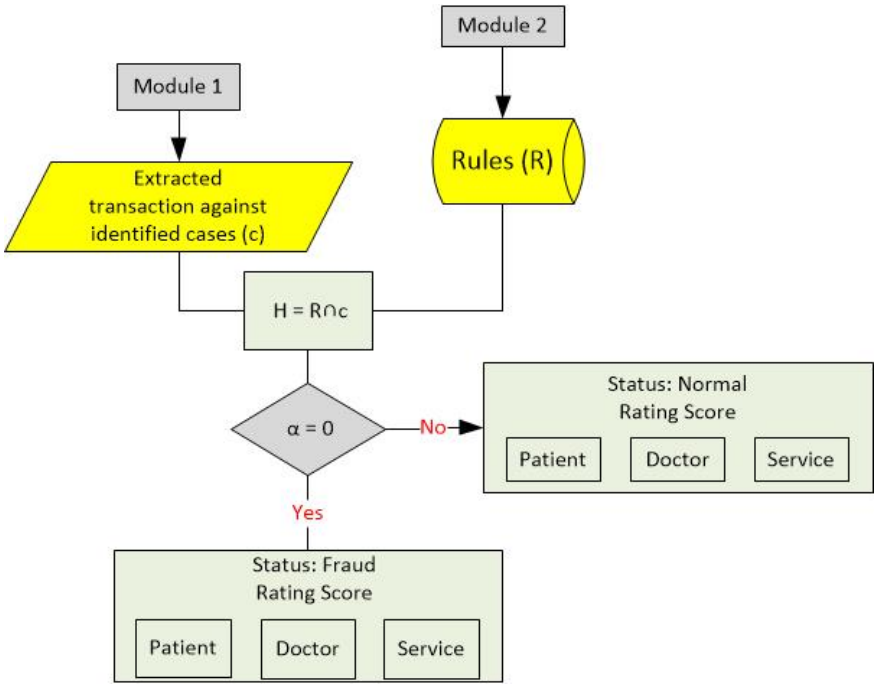


Figure 6. Third phase of Fraud detection model is described.

4. Results and Discussions

4.1. Case Study

The five years (2013, 2014, 2015, 2016, 2017, 2018, 2019) annotated insurance claim transactional data of employees of a local hospital is considered for this analysis. The addressed problem is the constant increase in employees insurance coverage expenditures in each year as depicted in Figure 7 and it can be easily predicted as exponential increment in coming years due to increase in healthcare frauds. Fraud detection model is applied to analyze this dataset and only few results are shown to add better understanding of the work and therefore only subsets of results are shown in the figures.

Figure 7. Yearwise insurance amount utilization.

4.1.1. First Phase

In the first phase, the association scores are computed between each pair of elements. Few of the cases are shown in this section to explain how association scores are actually computed. In this phase we identified two separate cases:

- Outliers
- Need to be investigated

Association score among service Optical Coherence Tomography OCT scan and patients are shown in Figure 8. Total 21 patients avail this service of OCT scan and an average of all association scores is 0.052. We set this average value as a threshold. It can be seen from the Figure 8 that two patients are identified as "need to be investigated", and rating of this service is decremented to 98 from 100. Total score of rating is 100. Similarly, association score for all services and patients are computed in the same manner and the rating score is also adjusted accordingly.

Service With Respect to Patient

OCT Scan

Search

- Service **OCT Scan** availed by **21** Patients
- **Average:** 0.0526
- **Rating:** 98

Mr No	No Of time Service Availed	Association score	Status
510051	2	0.0952	Need To Investigate
510230	1	0.0476	Normal
76279C	1	0.0476	Normal
16511564	1	0.0476	Normal
959490	1	0.0476	Normal
956429	1	0.0476	Normal
17498990	1	0.0476	Normal

Figure 8. Service with respect to Patient association scores.

Figure 9 explains doctor_id association scores with respect to patients. Total 36 number of patients are examined by the doctor_id 131. Two cases are identified as "need to be investigated" and the rating score of this doctor is decreased to 98 from 100.

Figure 9. Doctor with respect to Patient association scores.

Figure 10 shows association scores of services with respect to doctors. Service Routine Electroencephalogram "EEG" prescribed by 50 different doctors and six cases are identified as 'need to be investigated". The threshold value is 0.0476. The Rating score of this service is 94, which is decreased by 6. Complete output is shown in Appendix A.1.

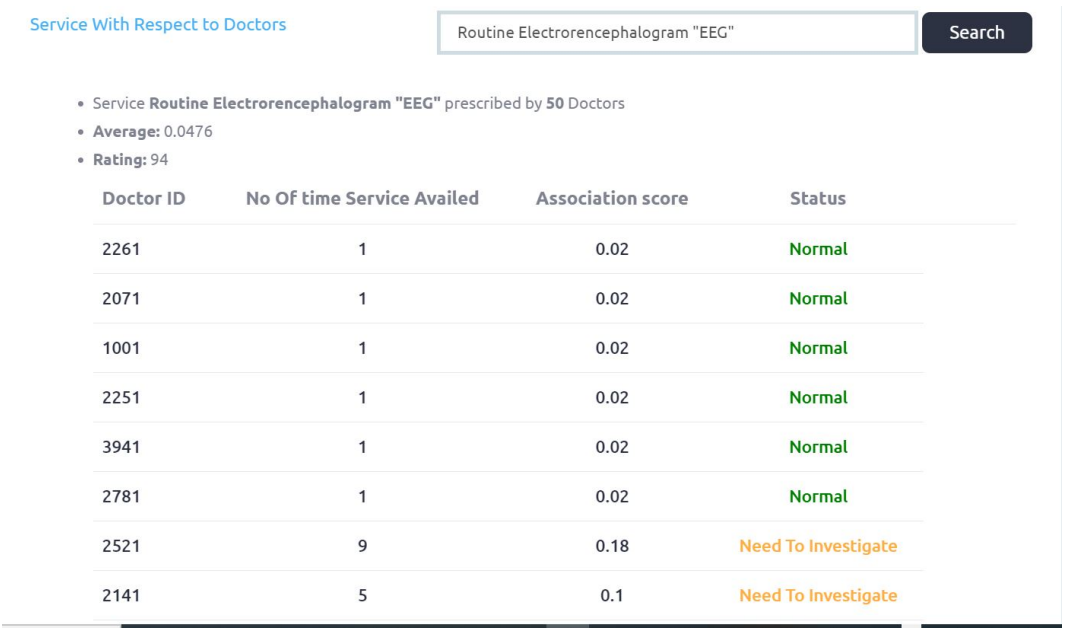


Figure 10. Service with respect to doctor association scores.

Figure 11 is explaining patients with respect to doctors association scores, Patient MR_no is 959705 visited 126 times to the hospital and he is examined by 12 different doctors. This patient visited doctor_id 1511, forty eight times. Four cases of "need to be investigated" are identified. The Rating score of this patient is decreased to 96.

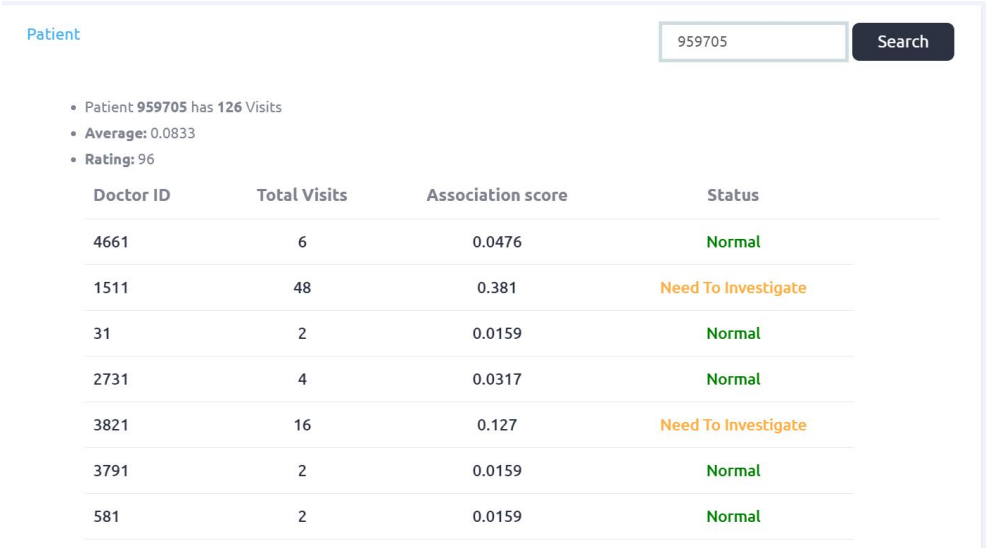


Figure 11. Patient with respect to doctor association scores.

4.1.2. Second Phase

All those records which are identified as "outlier" or "need to be investigated" are forwarded to the Rule Engine for a further investigation. Total 62 association rules are generated from this data set and separate rule is generated for each specialization and specialty_id is used to represent identifier for each specialization. Rule engine basically generate rules that describe which specialization can provide which specific service. We generated rule by computing confidence values for each service in particular specialization. By using this knowledge, we can evaluate each transaction whether it is normal or fraud. This can be done by applying the Similarity function. We have selected specialty *Urology* with specialty_id: 620 and we can get all services which are provided by this specialty_id as shown in Figure 12. It can be seen that there is confidence value of each service for each specialty_id.

Figure 12. The specialty_id 620 (Urology) is selected and confidence values of each service availed/provided in this specialty.

The relationship between service and specialty is depicted in Figure 13, in this plot confidence values of all service_ids for the specialty_ids are depicted. The value of confidence has provided us with an estimation, that what is the probability of prescription of considered service in this specialty_id. Based on this estimation, resources can be also allocated and budget can also be planned. Table 5, is depicting confidence values of few services for different specialties.

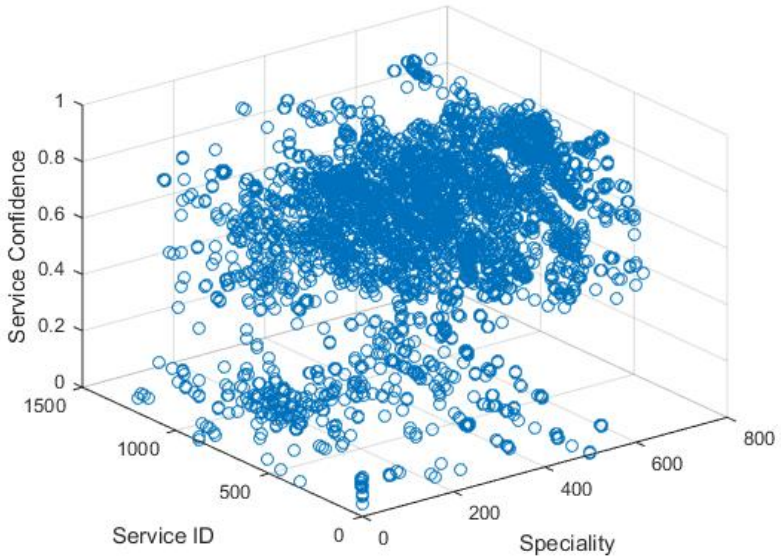


Figure 13. Scatter plot for all services confidence for all specialty.

Rule only contain service_ids whose confidence values are above 0.001 or user can define the threshold depending upon their scenarios. The rule generated by the rule engine can be explained with the help of example. Consider specialty name Pediatric Cardiology. Table 6 is showing services availed from this specialty and confidence values of these services are also provided.

Table 7 is depicting rule for this specialty. If in any transaction Abdomen upper service is availed. The similarity function first check whether this service is present in the rule of Pediatric Cardiology as shown in Table 6. This case is identified as fraud, and if in any transaction service is availed whose confidence value is less than 0.001, from the considered specialty it will also be identified as fraud and passed to analyst dashboard for further investigation. The rules are generated from the medical historical data.

Table 5. Services and their Confidence values for Pulmonologist, Orthopedic, Pediatrician, Neurologist, Urologist.

Specialty Name	Speciality_id	Service_id	Service Name	Service_Confidence
Pulmonologist	530	189	Bronchoscopy	0.08866593
Pulmonologist	530	301	Chest Xray 1 View	0.08277595
Pulmonologist	530	295	Humerus 2 Views	0.009384047
Pulmonologist	530	302	Chest Xray 2 Views	0.01450867
Pulmonologist	530	307	Thoracic Spine 2 Views	0.000765366
Pulmonologist	530	335	Liver & Gall Bladder	0.8180593
Pulmonologist	530	381	HBs Ag	0.9869222
Pulmonologist	530	17	Exercise Tolerance Tests	0.6025923
Pulmonologist	530	511	Chem 7	0.9742271
Orthopedic	410	133	Cast Up.Ext., Long Arm	0.9624833
Orthopedic	410	142	Splint Short Leg	0.6123378
Orthopedic	410	768	Foot Both 4 View	0.009955376
Orthopedic	410	312	Sacrum And Coccyx 2 Views	0.9772097
Orthopedic	410	280	Shoulders, both, 3 views	0.05383435
Orthopedic	410	245	Cervical Spine without Contrast	0.9958633
Pediatrician	490	337	Both Kidneys or GenitoUrinary Tract	0.992217
Pediatrician	490	342	Pelvis	0.9919294
Pediatrician	490	385	Hepatitis E Virus Ab (HEV)	0.9929713
Pediatrician	490	631	Rx Services-OPD	0.9913415
Pediatrician	490	732	RF Quantitative	0.9989808
Pediatrician	490	409	Vitamin B 12	0.5485373
Pediatrician	490	407	Growth Hormone	0.5484943

Table 5. *Cont.*

Specialty Name	Speciality_id	Service_id	Service Name	Service_Confidence
Neurologist	320	963	Cervical Spine-2 Views (AP, Lat)	0.001586921
Neurologist	320	18	Carotids, Ultrasound Doppler	0.6055873
Neurologist	320	59	OPG "Orthopentomogragh".	0.6076537
Neurologist	320	119	Pattern ShiftVisual Evo Pot."PSVEP"	0.7515727
Neurologist	320	510	Chem 7	0.9912891
Neurologist	320	513	LIPID Profile(HDL,LDL,Chlstrl,Trig)	0.9996347
Urologist	620	337	Both Kidneys or Genito Urinary Tract	0.200903
Urologist	620	359	Transplant Kidney With Doppler	0.911211
Urologist	620	330	Pelvis 1 View	0.030101
Urologist	620	308	Lumbar Spine 2 Views	0.231112
Urologist	620	301	Chest Xray 1 View	0.210001
Urologist	620	352	Abdomen Upper	0.230011
Urologist	620	276	Elbow 3 views	0.021121
Urologist	620	245	Cervical Spine without Contrast	0.951121
Urologist	620	229	Liver Dynamic (3 Phase)+Abdomen	0.210011
Urologist	620	211	Brain/Head (3-D Imaging)	0.200011
Urologist	620	228	Renal/Kidney(3 Phase)+Abdomen	0.234442
Urologist	620	200	Urinary Catheter without Consultant	0.900112

Table 6. Services and their confidence values for Pediatric Cardiology.

Service ID	Service Description	Confidence Values
513	LIPID Profile(HDL,LDL,Chlstrl,Trig)	0.8660617
9	ECG 12 Lead	0.5245009
11	Fetal Echo/Pediatric Echo	0.6903811
12	ECHO 2D & M Mode With Doppler	0.5506352
14	ECHO F/U with in one week	0.5219601
15	ECHO Transesophageal	0.6388385
16	24 Hour Holter	0.5183303
631	Rx Services-OPD	0.9303085
1055	Doctor-ID 16-OPd Initial Visit	0.814519
1056	Doctor-ID 16-OPd Follow-up Visit	0.7604356
1119	Doctor-ID 21-OPd Initial Visit	0.9771325
1120	Doctor-ID 21-OPd Follow-up Visit	0.8598911
21	24-Hour Ambulatory B.P. Monitor	0.6896552
342	Pelvis	0.9814882
487	BUN	0.9876588
488	Creatinine Serum	0.9938294
489	Uric Acid Serum	0.9967578

Table 7. Rule for Pediatric Cardiologist extracted from Table 6.

Pediatric Cardiologist Rule	If {Service = 513 AND confidence > 0.001 OR Service = 11 AND confidence > 0.001 OR Service = 12 AND confidence > 0.001, OR Service = 14 AND confidence > 0.001, OR Service = 9 AND confidence > 0.001, OR Service = 15 AND confidence > 0.001, OR Service = 16 AND confidence > 0.001, OR Service = 631 AND confidence > 0.001, OR Service = 1055 AND confidence > 0.001, OR Service = 1056 AND confidence > 0.001, OR Service = 1119 AND confidence > 0.001, OR Service = 1120 AND confidence > 0.001, OR Service = 21 AND confidence > 0.001, OR Service = 342 AND confidence > 0.001, OR Service = 487 AND confidence > 0.001, OR Service = 488 AND confidence > 0.001, OR Service = 489 AND confidence > 0.001}

4.1.3. Third Phase

The following example explains how the similarity bit is computed using already generated rule for specialty_id: 620. In Current transaction c, patient is availing three services from specialty_id: 620, it can be seen from Table 5, that there is no service with service_id: 2. Computation of similarity function and similarity bit value generation are shown below. The value of Similarity bit is 0 this means this transaction is a fraud case.

$$c = Transaction$$

$$c = \{1070, 1152, 2\}$$

$$Size(c) = 3$$

$$Similarity\ function = R \cap c$$

$$R = \{1070, 1152\}$$

$$Size(H) = 2$$

if $Size(c)$ and $Size(H)$ are equal then similarity bit will be $a = 0$. If the similarity bit is equal to 1 only then the current transaction is normal. This engine is generated from five years annotated transactional data, so based on the association scores we have identified cases and evaluated them against the rules, from which we have found already tagged fraud cases. After this analysis we have reached to final status and got rating of doctors, patients and services separately.

4.2. Detected Frauds

After the third phase fraud cases have been detected. Now, we are able to check status and rating of each element. As it is already mentioned that due to the large size of data only a subset of records are shown in screenshots to depict our system performance. One of the main point that must be clarified at this level is that we have considered employee insurance claim data, so detected cases are less in number because patients are either employees or their beneficiaries. We discussed detected fraud cases using the screenshots.

Figure 14 depicts the final status and ratings of doctors. It can be seen that three doctors have been identified by our system as a fraud. It can be seen that the rating score of these doctors are also adjusted finally. Doctor_id 2301, 551 and 31 are identifed as the fraud cases. Initial status and rating of the identified doctors, which are generated in the first phase of the proposed system, are depicted in Figures 15–17.

Figure 15 shows association score of doctor_id: 2301. Initial rating of this doctor is 95, as five identified cases of this doctor are forwarded to the Rule engine (second phase). Complete output of Figure 15 is provided in Appendix A.2. Third phase has identified doctor_id: 2301 as a fraud and updated rating score of this doctor is 99 as shown in Figure 14.

Figure 16 shows initial rating of doctor_id: 31 which is −9 (negative). More than a 100 identified cases of this doctor are forwarded to the rule engine (second phase). The doctor_id: 31 is also identified as fraud and final rating is 99 as shown in Figure 14. Figure 17 shows that first phase's initial rating of doctor_id: 551 is 7 and 93 cases of this doctor are forwarded to second phase for analysis. In the third phase this doctor is identified as fraud and updated rating score is 98. Figure 18 shows that eight cases of frauds are identified in service availing patterns. As it is the subset of complete output. We can also check initial ratings and association scores of each of these Patients in first phase as we have checked for the doctors. When a service element is considered, service_id 221 and service_id 250 are detected as fraud. It is shown in Figures 19 and 20.

So we have analysed this transactional data in terms of the three element of proposed framework and detected different number of already tagged fraud cases. The rule engine has been designed on

the basis of five years transactional data, each specialty_id (specialization like cardiology, urology etc.) has a set of services with confidence levels that define rules for it.

Doctor

Doctor ID	Rating	Status
4481	100	Normal
2561	100	Normal
1541	100	Normal
2461	100	Normal
31	99	Fraud
551	98	Fraud
1831	100	Normal
251	100	Normal
2301	99	Fraud
3711	100	Normal

Figure 14. Doctors rating and status.

Doctor 2301 Search

- Dr with id **2301** checked **116** Patients
- **Average:** 0.0588
- **Rating:** 95

Mr No	Total Visits	Association score	Status
06299E	4	0.0345	Normal
17510878	2	0.0172	Normal
20402E	6	0.0517	Normal
17510606	2	0.0172	Normal
17510440	2	0.0172	Normal
958925	10	0.0862	Need To Investigate
388493	8	0.069	Need To Investigate

Figure 15. Doctor_id 2301 initial rating in First phase.

Doctor 31 Search

- Dr with id **31** checked **3982** Patients
- **Average:** 0.0031
- **Rating:** -9

Mr No	Total Visits	Association score	Status
950657	8	0.002	Normal
20684E	10	0.0025	Normal
812895	30	0.0075	Need To Investigate
369351	22	0.0055	Need To Investigate
951623	6	0.0015	Normal
950654	6	0.0015	Normal
951863	10	0.0025	Normal

Figure 16. Doctor_id 31 initial rating in First phase.

Doctor

551 Search

- Dr with id **551** checked **1655** Patients
- **Average:** 0.0041
- **Rating:** 7

Mr No	Total Visits	Association score	Status
DB5717	6	0.0036	Normal
510296	4	0.0024	Normal
511141	6	0.0036	Normal
20604E	4	0.0024	Normal
812899	6	0.0036	Normal
916501	2	0.0012	Normal
01162E	8	0.0048	Need To Investigate

Figure 17. Doctor_id 551 initial rating in First phase.

Patient

MR No	Rating	Status
511142	100	Normal
510173	100	Normal
951863	99	Fraud
42001	100	Normal
18411836	99	Fraud
503304	100	Normal
99185	100	Normal
172950	99	Fraud
320269	99	Fraud
18173525	100	Normal
954646	99	Fraud
432726	100	Normal
18770185	100	Normal
191054	100	Normal
951260	100	Normal
196501	99	Fraud
957321	99	Fraud
BC3525	100	Normal
410377	100	Normal
17511187	99	Fraud

Figure 18. Patients Rating and status.

Service With Respect to Patient

Service ID	Rating	Status
1	100	Normal
1154	100	Normal
1027	100	Normal
221	99	Fraud
480	100	Normal
250	98	Fraud
380	100	Normal

Figure 19. Service wrt Patient Rating and status.

Service With Respect to Doctors

Service ID	Rating	Status
1	100	Normal
1154	100	Normal
1027	100	Normal
3	100	Normal
221	99	Fraud
480	100	Normal
250	98	Fraud
380	100	Normal
381	100	Normal
382	100	Normal

Figure 20. Service wrt doctors Rating and status.

5. Conclusions

Many countries have recently initiated government medical support programs and in such programs there is no tolerance for any fraudulent claims. There is a critical need of system for capturing and identifying fraud cases in day to day transactions in healthcare industry. Lots of research studies have been conducted in last decade but most of them are based on financial analysis and disease/medication analysis. We have proposed framework by considering patients, doctors (providers) and services as main elements. We computed relationships between these elements by calculating association scores. By learning from the historical transactional data, we have generated Rule engine. Firstly, dataset is filtered out based on elements association scores and then forwarded identified cases to the Rule engine for further analysis. The fraud cases are finally identified and the ratings of all three elements are updated after an evaluation from the rule engine. We have validated this framework for detecting fraudulent transactions from annotated local hospital transactional data and successfully identified eight fraud cases along patient element, two cases along service element and three cases along doctor element of proposed System. We communicated our findings to the hospital management.

In future, the proposed methodology can be further improved by extracting sequences of services availed from each specialty using some data mining techniques. Upon finding a set of sequences for every specialty, fraud detection will be more effective.

Author Contributions: Conceptualization, I.M. and S.K.; methodology, I.M. and S.K., software, I.M.; validation, I.M., S.K., H.u.R.; resources, H.u.R.; data curation, I.M.; writing—original draft preparation, I.M. visualization, I.M.; supervision, S.K.; Review and Proofreading F.H. and All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

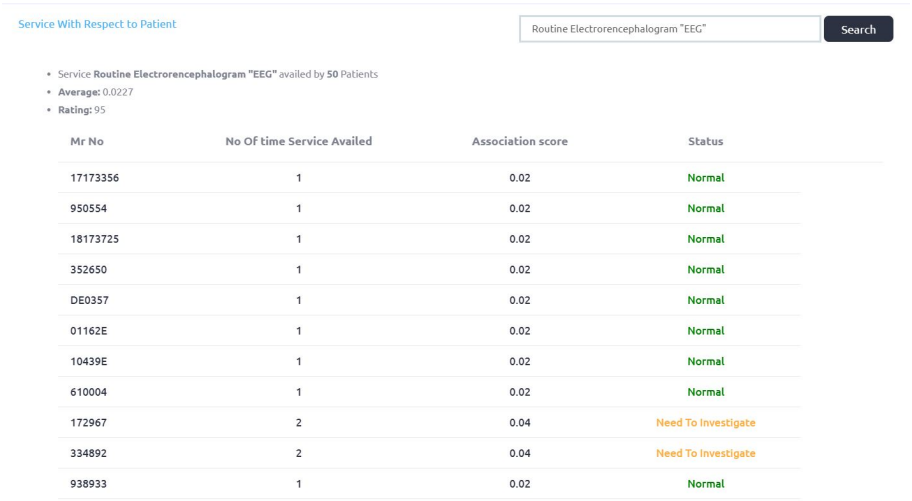
Acknowledgments: Special thanks to shifa international hospital, Islamabad, Pakistan for provding dataset for validation of proposed framework. This research is part of PM Task Force on IT and Telecom initiative for "Sehat Card Scheme".

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

Appendix A

Appendix A.1

Complete output of Figure 10 is shown in Figures A1–A4.



Mr No	No Of time Service Availed	Association score	Status
17173356	1	0.02	Normal
950554	1	0.02	Normal
18173725	1	0.02	Normal
352650	1	0.02	Normal
DE0357	1	0.02	Normal
01162E	1	0.02	Normal
10439E	1	0.02	Normal
610004	1	0.02	Normal
172967	2	0.04	Need To Investigate
334892	2	0.04	Need To Investigate
938933	1	0.02	Normal

Figure A1. Servuce wrt doctor Rating and status.

938933	1	0.02	Normal
950320	1	0.02	Normal
11287	1	0.02	Normal
10276	1	0.02	Normal
510824	1	0.02	Normal
194110	1	0.02	Normal
959771	1	0.02	Normal
951891	1	0.02	Normal
EC4141	1	0.02	Normal
10137	1	0.02	Normal
EC3935	1	0.02	Normal
18174746	1	0.02	Normal
45167	2	0.04	Need To Investigate
953799	1	0.02	Normal
EC3831	1	0.02	Normal
AE9340	1	0.02	Normal
BA6737	3	0.06	Need To Investigate

Figure A2. Servuce wrt doctor Rating and status.

EC3831	1	0.02	Normal
AE9340	1	0.02	Normal
BA6737	3	0.06	Need To Investigate
EC9619	1	0.02	Normal
956849	1	0.02	Normal
954494	1	0.02	Normal
EC7485	2	0.04	Need To Investigate
AC3102	1	0.02	Normal
168677	1	0.02	Normal
955643	1	0.02	Normal
957549	1	0.02	Normal
955723	1	0.02	Normal
17130073	1	0.02	Normal
11232	1	0.02	Normal
18174287	1	0.02	Normal
EC6336	1	0.02	Normal
20493E	1	0.02	Normal

Figure A3. Servuce wrt doctor Rating and status.

11232	1	0.02	Normal
18174287	1	0.02	Normal
EC6336	1	0.02	Normal
20493E	1	0.02	Normal
20186A	1	0.02	Normal
290857	1	0.02	Normal
903527	1	0.02	Normal

Figure A4. Servuce wrt doctor Rating and status.

Appendix A.2

Complete output of Figure 15, is depicted in Figures A5 and A6.

- Dr with id **2301** checked **58** Patients
- **Average:** 0.0588
- **Rating:** 95

Mr No	Total Visits	Association score	Status
06299E	2	0.0345	Normal
17510878	1	0.0172	Normal
20402E	3	0.0517	Normal
17510606	1	0.0172	Normal
17510440	1	0.0172	Normal
958925	5	0.0862	Need To Investigate
18232095	3	0.0517	Normal
388493	4	0.069	Need To Investigate
DC8446	3	0.0517	Normal
958698	4	0.069	Need To Investigate
18510440	1	0.0172	Normal
358395	2	0.0345	Normal
713183	15	0.2586	Need To Investigate

Figure A5. Doctor wrt Patient Rating and status.

20402E	3	0.0517	Normal
17510606	1	0.0172	Normal
17510440	1	0.0172	Normal
958925	5	0.0862	Need To Investigate
18232095	3	0.0517	Normal
388493	4	0.069	Need To Investigate
DC8446	3	0.0517	Normal
958698	4	0.069	Need To Investigate
18510440	1	0.0172	Normal
358395	2	0.0345	Normal
713183	15	0.2586	Need To Investigate
10495	3	0.0517	Normal
950286	2	0.0345	Normal
11387	1	0.0172	Normal
510424	7	0.1207	Need To Investigate

Figure A6. Doctor wrt Patient Rating and status.

References

1. Optum. *The Key to Detecting Fraud and Abuse in Medical Billing*; White Paper 12-28110 04/12; Optuminsight, Inc.: Eden Prairie, MN, USA, 2012.
2. Olsen, L.; Saunders, R.S.; Yong, P.L. *The Healthcare Imperative: Lowering Costs and Improving Outcomes: Workshop Series Summary*; National Academies Press: Washington, DC, USA, 2010.
3. Landon, B.E.; Keating, N.L.; Barnett, M.L.; Onnela, J.; Paul, S.; O'Malley, A.J.; Keegan, T.; Christakis, N.A. Variation in patient-sharing networks of physicians across the united states. *JAMA* **2012**, *308*, 265–273. [CrossRef] [PubMed]
4. Li, J.; Huang, K.Y.; Jin, J.; Shi, J. A survey on statistical methods for health care fraud detection. *Health Care Manag. Sci.* **2008**, *11*, 275–287. [CrossRef] [PubMed]
5. Joudaki, H.; Rashidian, A.; Minaei-Bidgoli, B.; Mahmoodi, M.; Geraili, B.; Nasiri, M.; Arab, M. Using data mining to detect health care fraud and abuse: A review of literature. *Global J. Health Sci.* **2015**, *7*, 194. [CrossRef] [PubMed]
6. Travaille, P.; Müller, R.M.; Thornton, D.; Hillegersberg, J.V. Electronic fraud detection in the us medicaid healthcare program: Lessons learned from other industries. In Proceedings of the 17th Americas Conference on Information Systems, AMCIS 2011, Detroit, MI, USA, 4–8 August 2011.
7. Ortega, P.A.; Figueroa, R.G.A.; Cristin, J. A medical claim fraud/abuse detection system based on data mining: A case study in chile. *DMIN* **2006**, *6*, 26–29.
8. Yang, W.; Hwang, S. A process-mining framework for the detection of healthcare fraud and abuse. *Expert Syst. Appl.* **2006**, *31*, 56–68. [CrossRef]
9. Thornton, D.; van Capelleveen, G.; Poel, M.; van Hillegersberg, J.; Mueller, R.M. Outlier-based health insurance fraud detection for us medicaid data. In Proceedings of the 16th International Conference on Enterprise Information Systems, ICEIS (2), Lisbon, Portugal, 27–30 April 2014; pp. 684–694.
10. Liu, Q.; Vasarhelyi, M. Healthcare fraud detection: A survey and a clustering model incorporating geo-location information. In Proceedings of the 29th World Continuous Auditing and Reporting Symposium (29WCARS), Brisbane, Australia, 21–22 November 2013.
11. Kose, I.; Gokturk, M.; Kilic, K. An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. *Appl. Soft Comput.* **2015**, *36*, 283–299. [CrossRef]
12. Thornton, D.; Mueller, R.M.; Schoutsen, P.; Hillegersberg, J.V. Predicting healthcare fraud in medicaid: A multidimensional data model and analysis techniques for fraud detection. *Procedia Technol.* **2013**, *9*, 1252–1264. [CrossRef]
13. Feldman, K.; Chawla, N.V. Does medical school training relate to practice? Evidence from big data. *Big Data* **2015**, *3*, 103–113. [CrossRef] [PubMed]
14. Herland, M.; Bauder, R.A.; Khoshgoftaar, T.M. Medical provider specialty predictions for the detection of anomalous medicare insurance claims. In Proceedings of the 2017 IEEE International Conference on Information Reuse and Integration (IRI), San Diego, CA, USA, 4–6 August 2017; pp. 579–588.
15. Bauder, R.A.; Khoshgoftaar, T.M.; Richter, A.; Herland, M. Predicting medical provider specialties to detect anomalous insurance claims. In Proceedings of the 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI), San Jose, CA, USA, 6–8 November 2016; pp. 784–790.
16. Bauder, R.A.; Khoshgoftaar, T.M. A probabilistic programming approach for outlier detection in healthcare claims. In Proceedings of the 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, USA, 18–20 December 2016; pp. 347–354.
17. Bauder, R.A.; Khoshgoftaar, T.M. A novel method for fraudulent medicare claims detection from expected payment deviations (application paper). In Proceedings of the 2016 IEEE 17th International Conference on Information Reuse and Integration (IRI), Pittsburgh, PA, USA, 28–30 July 2016; pp. 11–19.
18. Bauder, R.A.; Khoshgoftaar, T.M. The detection of medicare fraud using machine learning methods with excluded provider labels. In Proceedings of the Thirty-First International Florida Artificial Intelligence Research Society Conference (FLAIRS-31), Melbourne, FL, USA, 21–23 May 2018.
19. Chandola, V.; Sukumar, S.R.; Schryver, J.C. Knowledge discovery from massive healthcare claims data. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, IL, USA, 11–14 August 2013; ACM: New York, NY, USA, 2013; pp. 1312–1320.

20. Verma, A.; Taneja, A.; Arora, A. Fraud detection and frequent pattern matching in insurance claims using data mining techniques. In Proceedings of the 2017 Tenth International Conference on Contemporary Computing (IC3), Noida, India, 10–12 August 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–7.
21. Huang, Z.; Lu, X.; Duan, H. Anomaly detection in clinical processes. In Proceedings of the AMIA Annual Symposium Proceedings, Chicago, IL, USA, 3–7 November 2012; American Medical Informatics Association: Bethesda, MD, USA, 2012; Volume 2012, p. 370.
22. Okita, A.; Yamashita, M.; Abe, K.; Nagai, C.; Matsumoto, A.; Akehi, M.; Yamashita, R.; Ishida, N.; Seike, M.; Yokota, S.; et al. Variance analysis of a clinical pathway of video-assisted single lobectomy for lung cancer. *Surg. Today* **2009**, *39*, 104–109. [CrossRef] [PubMed]
23. de Klundert, J.V.; Gorissen, P.; Zeemering, S. Measuring clinical pathway adherence. *J. Biomed. Inform.* **2010**, *43*, 861–872. [CrossRef] [PubMed]
24. Gath, I.; Geva, A.B. Unsupervised optimal fuzzy clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **1989**, *11*, 773–780. [CrossRef]
25. Lenard, M.J.; Alam, P. Application of fuzzy logic to fraud detection. In *Encyclopedia of Information Science and Technology*, 1st ed.; IGI Global: Hershey, PA, USA, 2005; pp. 135–139.
26. Köppen, M.; Kasabov, N.; Coghill, G. *Advances in Neuro-Information Processing: 15th International Conference, ICONIP 2008, Auckland, New Zealand, November 25–28, 2008, Revised Selected Papers*; Springer: Berlin/Heidelberg, Germany, 2009; Volume 5507.
27. Peng, J.; Li, Q.; Li, H.; Liu, L.; Yan, Z.; Zhang, S. Fraud detection of medical insurance employing outlier analysis. In Proceedings of the 2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design (CSCWD), Nanjing, China, 9–11 May 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 341–346.
28. Anbarasi, M.S.; Dhivya, S. Fraud detection using outlier predictor in health insurance data. In Proceedings of the 2017 International Conference on Information Communication and Embedded Systems (ICICES), Chennai, India, 23–24 February 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–6.
29. Sun, C.; Yan, Z.; Li, Q.; Zheng, Y.; Lu, X.; Cui, L. Abnormal group-based joint medical fraud detection. *IEEE Access* **2018**, *7*, 13589–13596. [CrossRef]
30. Cui, H.; Li, Q.; Li, H.; Yan, Z. Healthcare fraud detection based on trustworthiness of doctors. In Proceedings of the 2016 IEEE Trustcom/BigDataSE/ISPA, Tianjin, China, 23–26 August 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 74–81.
31. Hristidis, V. *Information Discovery on Electronic Health Records*; CRC Press: Boca Raton, FL, USA, 2009.
32. Altaf, W.; Shahbaz, M.; Guergachi, A. Applications of association rule mining in health informatics: A survey. *Artif. Intell. Rev.* **2017**, *47*, 313–340.
33. Toti, G.; Vilalta, R.; Lindner, P.; Lefer, B.; Macias, C.; Price, D. Analysis of correlation between pediatric asthma exacerbation and exposure to pollutant mixtures with association rule mining. *Artif. Intell. Med.* **2016**, *74*, 44–52. [CrossRef] [PubMed]
34. Cai, R.; Liu, M.; Hu, Y.; Melton, B.L.; Matheny, M.E.; Xu, H.; Duan, L.; Waitman, L.R. Identification of adverse drug-drug interactions through causal association rule discovery from spontaneous adverse event reports. *Artif. Intell. Med.* **2017**, *76*, 7–15. [CrossRef] [PubMed]
35. Zeng, L.; Wang, B.; Fan, L.; Wu, J. Analyzing sustainability of chinese mining cities using an association rule mining approach. *Resour. Policy* **2016**, *49*, 394–404. [CrossRef]
36. Sowah, R.A.; Kuuboore, M.; Ofoli, A.; Kwofie, S.; Asiedu, L.; Koumadi, K.M.; Apeadu, K.O. Decision support system (dss) for fraud detection in health insurance claims using genetic support vector machines (gsvms). *J. Eng.* **2019**, *2019*, 1432597. [CrossRef]
37. Matloob, I.; Khan, S. A framework for fraud detection in government supported national healthcare programs. In Proceedings of the 2019 11th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), Pitesti, Romania, 27–29 June 2019; pp. 1–7.
38. Zhao, Z.; Guo, S.; Xu, Q.; Ban, T. G-means: A clustering algorithm for intrusion detection. In Proceedings of the International Conference on Neural Information Processing, Auckland, New Zealand, 25–28 November 2008; Springer: Berlin/Heidelberg, Germany, 2008; pp. 563–570.

39. Pelleg, D.; Moore, A.W. X-means: Extending k-means with efficient estimation of the number of clusters. In Proceedings of the ICML: Proceedings of the Seventeenth International Conference on Machine Learning, Stanford, CA, USA, 29 June–2 July 2000; Volume 1, pp. 727–734.
40. Ekina, T.; Leva, F.; Ruggeri, F.; Soyer, R. Application of bayesian methods in detection of healthcare fraud. *Chem. Eng. Trans.* **2013**, *33*. [CrossRef]