IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 19, 2020, accepted July 24, 2020, date of publication August 3, 2020, date of current version August 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3013962

Sequence Mining and Prediction-Based Healthcare Fraud Detection Methodology

IRUM MATLOOB[1], SHOAB AHMED KHAN[1], AND HABIB UR RAHMAN[2]

[1]Department of Computer and Software Engineering, College of Electrical and Mechanical Engineering (CEME), National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan
[2]Department of Cardiology, Shifa International Hospital, Islamabad 44000, Pakistan

Corresponding author: Irum Matloob (irum.matloob@ceme.nust.edu.pk)

This research is the part of Pakistan Government level initiative. Special thanks to Shifa International hospital.

ABSTRACT This article presents a novel methodology to detect insurance claim related frauds in the healthcare system using concepts of sequence mining and sequence prediction. Fraud detection in healthcare is a non-trivial task due to the heterogeneous nature of healthcare records. Fraudsters behave as normal patients and with the passage of time keep on changing their way of planting frauds; hence, there is a need to develop fraud detection models. The sequence generation is not the part of previous researches which mostly focus on amount based analysis or medication versus diseases sequential analysis. The proposed methodology is able to generate sequences of services availed or prescribed by each specialty and analyse via two cascaded checks for the detection of insurance claim related frauds. The methodology addresses these challenges and self learns from historical medical records. It is based on two modules namely ‘‘Sequence rule engine and Prediction based engine’’. The sequence rule engine generates frequent sequences and probabilities of rare sequences for each specialty of the hospital. The comparison of such sequences with the actual patient sequences leads to the identification of anomalies as both sequences are not compliant to the sequences of the rule engine. The system performs further in detail analysis on all non-compliant sequences in the prediction based engine. The proposed methodology is validated by generating patient sequences from last five years transactional data of a local hospital and identifies patterns of service procedures administered to patients using Prefixspan algorithm and Compact prediction tree. Various experiments have been performed to validate the applicability of the developed methodology and the results demonstrate that the methodology is pertinent to detect healthcare frauds and provides on average 85% of accuracy. Thus can help in preventing fraudulent claims and provides better insight into how to improve patient management and treatment procedures.

INDEX TERMS Anomaly, fraudsters, sequence mining, sequence prediction, probability.

I. INTRODUCTION

The fraud and abuse in healthcare systems are becoming crucial problem now a days. Healthcare insurance frauds are critical facilitater for the misutilization of public funds. There are two main categories of healthcare frauds. (I) Consumer related and (II) Provider related frauds. In this article, we consider consumers as patients and providers are doctors, hospitals etc. Consumer related frauds can be in the form of false claims, incorrect medical identity specification (using someone else medical benefit) and visiting multiple physicians to get opinions. Whereas provider related frauds can be in the form of incorrect billing (bill generated for non-availed service), pharmacy related frauds, charging patients by unbundling procedures or charging for expensive services which was not actually performed.

According to an estimate approximately 17 billion to 57 billion funds were misutilized via healthcare frauds under the healthcare supported scheme discussed by [1]. Such critical losses have motivated many researchers to focus on the development of fraud detection models. Most of the conventional fraud detection approaches rely on rule engines, designed by domain experts [2]. These approaches identify normal patients as anomaly because normal patient

The associate editor coordinating the review of this manuscript and approving it for publication was Dian Tjondronegoro.

sometimes deviates from the defined rules. Due to this fraudster gets the opportunity and try to get false benefits. This is the reason that all such approaches have high false positive rate.

Detection of fake patient claims or fraudsters is a non-trivial task and a challenging problem in all healthcare insurance programs due to the following factors :

1) Fraudster behave like normal patients and called Camouflage, defined in [3]. These fraudsters are difficult to identify as they are smart. Therefore, there is a need to design a model which can be used to distinguish normal patients from fraudsters.
2) Due to the longitudinal and heterogeneous nature of healthcare insurance data, fraud detection is exigent task. For example patients claim record consists of multiple services which he/she avail from different specialties of particular hospital along with the service date.
3) With the passage of time and with the advancement of technology, fraudsters are changing their behaviors and techniques for planting frauds. This makes task of the finding unusual patterns hectic.

To examine the aforementioned challenges in healthcare patient insurance claim data, there is a need to design effective methodology which can address all these issues and distinguish normal patients from fraudsters. Many previous researches are conducted to identify insurance claim frauds but most of them focus on either disease and medication related issues or consider one or two specialities for fraud detection. Many developing countries have started government medical support programs. Recently, Pakistan government initiated a first ever medical support program named ‘‘Sehat Card Scheme’’. There is a critical need that this and every other support program must not be affected by insurance claim frauds. By extensive studies over these programs, we observe that there is a dire need to analyse sequence of services which a patient avails from specific specialty of the hospital. For example, if any false claim is generated, patient sequence can be analyzed to detect anomaly or fraud in the process. We need standardized set of patient treatment sequences for each specialty to analyse patient sequences and our initial framework design is proposed in [4].

In this article we propose a novel fraud detection methodology to detect fraud claims in government initiated medical support program. In order to validate our methodology we use employee’s five year insurance claim data of a private hospital. Our proposed methodology is being considered as the pilot module for the above mentioned government level initiative. The main contribution of our research is that the designed methodology generates a set of sequences for each specialty after analyzing five years transactional data. The proposed methodology uses sequence mining and sequence prediction for fraud detection. Sequence mining is performed by creating sequence rule engine which is based on a set of frequent sequences and probabilities of rare sequences for each specialty. The design is capable of detecting anomalies using sequence matching. At initial step, the methodology generates patient time series traces and from these traces, sequences for each specialty are generated. The frequent sequences for each specialty are generated separately by adjusting minimum support value for each specialty based on the volume of transactions conducted in that specialty. The methodology also uses Bayes Theorem which is applied on rare sequences and probabilities of occurence of these sequences are genearted. For the detection of fraud cases, sequence matching of patient sequence is performed against sequence rule engine’s frequent sequences as well as against sequence rule engine’s rare sequences. Fig 1 exhibits the way transactional data is converted to time series traces. These time series traces generate the sequence database. All transactions are first sorted by MR_NO, if there are couple of records with same MR_NO then transactions are sorted by service_date, if there are couple of records with the same service_date then transactions are sorted by transaction_id. Afterwards, the obtained sequences are used by sequence rule engine for the generation of frequent and rare sequences. All the sequences which are not compliant to sequence rules are termed as anomalous sequences. Once we get anomalous sequences from sequence matching function, next step is to further analyse the anomalous sequences. All anomalous sequences are forwarded to prediction based engine (PBE) for in detail analysis. The main goal of second module is to identify the fraud cases for each specialty. The detailed functionality of sequence rule engine and prediction based engine are discussed in section III.

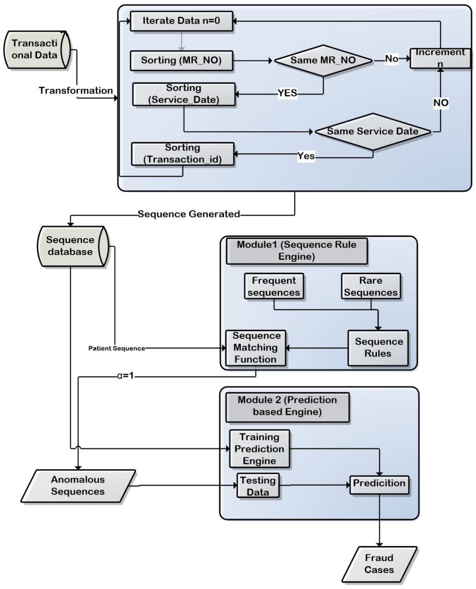


FIGURE 1. **Workflow of fraud detection methodology.**

1) RESEARCH CONTRIBUTIONS

This article proposes a novel fraud detection methodology which holds major contributions for the research field. It provides

1) Considerable understanding of recurring patient visits to each specialty(department) and the extraction of patients visits patterns. These patterns create awareness among doctors and hospital management regarding possible medical services which have been availed on subsequent patient visits in each specialty. It also enables service providers to take preventive actions in the case of anomalous behavior. The availed service could be either regular treatment or special treatment depending on the risk of diseases.
2) Ability to analyse more than 62 specialities of a hospital and can predict sequence and predicted sequence for each specialty separately where as in most of the previous researches authors considered one to two specialties only.
3) Specialty specific sequence generation using sequence mining algorithm.
4) A significance of using patient time series traces which basically depicts healthcare utilization details of patient. With the help of time series traces and above mentioned approaches, the sequence rule engine is generated based on the frequent and rare sequence, due to which, we are able to find out misutilization of health care services. Prediction based engine enable us to identify fraud cases and predict what services a patient can avail in near future, it is basically providing insight for the prediction of clinical events. All findings are communicated to hospital management.

II. RELATED WORK

Various researches have been conducted on health examination procedures. Such studies have attracted more attention as compared to studies which were conducted to predict patient behavior patterns. In general medicine many researches are based on predicting demands of patients. McCarthy et al [5] predicted demands of patients in emergency departments. AlNuaimi [6] performed prediction of demand in healthcare services and Batal *et al* [7] considered urgent care units and discussed predictions of patient visits in these units. These studies highlighted that strategic planning is an important aspect for predicting demand. Liao *et al.* [8] considered that examination of patients visit patterns and health records using data mining techniques is the efficient way of extracting useful information. Koh and Tan [9], Taneja *et al* [10], Ito *et al* [11] applied data mining techniques and algorithms for predicting diseases and for analyzing medical records. Few studies applied association patterns concept on outpatient medical record for predicting patterns of hospital visits. Kontio *et al* [12] utilized machine learning algorithms for predicting patient needs during his hospital visit. Ohara *et al* [13] utilized sequential pattern mining for prediction of diseases. Ou-Yang *et al* [14] perfomed association rule mining on doctors prescriptions. Sequential pattern mining is basically used in medical field to identify frequent patterns or behaviours. According to the most of the previous research studies sequential pattern mining and association rule mining. Both methods are suitable for patient data analysis. Sequential pattern mining used sequential data as the input where as association rule mining generate association among features to predict future visits. Ou-Yang *et al* [3] has combined these both methods for predicting patients future visits. Fraud detection in healthcare insurance has gained significant focus in recent literature. Data mining algorithms have been identified as an immediate solution to healthcare frauds. With the advent of technologies, volume of data has increased drastically and it is not possible to analyse this data using conventional methods. There is a need to apply sequential based algorithms to analyse such data. Operational efficiency can be achieved in improved manner via data mining based anaylsis by [15]. Musal [16] conducted fraud detection via geographical analysis using clustering algorithms. Yang and Hwang [17] proposed framework based on clinical pathways for automatic generation of fraud detection models. Graph theory based analysis [18] conducted for identification of fraud cases in healthcare insurance records. Knowledge is extracted by searching out relationships between doctors, patients, pharmacies and insurance claims. Based on extracted knowledge anomalous relationships are identified. The case study of Chinese healthcare insurance claim considered in [3]. Users relationship are not possible as users can enter their claims on single platform and there is no possibility of interactions with other stakeholders(doctors pharmacies insurance companies etc). Camouflage behaviours are not easy to detect using above mentioned approaches temporal data mining can fulfil this purpose. There are many researches on continuous time series data analysis [19]–[22]. Some recent studies were conducted by [23]–[25] to analyse discrete event based sequential data. Sequence of physician orders are used to perform temporal sequence analysis. Outlier detection algorithms [26] are generally classified into two broad categories: First class of algorithms focuses on identification of anomalies in individual data points and the second class of algorithms considers the data as sequence in developing the model. Almost all the algorithms which are implemented in beymani belong to the first category. Fraud detection in real time is only possible when algorithm generates a model which can be used by real time fraud detection. Proximity based algorithms scans through whole database for detecting fraud but such approaches are not recommendable in real time environment. Fraud detection approach using SSIsomap activity clustering method is proposed by Yangchang [27]. Jurgovsky [28] utilized the concept of sequence classification for detecting credit card fraud. The unbalanced classification of data is the major issue which decreases the performance of machine learning algorithms while detecting frauds addresssed by [29]. Fraud detection in E-commerce industry transactions was performed by using a prudential Multiple Consensus model [30]. Novel LSTM based approach is proposed and applied by [31] on telecomunication dataset for fraud detection.

We propose a novel fraud detection methodology based on sequences mining and sequence prediction. The system considers sequence of services availed by each patient in last five years. For this purpose, we generate sequence of transactions of each patient to investigate specific set of services being availed by each patient from each specialty. All such sequences are then analyzed by the sequence rule engine to identify non-compliant sequences which are considered as anomalies which are further analyzed by the prediction based rule engine. Core objective of our research work is identification of anomalies based on historical medical sequential records.

III. PROPOSED METHODOLOGY

This section describes the methodology for insurance claim related fraud detection in healthcare systems. The three main elements of proposed framework [4] are Patients, Providers (doctors, Pharmacy and hospitals) and Services. Patients are availing services from providers. Providers can be doctors, hospitals and pharmacies. Services are availed by patients and provided by doctors, hospitals or pharmacies. These three elements are actually associated with each other. First we get time series traces of patients, to analyse the behavioral patterns of patients. The system consists of two cascaded modules which can identify anomalies and frauds.

A. SEQUENCE RULE ENGINE(SRE)

The Sequence rule engine (SRE) consists of following steps :

1) Convert transactional data into time series sequence database
2) Generate frequent sequence based rule engine
3) Generate probabilities of rare sequences
4) Sequence matching to detect anomaly

We cannot detect anomalies directly from patient time series traces, there is a need to dig out details. So we find out sequence of transactions of patients in each specialty separately. For each specialty sequence of services availed by each patient are captured. Once sequences for all specialties are computed then prefixspan algorithm is applied on each specialty. The main objective of applying prefixspan (frequent sequence mining algorithm) on each specialty is to get frequent sequences for all specialties. Secondly, the system considers different values of minimum support and minimum pattern length for each specialty. Based on the value of minimum support, length of the sequence get reduced. The comparison of prefixspan algorithm with other pattern sequence mining algorithms is provided in table 1. GSP (Generalized Sequential Patterns) is Apriori based approach but we use prefixspan which is based on pattern growth approach. The comparison of prefixspan algorithm with other pattern sequence mining algorithms is provided in table 1.

1) DEFINITION 1 (CLINICAL SERVICE EVENT)

Let S be a set of clinical service, T be the date of service availed and ε be the universal set containing all service event i-e the set of all possible service event identifiers. We assume that service events are characterized by multiple attributes. For instance, clincal service event has a service date, specialty name where this event has taken place and medical experts or doctors who have prescribed services.

2) DEFINITION 2 (PATIENT TIME SERIES TRACE)

A patient time series trace is represented as a sequence of service events. Each service event can appear more than once and for that time is non-decreasing. We consider Patient information detail P which contains Patient MR_No P_m, service date S_d and service event type S_t are defined in equation 1.

$$\forall P = (P_m, S_d, S_t) \tag{1}$$

We consider two main attributes clinical service type and service date and their functions are $\alpha_s \in \varepsilon \rightarrow$ S and $\alpha_t \in \varepsilon \rightarrow$ T respectively. So, e = $\{\alpha_s,\alpha_t\}$. The patient sequence ϵ and patient time series trace γ are defined in equation 2 and 3 respectively.

$$\epsilon = \{e_1, e_2, e_3 \ldots\ldots e_n\} \tag{2}$$
$$\gamma = \{\epsilon_1, \epsilon_2, \epsilon_3 \ldots .\epsilon_n\} \tag{3}$$

Patient sequences are basically services availed by patients. Patient time series trace is a set of all patient sequences in different specialties. In the time series trace, if events occur at the same date, they are ordered by transaction_id. The patient time series trace of employee_id 12838 is shown in Fig 2. The colors of bars in Fig 2 are depicting names of all specialties from which this employee availed services. Y-axis is the number of transactions availed by the patient in each specialty. X-axis is showing date on which particular services are availed.

3) DEFINITION 3 (SPECIALTY SEQUENCE)

Let L be a specialty log and Sim(ϵ, p) be the similarity measure for any two sequences ϵ and p in L. The L can be partitioned into multiple specialties, in our case there are 62 specialties $\varphi_1, \ldots, \varphi_{62}$. Specialty sequence φ is shown by equation 4,5 and 6.

$$\varphi_1 = \{\epsilon_1, \epsilon_2 \ldots \epsilon_n\} \tag{4}$$
$$\varphi 2 = \{\epsilon_1, \epsilon_2 \ldots \epsilon_n\} \tag{5}$$
$$\vdots$$
$$\varphi_{62} = \{\epsilon_1, \epsilon_2 \ldots \epsilon_n\} \tag{6}$$

Therefore, $\epsilon_i \in \varphi_i$ where i represents number of specialties. As shown in Fig 3, that there are different patient traces ϵ in medical specialty φ. Each patient trace consists of α_a and α_t. For each patient, sequences are generated for specific specialty. In Fig 3, colors of bars are showing services which are availed by the patient from this specialty. Y-axis is the number of transactions of this patient for each service.

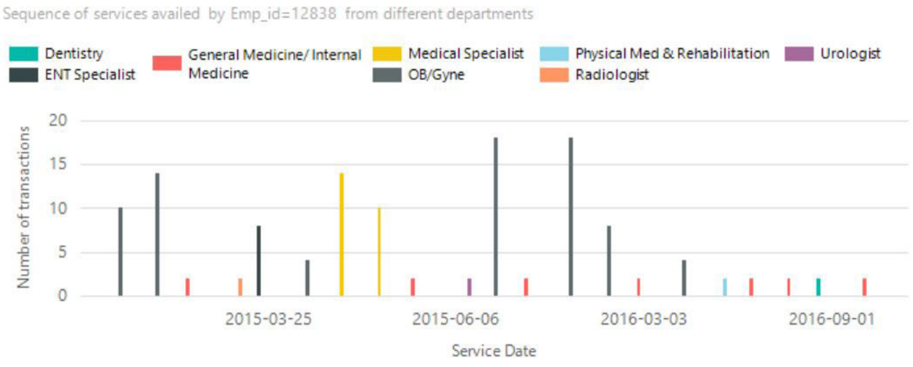


FIGURE 2. Sequence of services availed by Empid 12838.

TABLE 1. Comparison of sequence mining algorithms.

Algorithm	Freespan [32], [33]	Prefixspan [34], [35]	GSP [33]–[36]	SPADE [32], [34], [37]	Apriori [38]–[40]
Generate and Test	No	No	Yes	Yes	Yes
MultiScan of Database	No	No	Yes	No	Yes
Candidate Sequence Pruning	No	Yes	Yes	Yes	-
Sampling and/or compression	No	No	No	No	Yes
DFS based approach	Yes	Yes	No	No	-
BFS based approach	No	No	Yes	Yes	-
Top-down search approach	Yes	Yes	No	No	-
Bottom-up search approach	No	No	Yes	Yes	-
Prefix growth approach	No	No	No	No	-
Search Space Partitioning	Yes	Yes	No	Yes	Yes
Database vertical projection	No	No	No	Yes	-
Support counting avoidance	No	No	No	No	-
Position coded avoidance	No	No	No	No	-

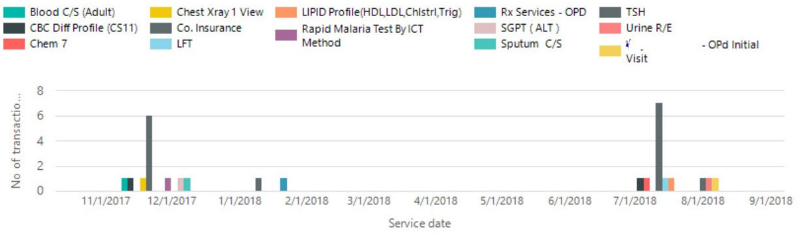


FIGURE 3. Sequence of services availed by employee 11757 in medical specialist specialty.

The workflow of sequence rule engine is explained in Fig 4. Sequence database is generated from transactional database. Four sequences {201,567,345}, {301, 201,434,567}, {301,201,567,345} and {201,301,567}are shown as input to prefixspan algorithm. The prefix 301, 201, 567 are chosen by the algorithm, as their support is greater than the mentioned minimum support 2. For pefix 301, algorithm checks the next service 301, 434 comes together in only one sequence but {301, 201} comes together in two sequences therefore support of pattern {301, 201} is 2.

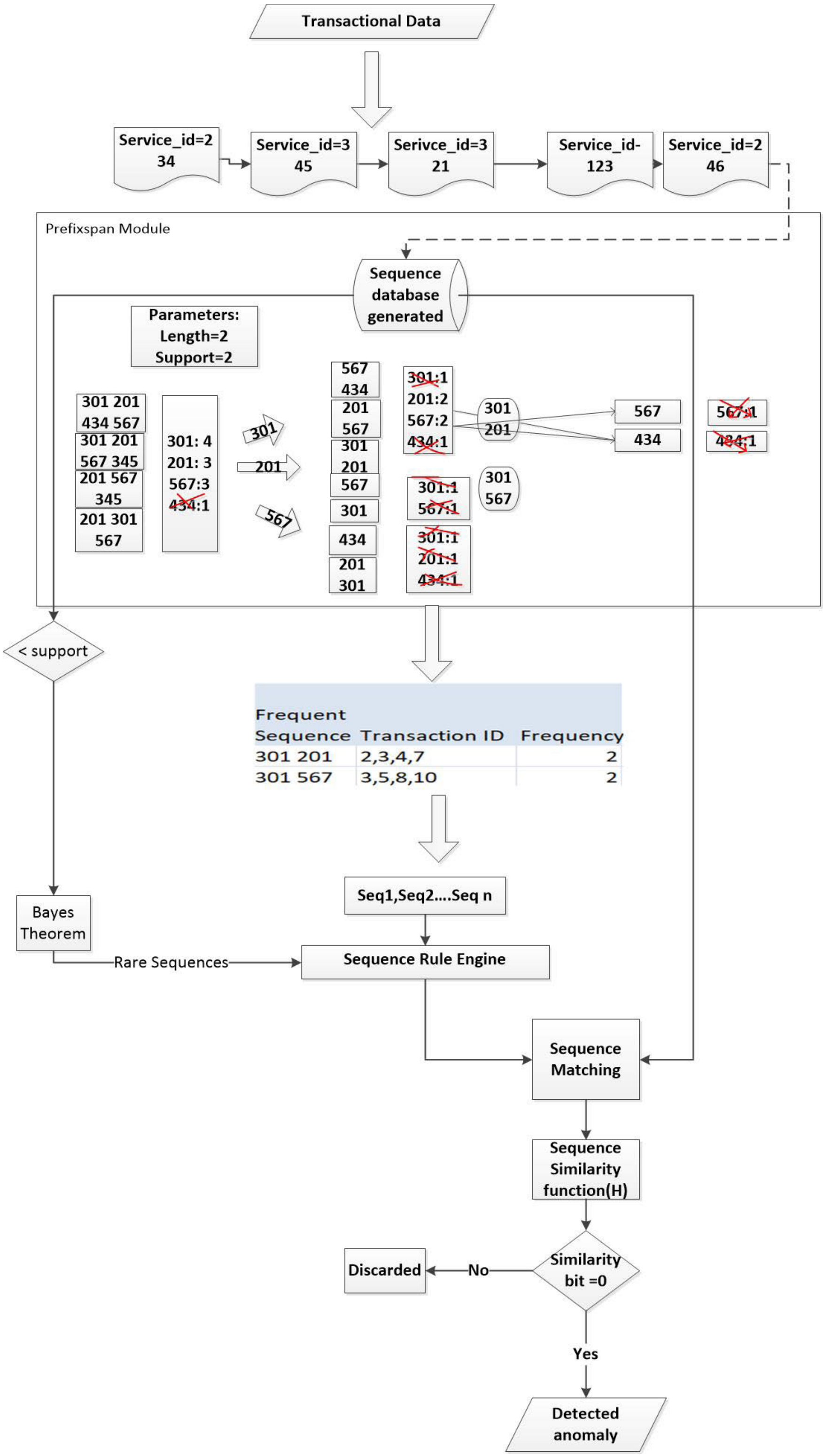


FIGURE 4. **Workflow of sequence rule engine.**

Similarly, {301, 567} comes in two sequences, support of this pattern is also 2. All other patterns with prefix 301 : {301, 434} are discarded {shown in Fig 4 by red cross} and not selected as frequent sequences because their support is less than minimum support value. Same process is repeated for all other two prefixes also. The output of prefixspan is frequent sequences. As it is already mentioned that minimum support is the input parameter of prefixspan, we set minimum support value to 2 in Fig 4. We get all frequent sequences greater and equal to 2. All those sequences whose support is less than 2 are further processed using Bayes Theorem. Whatever the value of minimum support, all the sequences whose support is less than the minimum support value, are further processed.We use Bayes theorem for computing probability of every such sequence in the sequence database. There are many other alternatives like Instance-based methods, decision tree algorithms etc. The result of Bayesian theorem depends strongly on prior probabilities [41] and is useful in our case. The use of other alternatives is not suitable for this scenario. The affect of using Bayes Theorem is that accuracy of our methodology gets improve because if the probability of any sequence is too low, it means it may occur just once in the last five years so it is not rare sequence instead it can be an anomaly. These frequent sequences and rare sequences, are fed into the Sequence matching function. Sequence rule engine is generated based on two types of sequences :

- Frequent sequences
- Rare sequences with probabilities.

We determine posterior probabilities of all the rare sequences in specific specialty by using Bayes Theorem. The average of such probabilities are computed for each specialty and set the said average as a threshold. The Bayes Theorem can be applied only when we already know other probabilities:

$$P(\varphi|\epsilon) = \frac{P(\varphi)P(\epsilon|\varphi)}{P(\epsilon)} \tag{7}$$

For each specialty φ, we calculate probability of each less frequent/rare sequence ϵ.

- When specialty φ probability that this ϵ will occur is denoted as $P(\varphi|\epsilon)$,
- When sequence is this ϵ probability that it will occur in this φ is denoted as $P(\epsilon|\varphi)$.
- Probability of occurrence of this specailty φ in whole data is denoted as $P(\varphi)$.
- Probability of occurrence of this sequence ϵ in whole data is denoted as $P(\epsilon)$.

Patient sequences from sequence database, are entered as an input to Sequence matching function. If patient sequences donot match any sequence from sequence rule engine. It will be identified as anomaly. When any anomalous sequence is found we check its similarity with these rare sequences. There are two possibilities, firstly, there is a possibility that anomalous sequence matches with one of the rare sequences, in that case we check probability of that rare sequence, if the probability of that sequence is less than the threshold, only then that anomalous sequence is forwarded to prediction based engine otherwise not.

Second possibility is that anomalous sequence is not matched with any one of these rare sequences, in that case it is passed to Prediction based engine for a further analysis.

B. PREDICTION BASED ENGINE

Prediction based engine(PBE) performs following steps:

1) Once anomaly is detected by first module, identified anomalous cases are forwarded to prediction based engine for a further analysis.
2) PBE takes test case as a testing data and breaks down each test case into test sets.
3) When prediction engine generate null value for any test set, that particular case will be identified as Fraud.

Each identifed sequence is entered as a test case for Pre diction based engine. Test case is representing anomalous sequence and test set is representing each service in that sequence as shown in Fig 5. Each test case is a vector of test sets. Prediction based engine predicts next service for each test set of considered test case. When Prediction based engine predicts null value for any test set this means that particular test case is fraud.

1) LOOP IN CPT

There are three structures in Compact Prediction Tree

1) Prediction tree
2) Inverted index
3) Count table

The training phase is performed by using Sequence database as a training set. Services (items of sequences) are inserted one by one simultaneously in Prediction tree and Inverted index table. Prediction tree is composed of nodes. A node contains list of child nodes that are pointing towards parent node. The sequence in prediction tree is represented by full branch or partial branch of a tree which is starting from child node of a root node.

The prediction tree is constructed in the following way, for each given training sequence, it checks if the considered node (the root) has a direct child similar with the first item of this sequence. If it does not match then a new child is added to the root node with this item's value. After this, the pointer is moved to the recently created child and same process is repeated for each item in the training sequence. This structure creates hash table which contains key for each unique item it found during the training phase. Each key contains a bitset that provides reference of the sequences in which the item appears. The size of the bitset is N which is representing number of sequences used at training phase. Lookup table structure basically links Prediction tree and inverted index. For each sequence reference id, lookup table points to the last item of the sequence in the prediction tree. The main function of lookup table is to retrieve sequence from the prediction tree based on the sequence reference id. Each time sequence is added to the prediction tree, lookup table is updated. Inverted index is the structure which is basically used to find number of sequences which contain given set of items. In Prediction phase, given service 'S' is matched with the similar sequences that have been generated and stored in the lookup table along with their associated frequency values. This structure holds these frequency values for a specific prediction and hence is unique for each individual prediction task. In this way we

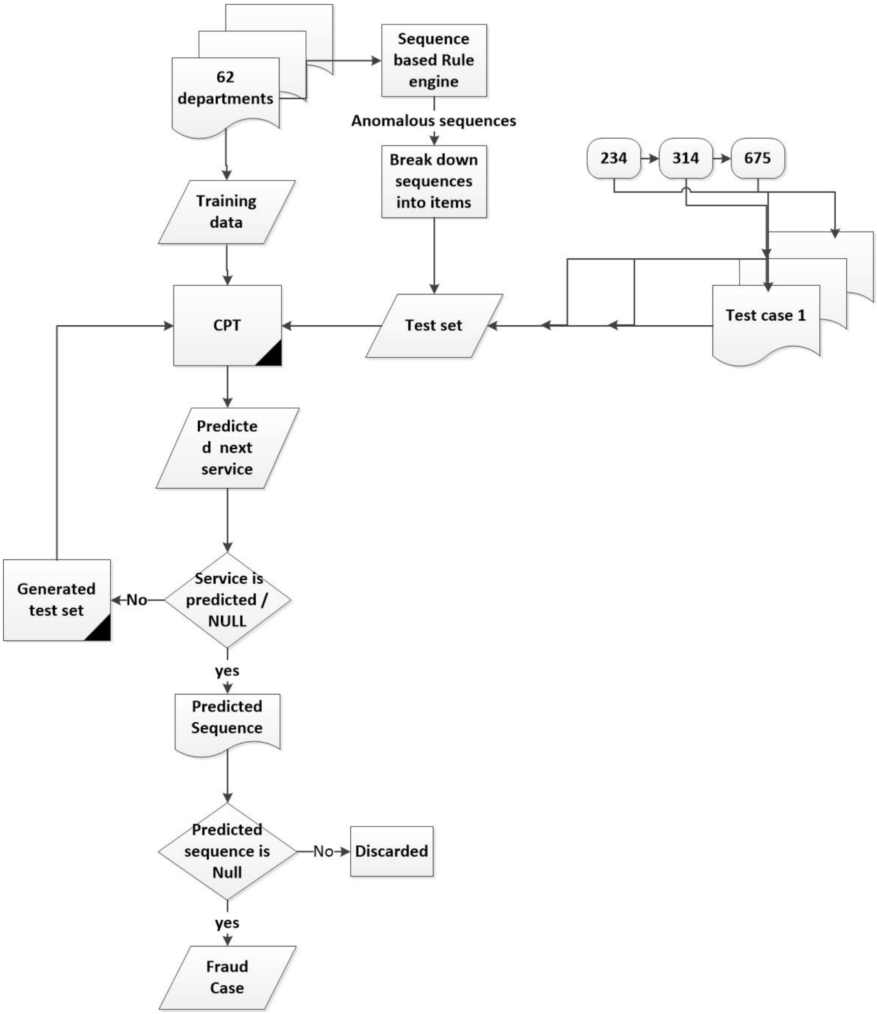


FIGURE 5. **Prediction based engine.**

are able to check each service in anomalous sequence using prediction technique whether services availed in anomalous sequence can be availed from specific specialty. For this purpose after sequence matching step, all anomalous sequences are forwarded as input to com pact prediction tree as shown in Fig 5. Compact prediction tree (CPT plus) CPT takes training set and testing set as an input. So we entered traninig set and anomalous sequences for each specialty as test case depicted in Fig 5. Then in each iteration we concatenate each record in test case with generated predicted set and then used this as new test case for next iteration. This loop continues until we get null value in predicted set as depicted in Fig 5. As a result, we achieved group of services which can be availed after selected test set. But if prediction engine predicts null sequence for any test set then associated test case of this test set is identifed as fraud. Our Prediction based engine is generated that is trained on sequences for each specialty of local hospital. The Prediction based engine detects fraudulent behaviour from anomalous sequences. In addition, Prediction based engine can be used for controlling budget related issues in hospitals as well as for predicting next event in clinical procedures.

It's evident that CPT is an incremental model as each time it selects the next event in the sequence, its always based on the frequency of that event in the training set with respect to the first event.

Algorithm 1, describes the overall functionality of sequence rule engine. Input parameters are patient sequences from the sequence database denoted by ϵ. ϵ_1 is the set of sequences denotes specialty and the minimum support. The algorithm runs for 62 specialties and for each specialty we set minimum support value. Prefixspan algorithm is applied on all sequence whose support is greater than or equal to the entered minimum support value. F_x frequent sequences are generated. The all sequences whose support is less than minimum support are K. Bayes Theorem is applied using equation 7. Once probability of all K sequences are computed then Mean of all is computed which decides the threshold value. Rare K sequences are all those sequences whose probabilities are greater than the threshold. The similarity function is based on the value of similarity bit. Each patient sequence is matched with the frequent sequence F_x, if sequence is matched, similarity bit value is set to 1. As a result, all the sequences ϵ' are generated with their MR_NO. These sequences ϵ' are matched with the rare sequences K. Only then α is equal to zero otherwise 1. When α is 1, then sequence is anomalous and in result all anomalous sequences ϵ'' are generated along with MR_NO. All these anomalous sequences are entered as test data to the Algorithm 2.

In Algorithm 2, each anomalous sequence is converted into column vector V. Each item R is test set. On each Test set compact prediction tree is applied. If the value of prediction P_t is NULL then test case is detected as fraud otherwise not.

Notation	Description
φ	Specialties
min_sup	Minimum support
u	Counter for specialties
x	Counter for rare sequences
Q	Sequential pattern
L	Length of Q
φ| Q	Q-projected database
F	Frequent sequence
j	Number of frequent sequences
Q’	New sequential pattern
K	Rare sequences
g	Number of Rare sequences
ϵ	Patient sequences
α	Similarity bit
ϵ'	Set of anomalous sequences
ϵ''	Set of anomalous sequences forwarded to PBE
i	Loop counter
l	length of ϵ''_n
V	Column vector
R	Each item in column vector(test set)
D	Threshold
T	Test case

IV. RESULTS AND DISCUSSION

Before starting our analysis we compared our current methodolgy with few related research studies as listed in Table 2.

A. CASE STUDY

The proposed framework is evalauted on five years[2013, 2014, 2015, 2016, 2017, 2018, 2019] insurance claim transactional data of 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. Table 3, is showing size of transactional dataset. The considered attributes for this framework are mentioned. Table 4 is showing set of attributes which are providing details about the availed and provided services.

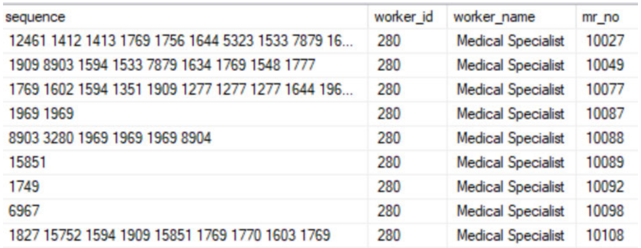
Service_major_description includes clinic, laboratory, Radiology, Miscellaneous, Supplies, Consultation and Pharmacy. Service_id 13556(Co_insurance) and 1969(Rx_services), these two services are being used under service_major_description Miscelleneous and Pharmacy. In Miscelleneous, all pateints are using Co-insurance service. In co-insurance bill amount is divided to some percentage between hospital and patient. Rx_services under service_major_description Pharmacy, is the type of treatment to patient. So pharmacy

Algorithm 1 Sequence Rule Engine

```
Input: $\varphi_1, \varphi_2 \ldots \varphi_{62}$ and $min\_sup$
Output: $\epsilon'$ $\epsilon''$ sequences with $MR\_no$

for $u \leftarrow 1$ to 62 do
    if ($\epsilon > min\_sup$) then
        Call function Prefixspan(Q, L, $\varphi$|Q)
        if ($Q \neq \phi$) then
            Scan $\varphi_u|Q$
            Find frequent sequence $F_x$
        else
            sequence database $\varphi_u$
        end
        foreach $F_x$ do
            Append $F_x$ to $Q$
            Generate new $Q'$
        end

        foreach $Q'$ do
            Generate projected database
            Call function Prefixspan(Q’, L+1, $\varphi$|Q’)
        end
    end
    else
        for $K \leftarrow 1$ to $g$ do
            $P(\varphi_u|K) \leftarrow P(\varphi_u)P(K|\phi)/P(K)$
        end
    end
end
Calculate Mean $P(\varphi_u|K)$
Set Mean as D
// Perform Sequence matching for Patient sequences $\epsilon$
if ($\epsilon == F_x$) then
    $\alpha \leftarrow 1$
else
    $\alpha \leftarrow 0$
end
end
Generate $\epsilon'$ with $MR\_no$
if ($\epsilon' == K$) && (Mean(K) > D ) then
    $\alpha \leftarrow 0$
else
    $\alpha \leftarrow 1$
end
end
Generate $\epsilon''$ with $MR\_no$
end
```

Algorithm 2 Prediction Based Engine

```
Input: Training set : $\{\varphi_1, \varphi_2 \dots \varphi_{62}\}$
and Testing set : $\epsilon''_i$ where $i = \{1, 2, 3, ..62\}$

for ($i \leftarrow 1$ to 62) do
    Train CPT with $\varphi_i$
    // Prepare Test Case
    for ($\epsilon'' \leftarrow 1$ to $l$) do
        Convert $\epsilon''_n$ into $V$
        Each $R$ in $V$ entered as $T$
        Check using CPT Prediction
    end
    if ($P_t$ == NULL) then
        $T$ is Fraud
        else
            $T$ is Normal case
        end
    end
end
```



sequence	worker_id	worker_name	mr_no
12461 1412 1413 1769 1756 1644 5323 1533 7879 16...	280	Medical Specialist	10027
1909 8903 1594 1533 7879 1634 1769 1548 1777	280	Medical Specialist	10049
1769 1602 1594 1351 1909 1277 1277 1277 1644 196...	280	Medical Specialist	10077
1969 1969	280	Medical Specialist	10087
8903 3280 1969 1969 1969 8904	280	Medical Specialist	10088
15851	280	Medical Specialist	10089
1749	280	Medical Specialist	10092
6967	280	Medical Specialist	10098
1827 15752 1594 1909 15851 1769 1770 1603 1769	280	Medical Specialist	10108

FIGURE 6. Subset of patient sequences of services availed in medical specialist.

and Miscelleneous both service_major_descriptions are availed from almost all specailizations.

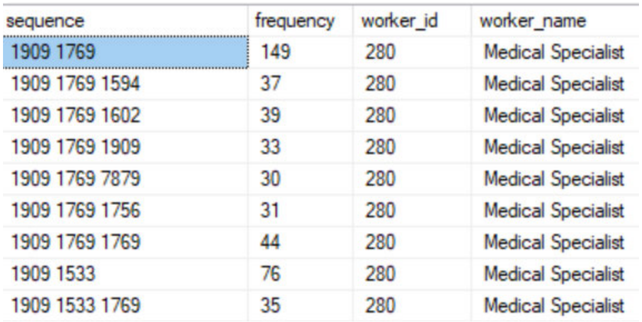
1) EXPERIMENTATION

After the conversion of transactional data into sequence database, we get patient sequences for every specialty. Subset of sequences for medical specialist specialty is shown in Fig 6. The sequence database is entered as an input to Sequence rule engine, firstly prefixspan algorithm executes and it generates frequent sequences for *medical specialist* specialty as shown in Fig 7.

One of the frequent patterns in *cardiology* specialty is shown in Fig 8.

Two of the frequent sequences of *ENT specialty* are shown in Fig 9.

Frequent sequences generated in *Pediatrician* are shown in Fig 10. This proposed methodology has been validated on five years transactional data. Subsets of services which can be availed are provided in Table 5. The subset of frequent sequences of some specialties with service description and service ids are provided in Table 6. The SRE is generated based on the frequent and rare sequences of services for each specailty. Frequent sequences are all those sequences whose support is greater than minimum support value. And rare sequences are all those whose support is less than the minimum support value. Table 7 depicts subsets of rare sequences for few specialties. These sequences are basically defining a rule for each specialty, rule is based on all frequent



sequence	frequency	worker_id	worker_name
1909 1769	149	280	Medical Specialist
1909 1769 1594	37	280	Medical Specialist
1909 1769 1602	39	280	Medical Specialist
1909 1769 1909	33	280	Medical Specialist
1909 1769 7879	30	280	Medical Specialist
1909 1769 1756	31	280	Medical Specialist
1909 1769 1769	44	280	Medical Specialist
1909 1533	76	280	Medical Specialist
1909 1533 1769	35	280	Medical Specialist

FIGURE 7. Frequent sequence of services availed in medical specialist.

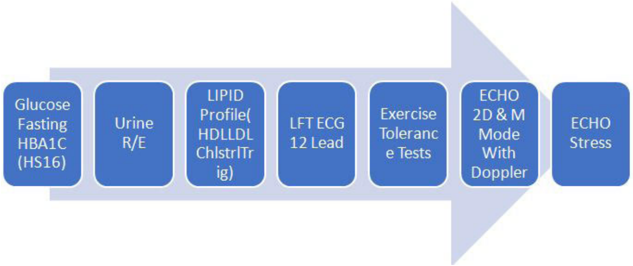

Glucose Fasting HBA1C (HS16)
Urine R/E
LIPID Profile(HDLLDL ChlstrlTr ig)
LFT ECG 12 Lead
Exercise Toleranc e Tests
ECHO 2D & M Mode With Doppler
ECHO Stress

FIGURE 8. Frequent sequence of services availed in Cardiology.

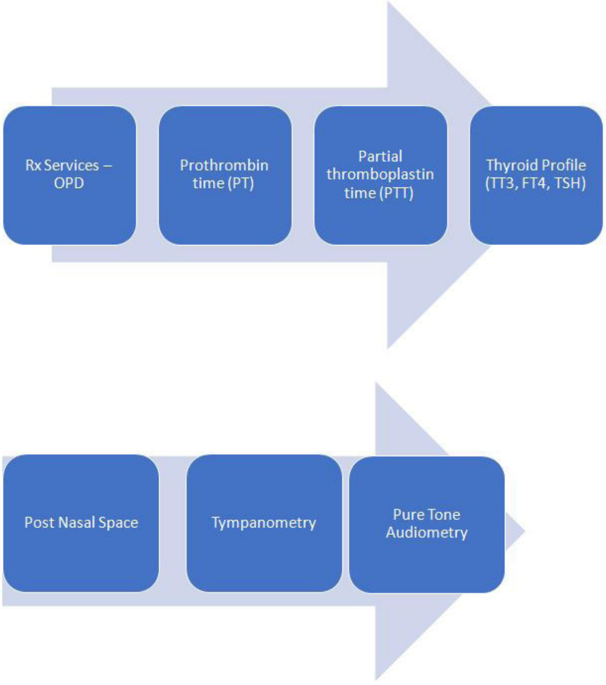
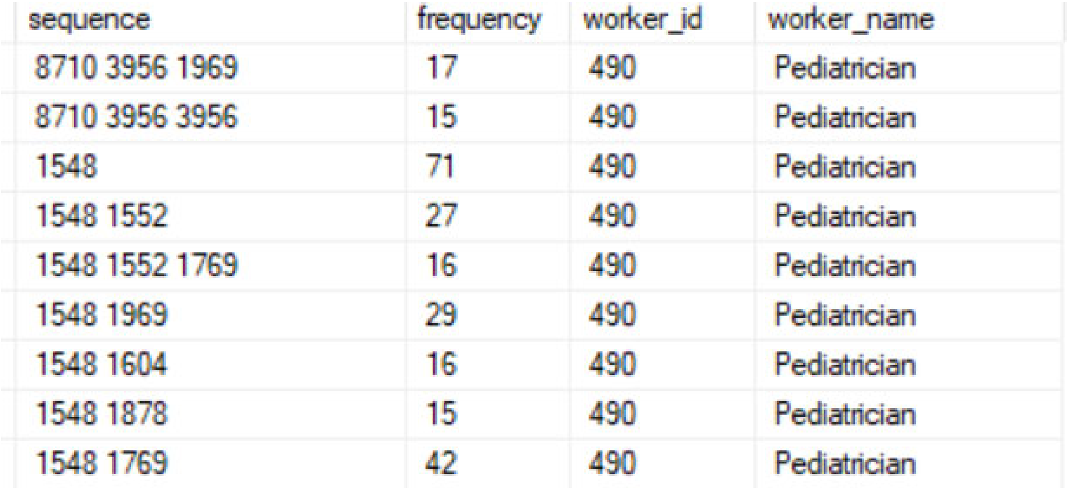


FIGURE 9. Frequent sequence of services availed in *ENT* specialty department.

TABLE 2. comparison of proposed methodology with other studies.

Study name and References	Technique used	Comments
Identifying frauds and anomalies in Medicare-B dataset [42]	A similarity graph and Page rank algorithm	Provider level frauds are detected. Similarity graph between the Prescriptions of doctors of same specialty are created and then page rank algorithm is utilized to detect anomalies. Whereas, our methodology is using patient sequences for each specialty for the identification of anomalies. We generated sequence rule engine for the identification of anomalies. Furthermore, these anomalies are analyzed using Prediction based engine.
Sequential pattern mining of electronic healthcare reimbursement claims: Experiences and challenges in uncovering how patients are treated by physicians [43]	Sequence mining approach	Heart Disease, Breast cancer and autism spectrum disease Procedures administered to patients are analyzed but fraudulent activities are detected by considering costs of these procedures. But our methodology self learns from historical medical data and generate frequent and rare sequences and based on these sequences anomalous sequences are detected for 62 specialties.
Fraud detection and frequent pattern matching in insurance claims using data mining techniques [44]	Kmeans clustering and association rule mining and Gaussian distribution.	Period based anomalies and disease based anomalies are detected in this research. The payment and time labs are analyzed for detecting frauds. But we are detecting frauds without considering payments of availed services or procedures.



sequence	frequency	worker_id	worker_name
8710 3956 1969	17	490	Pediatrician
8710 3956 3956	15	490	Pediatrician
1548	71	490	Pediatrician
1548 1552	27	490	Pediatrician
1548 1552 1769	16	490	Pediatrician
1548 1969	29	490	Pediatrician
1548 1604	16	490	Pediatrician
1548 1878	15	490	Pediatrician
1548 1769	42	490	Pediatrician

FIGURE 10. Sequence availed from pediatrician.

TABLE 3. Attributes in dataset.

Attributes	Value
Unique number of serviceIds	1206
Unique number of Doctors	486
Unique number of specailityId	62
Total number of transactions	441506

and rare sequences which have been availed from this specialty in last five years.

The proposed fraud detection methodology is able to identify possible anomalous behaviours from employee transactional data. It performs two possible cascaded checks at two different levels :

1) Similarity verification from sequence rule engine
2) Fraud detection using Prediction based engine

The main challenge is to set minimum support value for each specialty. The need is to get maximum number of frequent sequences with two to three length patterns. This fact can be expalined by considering case of specialty *cardiology*. The frequent sequence in *cardiology* specialty with support 2 are listed in Table 8. When we increase the minimum support

TABLE 4. Each transaction 's attributes in data set.

Attributes	DataTypes
MRNO	Varchar(255)
Serviceid	Varchar(255)
Service description	Char
Service Major Description	Char
Service Minor description	Char
Doctor	Varchar
Specialty	Varchar
Category	Char

TABLE 5. Service description.

SERVICE_ID	SERVICE_DESCRIPTION
1	Allergy Skin Test (Airborne Allergens)
3	Level C Allergy Skin Test
4	Level D Allergy Skin Test
7	Level B Allergy Vaccination
8	Level C Allergy Vaccination
9	Level D Allegy Vaccination
10	Allergy Skin Test (Food Allergens)
18	Misc. Procedure 2A
101	ECG 12 Lead
102	ECG 12 Lead with Right Sided Chest
104	Fetal Echo / Pediatric Echo
105	ECHO 2D & M Mode With Doppler
106	ECHO Stress
107	ECHO F/U with in one week
108	ECHO Transesophageal
109	24 Hour Holter
110	Exercise Tolerance Tests
115	Carotids, Ultrasound Doppler
118	Extremities Arter.U/S Dopp.Lower
120	Extremities Venous U/S Dopp. Lower
125	24-Hour Ambulatory B.P. Monitor
167	Small Procedure set

(input parameter) we get sequences shown in Table 9. In all the tables subsets of frequent sequences have been shown.

It have been observed that length and number of sequence gets lesser whenever there is an increase in minimum support value. But this depends on number of sequences extracted from transactional data for specific specialty. The objective of this module of proposed methodology is to generate frequent and rare sequences based rule Engine. Table 10 reflects frequent sequences for *dentistry* specialty. If any patient visits dentist and he/she deviates from these generated sequences, proposed methodology's first module will identify it as anomaly. There are different cases of anomalies which can be detected by proposed methodology.

2) CASE 1 : SERVICES AVAILED BUT NOT COMPLIANT WITH SEQUENCE GENERATED BY SRE

For instance, if a patient avails service_ID 317 (Removal of Impected tooth - Simple) first and then avails service_ID 314 (OPG "Orthopentomogragh"), it will be identified as anomalous. The first module identifies all such anomalous cases from all specialties. Once the anomalous sequence has been identified it will be treated as a test case and 314 is one test set and 317 is the second test set. The Prediction based engine, checks each test set of this test case. If in any test case it predicts null value then whole case is identifed as fraud.

3) CASE 2 : REPITITION IN AVAILED SERVICES

The sequence is availed by MR_NO from Dentistry specialty. This sequence is identified as anomaly by module 1 as shown in Fig 11. The service_description of services which are availed are Composite_Large, 45 OPd Initial Visit, Removal of Impected tooth Simple, Removal of Impected tooth Simple. This anomalous sequence is forwarded to prediction based engine. This sequence is identified as fraud by module 2 also as shown in Fig 12. The Removal of Impected tooth Simple service is availed twice. The PBE predicts null for this service as engine finds no other sevice availed after this specified service and this whole sequence is identified as fraud.

4) CASE 3 : FEW SERVICES IN A SEQUENCE ARE ANOMALOUS

There is a possibility that first few services in a patient sequence are compliant to rule engine sequences and can be availed from specified specialty but second half of services are not compliant. Fig 13 depicts this case service ID 7910 is with service_description "Surgical Extraction by Division", service ID 229 with service description "Pulpotomy", Service ID 246 with service_description "Temporary Filling with Sedative Dre", service ID 1034 with service description" Brain/Head (3-D Imaging)" and last service ID 1034 with service_description" Brain/Head (3-D Imaging)". The first three services of this sequence are compliant with sequence rule engine's sequences but last two services are not so this sequence is identified as anomaly by module 1 as shown in Fig 13. The identified anomalous sequence is forwarded to module 2 PBE for further analysis. Fig 14 depicts that module 2 finally identified this sequence as fraud because PBE predicts null for service ID 1034 and finally identified whole sequence as fraud.

5) CASE 4 : SERVICES AVAILED FROM SPECIALTIES WHICH ARE NOT PERMITTED FROM THIS SPECIALTY

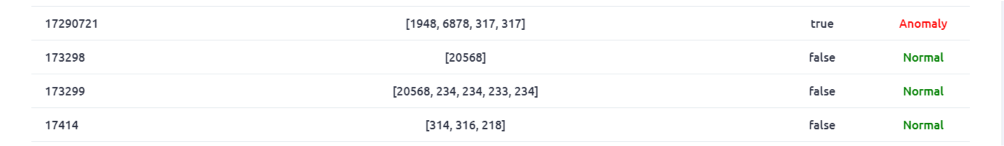
Fig 15 depicts anomalous sequences detected in module 1, sequence [1756, 1762, 1769, 1909] is identied as anomaly. These are subsets of results. The output of our methodology is shown in Fig 16, due to the limitation of space we are discussing just one such case. The methodology results are computed for *dentistry* in Fig 16 which shows the subsets of sequence generated for this speciality. The sequence [1756, 1762, 1769, 1909] is identified as a fraud, the service_id 1762 is for " Erythrocyte sedimentation rate (ESR)", it is a test to check heart functionality and after this patient availed service_id = 1762, which is the service "Peripherial Flim", service_id = 1769 which is the service "CBC Diff ProfileCS11)" and service_id = 1909 which is "Urine RE". The fraud is identified because prediction engine predicts null for each service of this sequence and in actual the dentist cannot prescibe these services, therefore this sequence is identified as a fraud. Many similar cases are detected by our methodology in some other specialties. At this stage the identified frauds are included in the type of patient level fraud that patient avails the service and gets insurance claim against it. Hence, it is crystal clear that this service cannot be availed from this specialty and generated sequence is not present in sequence rule engine so it is identified as an anomaly by module 1 and identified as fraud by module 2 because there is no sequence in which service 1756 is availed as a first service and there is no sequence in which service_id 1769 is availed after service_id 1756 and same is true for

TABLE 6. Subsets of frequent sequences of few specialties.

Specialty_name	Frequent Sequence using Service_id	Frequent sequence using service_description	Support
Medical Specialist	1587 1602 1637 1769 1909	Glucose Fasting, Blood Urea nitrogen, LIPID Profile, Anti Thrombin(111), Urine R/E	26
	1587 1602 1909	Glucose Fasting, Blood urea nitrogen, Urine R/E	55
	1533 1769 1909	Thyroid stimulating hormone(TSH), AntiThrombin, Urine R/E	101
Pediatrician	19015 1769 1629	Blood C/S (Peads), CBC Diff Profile (CS11), C-Reactive Protein(CRP) High Sensitivity	25
	1878 1909	Stool Routine Examination, Urine R/E	38
	1533 1532 4025	thyroid stimulating hormone, T3,490-OPD Initial visit	83
	1769 1909	CBC Diff Profile (CS11), Urine R/E	206
Urologist	1334 1826 1909	Both Kidneys or GenitoUrinary Tract, Urine C/S , Urine R/E	70
	1602 1909	Creatinine Serum, Urine R/E	91
ENT Specialist	1718 1715 1769	Activated partial thromboplastin time (APTT), PT (Prothrombin Time) ,CBC Diff Profile (CS11)	30
	370 371	Pure Tone Audiometry, Tympanometry	78
Nephrologist	1412 1413	HBs Ag, Hepatitis C Virus Ab (HCV) Hepatitis C Virus Ab (HCV)	26
OB/Gyne	1613 1769 1909 7840 1909 7840	GTT 2 hrs. 75 gm Glucose, CBC Diff Profile , Urine R/E, 35OPd Follow-up Visit, Urine R/E, 35_OPd Follow-up Visit	54
	1613 1769 1909 680	GTT 2 hrs. 75 gm Glucose, CBC Diff Profile (CS11) , Urine R/E, Tetanus Toxoid Tetta vax "0.5ml P/D"	98
Neurologist	1756 1769	Erythrocyte sedimentation rate (ESR), CBC Diff Profile (CS11)	41
	1174 1602	Brain with Contrast, Creatinine Serum	28
	1594 1769	SGPT (ALT), AntiThrombin	39
Orthopedic	1603,1323	25-Hydroxy Vitamin D, Knee Ap/Lat/Skyline	25
	1603 1756 1769	Uric Acid Serum, ESR, CBC Diff Profile (CS11)	30

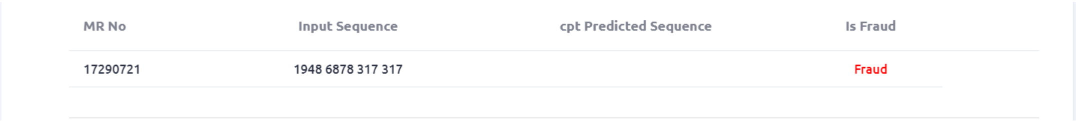
TABLE 7. Subset of rare sequences with probability.

Sequence	Sequence_name	Support	Specialty_id	Specialty_name	Probability
1969 1769 1769	Rx Services – OPD ,CBC Diff Profile (CS11) ,CBC Diff Profile (CS11)	2	320	Neurologist	0.0008
1715 1769 1909	PT (Prothrombin Time), CBC Diff Profile (CS11), Urine R/E	2	320	Neurologist	0.0008
1548 1552 1969	Iron Serum, TIBC, Rx Services – OPD	9	280	Medical Specialist	0.0035
1548 1604 1769	Iron Serum, Calcium Serum, CBC Diff Profile (CS11)	7	280	Medical Specialist	0.0027
1073 3899 3900	Sinuses Axial Coronal W/O, 67- OPd Initial Visit, 67- OPd Followup Visit	5	160	ENT Specialist	0.0077
1075 1079 1602	Brain without Contrast, Non Ionic Contrast Medium, Creatinine Serum	4	160	ENT Specialist	0.0031
3899 12963	67- OPd Initial Visit , 3771 - OPd Initial Visit	4	160	ENT Specialist	0.0015
1351 1769	Abdomen Upper, CBC Diff Profile (CS11)	3	5	Cardiologist	0.0046
1637 1644 5323	LIPID Profile(HDL,LDL,Chlstrl,Trig), LFT, HBA1C (HS16)	6	5	Cardiologist	0.0069
1412 1413 1533	HBs Ag, Hepatitis C Virus Ab (HCV). TSH	3	5	Cardiologist	0.0012
1537 1538 1542 1337	luteinizing hormone(LH), Follicle-stimulating hormone(FSH), Testosterone, Scrotum	3	620	Urologist	0.0012
1602 1603 1909	Creatinine Serum, Uric Acid Serum, Urine R/E	10	620	Urologist	0.0039



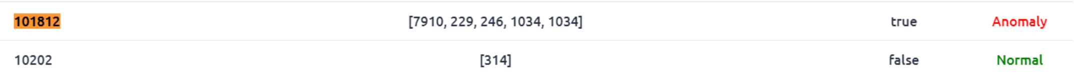
17290721	[1948, 6878, 317, 317]	true	Anomaly
173298	[20568]	false	Normal
173299	[20568, 234, 234, 233, 234]	false	Normal
17414	[314, 316, 218]	false	Normal

FIGURE 11. Anomalous sequences after sequence matching.



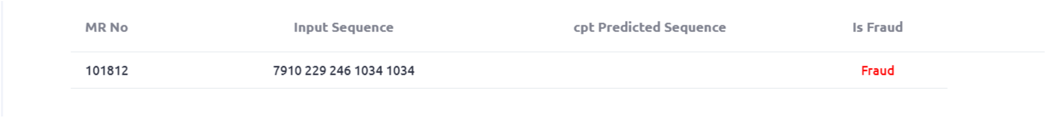
MR No	Input Sequence	cpt Predicted Sequence	Is Fraud
17290721	1948 6878 317 317		Fraud

FIGURE 12. Anomalous sequences after prediction.



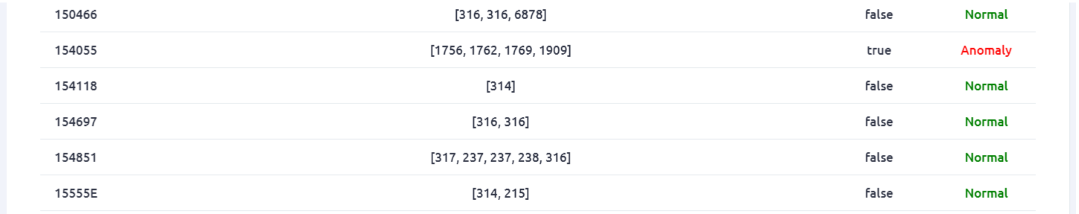
101812	[7910, 229, 246, 1034, 1034]	true	Anomaly
10202	[314]	false	Normal

FIGURE 13. Anomalous sequence for case 3.



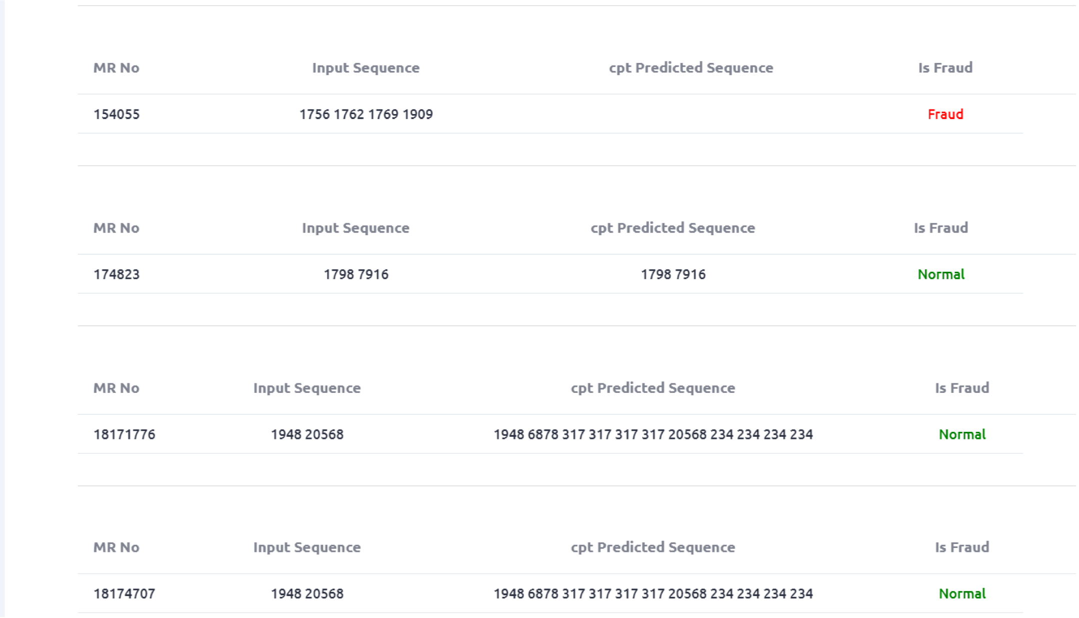
MR No	Input Sequence	cpt Predicted Sequence	Is Fraud
101812	7910 229 246 1034 1034		Fraud

FIGURE 14. Final decision for case 3.



150466	[316, 316, 6878]	false	Normal
154055	[1756, 1762, 1769, 1909]	true	Anomaly
154118	[314]	false	Normal
154697	[316, 316]	false	Normal
154851	[317, 237, 237, 238, 316]	false	Normal
15555E	[314, 215]	false	Normal

FIGURE 15. Anomalous sequences after sequence matching for dentistry.



MR No	Input Sequence	cpt Predicted Sequence	Is Fraud
154055	1756 1762 1769 1909		Fraud

MR No	Input Sequence	cpt Predicted Sequence	Is Fraud
174823	1798 7916	1798 7916	Normal

MR No	Input Sequence	cpt Predicted Sequence	Is Fraud
18171776	1948 20568	1948 6878 317 317 317 317 20568 234 234 234 234	Normal

MR No	Input Sequence	cpt Predicted Sequence	Is Fraud
18174707	1948 20568	1948 6878 317 317 317 317 20568 234 234 234 234	Normal

FIGURE 16. Final results for specialty dentistry.

TABLE 8. Minimum support = 2.

Sequence	Frequency	Specialty_id	Specialty_name
1536	2	5	Cardiologist
1536 5330	2	5	Cardiologist
7177	11	5	Cardiologist
7177 1602	2	5	Cardiologist
7177 1637	2	5	Cardiologist
1553	2	5	Cardiologist
7199	17	5	Cardiologist
7199 101	3	5	Cardiologist
7199 110	2	5	Cardiologist
1827	2	5	Cardiologist
1583	12	5	Cardiologist
1583 1602	10	5	Cardiologist
1583 1602 1637	5	5	Cardiologist
1583 1602 1637 1769	4	5	Cardiologist

TABLE 9. Minimum support = 20.

Sequence	Frequency	Specialty_id	Specialty_name
12339	29	5	Cardiologist
1602	51	5	Cardiologist
1604	22	5	Cardiologist
1630	25	5	Cardiologist
1634	31	5	Cardiologist
1637	73	5	Cardiologist
1637 1644	21	5	Cardiologist
101	100	5	Cardiologist
101 1637	21	5	Cardiologist
101 105	32	5	Cardiologist

all other services. That's why prediction engine gives null value for this test case. Validation of proposed methdology based on hospital data has revealed interesting anomalies. Few of these anomalies have been identified as fraud cases by our methodology. Accuracy A is computed for the system in terms of percentage using equation 8 :

$$Accuracy\ A = \frac{TP + TN}{Total\ number\ of\ \epsilon in\ \varphi_u} * 100 \qquad (8)$$

where TP is true positive, which means sequence is fraud sequence and identified as fraud sequence. TN is true negative which means sequence is normal sequence and

TABLE 10. Subset of frequent sequences for dentistry.

Sequence	Frequency	Specialty_id	Specialty_name
314 239	4	100	Dentistry
314 1969	19	100	Dentistry
314 1969 316	2	100	Dentistry
314 314	8	100	Dentistry
314 315	8	100	Dentistry
314 316	65	100	Dentistry
314 316 233	2	100	Dentistry
314 316 238	3	100	Dentistry
314 316 314	2	100	Dentistry
314 316 316	13	100	Dentistry
314 316 316 316	5	100	Dentistry
314 316 316 316 316	2	100	Dentistry
314 316 317	6	100	Dentistry
314 316 6878	2	100	Dentistry
314 317	26	100	Dentistry

TABLE 11. Different minimum support values for prediction based engine.

Minimum support	Text set	Predicted Sequence
2	1	1 7 7 7
15	1	1 10 7 4412 4413 1969
25	1	1 10 7 4412

identified as normal. ϵ is representing sequences in specialty φ. U is the counter for specialties as we are considering 62 specialties. Average accuracy for all specialties is 85%. Each identifed case is forwarded to analyst dashboard for further analysis. Analyst can append, approve and reject the identified case. The purpose of generating frequent sequence based rule engine is to identify the existence of any anomaly in case any patient has availed one sequence of services from any specialty for many or one time only. But at the same time there is possibility that identified anomaly is not infact an anomaly. By considering this point, we have designed prediction based engine which performs fraud detection by prediciting group of services for each considered test set.

B. OBSERVATIONS

1) Firstly, for the sake of experimentation we trained Prediction based Engine on frequent sequences for each Specialty. During experimentation we found that whenever there is change in minimum support values for generating frequent sequences and fed these frequent sequences as an input to Prediciton based engine and trained it, Prediction based engine predicts items accordingly. Table 11 exhibits this observation more clearly. After this experiment, we provided sequence database as a training set to Prediction based engine.
2) The one observation in proposed methodology is that if there is an anomalous service availed within the sequence, after that service other services can be availed from considered specialty then sequence rule engine identify it as anomalous sequence but prediction based engine can mark it as normal because it will not predicts null in that case. To handle this issue, we introduced the step of entering sequence in the form of vector as input to compact prediction tree in prediction based engine.
3) When number of records increases, minimum support value increases accordingly and it is observed that number of detected anomalies also increases as depicted in Fig 17.
4) The one more observation which has been identified during analysis, When we apply frequent mining algorithm on sequence database and keep any support values, we get sequences which mostly contain service_id = 13556 and service_id = 1969. As already mentioned these two serviceids are Co-insurance and Rx-services OPD with service -Major -description Miscellaneous and Pharmacy respectively. There is a possibility that these are going to be availed in almost all specialties but strange thing is that they are among most frequently availed services in all sequences. We communicated this anomalous behavior to hospital management.
We have included services related to clinics, laboratory and radiology and consultation.

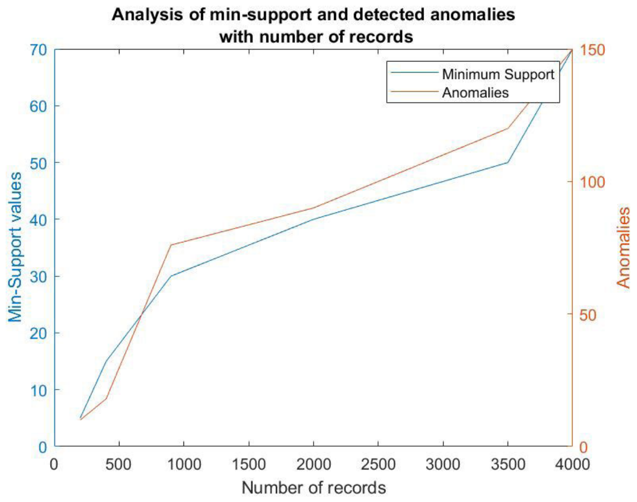


FIGURE 17. Relationship among input and output parameters.

Few cases of frauds are shown in Table 12. Our sequential mining methodology has revealed that it is practicable to utilize the concept of frequently occurring sequences of clinical services which are being availed by patients and administered to patients.

Although, the length of these clinical service sequence patterns is typically not fixed and significant heterogeneity has been noticed in the way clinical services are delivered across the same specialty. However, relating each visit of patient with his/her previous visit in each specialty and then extracting and analyzing anomalies by using rule engine and prediction based engine not only facilitates decision making within healthcare delivery

TABLE 12. Subsets of fraud cases are depicted in few specialties.

Specialty name	MR_No	Sequence Identified as Fraud
Cardiology	20382E	Pelvis exam, Insulin, thyroid stimulating hormone test, prolactin (PRL) test, Thyroid Profile (TT3, FT4, TSH)
	954703	Vitamin B 12, thyroid stimulating hormone test, 25-Hydroxy Vitamin D, CBC Diff Profile (CS11), ECG 12 Lead, Creatine Kinase, Cardiac Profile (CPKMB, Troponin-I), Cardiac Profile (CPKMB, Troponin-I), Cardiac Profile (CPKMB, Troponin-I)
ENT specialist	03421C	Glucose Fasting, LIPID Profile(HDL, LDL, Chlstrl, Trig), CBC Diff Profile (CS11)
Gastroenterogist	50280	Chest Xray 1 View, thyroid stimulating hormone test, LFT, ESR

and practice but also helps in detection of fraudulent practices.

V. CONCLUSION

Many developing countries have recently initiated government medical support programs which incorporate less tolerance for any fraudulent claims. Fraud detection in healthcare is a non-trivial task due to the heterogeneous nature of healthcare records. Fraudsters behave as normal patients and with the passage of time keep on changing their way of planting frauds. Therefore, there is a critical need to design a system that is able to capture and identify fraud cases in day to day transactions in the healthcare industry To the best of our knowledge, only few studies have utilized sequential pattern mining for predicting and detecting frauds in healthcare industry. Rest of the researches have utilized financial information for detecting frauds. However our proposed methodology relies on the novel idea of analyzing patient time series traces for detecting fraud at each specialty level. We proposed a framework for fraud detection in clinical service processes. For this purpose, we have used prefixspan sequence mining approach and bayes rule for populating frequent and rare sequences in Sequence rule Engine based on sequence database of patient time series traces and particular patient traces for specific specialty. Analysis of medical behaviors in clinical processes has led to identification of anomalous sequences as patients deviate from sequences contained in Sequence rule Engine. In other words, we are able to detect sequences that deviate from frequent medical behaviors or it can be less frequent behavior. Once anomalous sequences are identified, these sequences are further analyzed by Prediction based Engine to detect fraudulent cases. Various meetings have been arranged with medical domain experts for upgrading, and evaluating the proposed methodology in clinical settings. The results of validation of this methodology, combined with the concept that both patient and physician can commit the fraud, have shown that our proposed methodology is efficient and capable of identifying even such fraudulent cases that are not detected by existing or manually constructed detection model. The design is able to provide average accuracy upto 85% in detecting frauds. The data set we used to validate the proposed methodology was difficult to obtain as it contains private and confidential information related to employee's insurance data. The data set was in raw form and to handle the missing and redundant information was also time consuming. We use five years transactional dataset but to check the competence of the methodology, larger datasets would be more useful and will show better visualization and strength of proposed methodology in larger perspective.

ACKNOWLEDGMENT

The authors would like to thank Shifa International Hospital, Islamabad, Pakistan for providing the employee's insurance data to validate our proposed methodology. This work is part of PM Task Force on IT and Telecom initiative for "Sehat Card Scheme" so that genuine people can easily get medical benefits from this government level initiative and is presented as an initial proposed methodology to the taskforce for fraud detection.

REFERENCES

[1] N. Aldrich, J. Crowder, and B. Benson, "How much does medicare lose due to fraud and improper payments each year?" *Sentinel*, 2014.

[2] A. K. Jain, "Data clustering: 50 Years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, Jun. 2010.

[3] C. Ou-Yang, C. P. Wulandari, R. A. R. Hariadi, H.-C. Wang, and C. Chen, "Applying sequential pattern mining to investigate cerebrovascular health outpatients' re-visit patterns," *PeerJ*, vol. 6, p. e5183, Jul. 2018.

[4] I. Matloob and S. Khan, "A framework for fraud detection in government supported national healthcare programs," in *Proc. 11th Int. Conf. Electron., Comput. Artif. Intell. (ECAI)*, Jun. 2019, pp. 1–7.

[5] M. L. McCarthy, D. Aronsky, I. D. Jones, J. R. Miner, R. A. Band, J. M. Baren, J. S. Desmond, K. M. Baumlin, R. Ding, and R. Shesser, "The emergency department occupancy rate: A simple measure of emergency department crowding?" *Ann. Emergency Med.*, vol. 51, no. 1, pp. 15–24, 2008.

[6] N. A. Nuaimi, "Data mining approaches for predicting demand for healthcare services in abu dhabi," in *Proc. 10th Int. Conf. Innov. Inf. Technol. (IIT)*, Nov. 2014, pp. 42–47.

[7] H. Batal, J. Tench, S. Mcmillan, J. Adams, and P. S. Mehler, "Predicting patient visits to an urgent care clinic using calendar variables," *Academic Emergency Med.*, vol. 8, no. 1, pp. 48–53, Jan. 2001.

[8] S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, "Data mining techniques and applications–A decade review from 2000 to 2011," *Expert Syst. Appl.*, vol. 39, no. 12, pp. 11303–11311, 2012.

[9] D. P. K. Ng, B. C. Tai, D. Koh, K. W. Tan, and K. S. Chia, "Angiotensin-I converting enzyme insertion/deletion polymorphism and its association with diabetic nephropathy: A meta-analysis of studies reported between 1994 and 2004 and comprising 14,727 subjects," *Diabetologia*, vol. 48, no. 5, pp. 1008–1016, May 2005.

[10] A. Taneja, "Heart disease prediction system using data mining techniques," *Oriental J. Comput. Sci. Technol.*, vol. 6, no. 4, pp. 457–466, 2013.

[11] R. Ito *et al.*, "Comparison of cystatin C-and creatinine-based estimated glomerular filtration rate to predict coronary heart disease risk in Japanese patients with obesity and diabetes," *Endocrine J.*, vol. 62, no. 2, pp. 201–207, 2014, doi: 10.1507/endocrj.EJ14-0352.

[12] J. Tuomi, K.-S. Paloheimo, J. Vehviläinen, R. Björkstrand, M. Salmi, E. Huotilainen, R. Kontio, S. Rouse, I. Gibson, and A. A. Mäkitie, "A novel classification and online platform for planning and documentation of medical applications of additive manufacturing," *Surgical Innov.*, vol. 21, no. 6, pp. 553–559, Dec. 2014.

[13] Y. Tokura, O. Yoshino, S. Ogura-Nose, H. Motoyama, M. Harada, Y. Osuga, Y. Shimizu, M. Ohara, T. Yorimitsu, O. Nishii, S. Kozuma, and T. Kawamura, "The significance of serum anti-Müllerian hormone (AMH) levels in patients over age 40 in first IVF treatment," *J. Assist. Reproduction Genet.*, vol. 30, no. 6, pp. 821–825, 2013.

[14] C. Ou-Yang, S. Agustianty, and H.-C. Wang, "Developing a data mining approach to investigate association between physician prescription and patient outcome–A study on re-hospitalization in Stevens–Johnson Syndrome," *Comput. Methods Programs Biomed.*, vol. 112, no. 1, pp. 84–91, 2013.

[15] X. Li, H. Cao, E. Chen, H. Xiong, and J. Tian, "BP-growth: Searching strategies for efficient behavior pattern mining," in *Proc. IEEE 13th Int. Conf. Mobile Data Manage.*, Jul. 2012, pp. 238–247.

[16] R. M. Musal, "Two models to investigate medicare fraud within unsupervised databases," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8628–8633, Dec. 2010.

[17] W.-S. Yang and S.-Y. Hwang, "A process-mining framework for the detection of healthcare fraud and abuse," *Expert Syst. Appl.*, vol. 31, no. 1, pp. 56–68, Jul. 2006.

[18] J. Liu, E. Bier, A. Wilson, J. A. Guerra-Gomez, T. Honda, K. Sricharan, L. Gilpin, and D. Davies, "Graph analysis for detecting fraud, waste, and abuse in healthcare data," *AI Mag.*, vol. 37, no. 2, pp. 33–46, 2016.

[19] I. Batal, D. Fradkin, J. Harrison, F. Moerchen, and M. Hauskrecht, "Mining recent temporal patterns for event detection in multivariate time series data," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2012, pp. 280–288.

[20] C. Liu, F. Wang, J. Hu, and H. Xiong, "Temporal phenotyping from longitudinal electronic health records: A graph based framework," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, pp. 705–714, 2015.

[21] A. Arora, A. Srivastava, and S. Bansal, "Business competitive analysis using promoted post detection on social media," *J. Retailing Consum. Services*, vol. 54, May 2020, Art. no. 101941.

[22] A. Taneja, P. Gupta, A. Garg, A. Bansal, K. P. Grewal, and A. Arora, "Social graph based location recommendation using users' behavior: By locating the best route and dining in best restaurant," in *Proc. 4th Int. Conf. Parallel, Distrib. Grid Comput. (PDGC)*, 2016, pp. 488–494.

[23] G. E. A. P. A. Batista, X. Wang, and E. J. Keogh, "A complexity-invariant distance measure for time series," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2011, pp. 699–710.

[24] C. Liu, K. Zhang, H. Xiong, G. Jiang, and Q. Yang, "Temporal skeletonization on sequential data: Patterns, categorization, and visualization," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 1, pp. 211–223, Jan. 2016.

[25] L. Sun, C. Liu, C. Guo, H. Xiong, and Y. Xie, "Data-driven automatic treatment regimen development and recommendation," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1865–1874.

[26] M. Gupta, J. Gao, C. Aggarwal, and J. Han, "Outlier detection for temporal data," *Synth. Lectures Data Mining Knowl. Discovery*, vol. 5, no. 1, pp. 1–129, Mar. 2014.

[27] J. Yang, C. Liu, M. Teng, H. Xiong, M. Liao, and V. Zhu, "Exploiting temporal and social factors for B2B marketing campaign recommendations," in *Proc. IEEE Int. Conf. Data Mining*, Nov. 2015, pp. 499–508.

[28] J. Jurgovsky, M. Granitzer, K. Ziegler, S. Calabretto, P.-E. Portier, L. He-Guelton, and O. Caelen, "Sequence classification for credit-card fraud detection," *Expert Syst. Appl.*, vol. 100, pp. 234–245, Jun. 2018.

[29] R. Saia, "Unbalanced data classification in fraud detection by introducing a multidimensional space analysis," in *Proc. IoTBDS*, 2018, pp. 29–40.

[30] S. Carta, G. Fenu, D. Reforgiato Recupero, and R. Saia, "Fraud detection for E-commerce transactions by employing a prudential multiple consensus model," *J. Inf. Secur. Appl.*, vol. 46, pp. 13–22, Jun. 2019.

[31] G. Liu, J. Guo, Y. Zuo, J. Wu, and R.-Y. Guo, "Fraud detection via behavioral sequence embedding," *Knowl. Inf. Syst.*, vol. 62, pp. 2685–2708, Jan. 2020.

[32] T. Slimani and A. Lazzez, "Sequential mining: Patterns and algorithms analysis," 2013, *arXiv:1311.0350*. [Online]. Available: http://arxiv.org/abs/1311.0350

[33] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M.-C. Hsu, "FreeSpan: Frequent pattern-projected sequential pattern mining," in *Proc. 6th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2000, pp. 355–359.

[34] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu, "Mining sequential patterns by pattern-growth: The PrefixSpan approach," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 11, pp. 1424–1440, Nov. 2004.

[35] N. R. Mabroukeh and C. I. Ezeife, "A taxonomy of sequential pattern mining algorithms," *ACM Comput. Surveys*, vol. 43, no. 1, pp. 1–41, Nov. 2010.

[36] C. Antunes and A. L. Oliveira, "Sequential pattern mining algorithms: Trade-offs between speed and memory," INESC-ID, Dept. Inf. Syst. Comput. Sci., MGTS, Lisboa, Portugal, 2004.

[37] M. J. Zaki, "Spade: An efficient algorithm for mining frequent sequences," *Mach. Learn.*, vol. 42, nos. 1–2, pp. 31–60, Jan. 2001.

[38] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Proc. 11th Int. Conf. Data Eng.*, Mar. 1995, pp. 3–14.

[39] J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent pattern mining: Current status and future directions," *Data Mining Knowl. Discovery*, vol. 15, no. 1, pp. 55–86, Jul. 2007.

[40] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. 20th Int. Conf. Very Large Data Bases (VLDB)*, vol. 1215, 1994, pp. 487–499.

[41] M. J. Islam, Q. M. Jonathan Wu, M. Ahmadi, and M. A. Sid-Ahmed, "Investigating the performance of Naive- bayes classifiers and K- nearest neighbor classifiers," in *Proc. Int. Conf. Converg. Inf. Technol. (ICCIT)*, Nov. 2007, pp. 1541–1546.

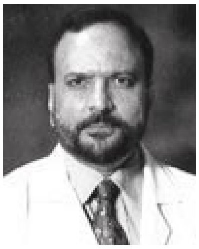
[42] J. Seo and O. Mendelevitch, "Identifying frauds and anomalies in Medicare-B dataset," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2017, pp. 3664–3667.

[43] K. Malhotra, T. C. Hobson, S. Valkova, L. L. Pullum, and A. Ramanathan, "Sequential pattern mining of electronic healthcare reimbursement claims: Experiences and challenges in uncovering how patients are treated by physicians," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Oct. 2015, pp. 2670–2679.

[44] A. Verma, A. Taneja, and A. Arora, "Fraud detection and frequent pattern matching in insurance claims using data mining techniques," in *Proc. 10th Int. Conf. Contemp. Comput. (IC3)*, Aug. 2017, pp. 1–7.

IRUM MATLOOB received the M.S. degree in computer software engineering from the National University of Sciences and Technology (NUST), Islamabad, in 2012, where she is currently the pursuing the Ph.D. degree. She has been a permanent Lecturer with Fatima Jinnah Women University, since 2014. Her research interests include data mining, health informatics, trend analysis, system design and testing, and machine learning algorithms.

SHOAB AHMED KHAN received the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA. He is currently a Professor of computer and software engineering with the College of Electrical and Mechanical Engineering, National University of Sciences and Technology (NUST). He is an inventor of five awarded U.S. patents and has over 260 international publications. His book on digital design is published by John Wiley and Sons and is being followed in national and international universities. He has more than 22 years of industrial experience in companies in USA and Pakistan. He has also served as a member of the National Computing Council and the National Curriculum Review Committee. He received the TamgheImtiaz (Civil), the Highest National Civil Award in Pakistan, the National Education Award, in 2001, and the NCR National Excellence Award in Engineering Education. He is the Founder of the Center for Advanced Studies in Engineering (CASE) and the Center for Advanced Research in Engineering (CARE). CASE is a primer engineering institution that runs one of the largest postgraduate engineering programs in the country and has already graduated 50 Ph.D. students and more than 1800 M.S. students in different disciplines in engineering, whereas CARE, under his leadership, has risen to be one of the most profound high technology engineering organizations in Pakistan developing critical technologies worth millions of dollars for organizations in Pakistan. CARE has made history by winning 13 PASHA ICT awards and 11 Asia Pacific ICT Alliance Silver and Gold Merit Awards while competing with the best products from advanced countries like Australia, Singapore, Hong Kong, Malaysia, and so on. He has served as the Chairman of the Pakistan Association of Software Houses (PASHA) and a member of the Board of Governance of many entities in the Ministry of IT and Commerce.



HABIB UR RAHMAN received the M.B.B.S. degree from the King Edward Medical College, Lahore, Pakistan, in January 1971. He is currently a Diplomate of the American Board of Internal Medicine and the American Board of Cardiovascular Diseases. He is a Highly Qualified Doctor and practices as a Cardiologist with Shifa International Hospital Ltd., Islamabad, Pakistan. He was a Fellow of American College of Cardiology and also received the Fellowship of the Mount Sinai Hospital, Hartford, CT, USA. He has memberships of the American College of Physicians and Pakistan Cardiac Society Life Savers Foundation. He is Chief/Consultant with the Department of Cardiology Shifa International Hospitals Ltd., Islamabad, since September 1995. He went to Byrd Regional Hospital, Leesville, LA, USA, and Lake Charles Memorial Hospital, Lake Charles, LA, USA, from November 1989 to July 1995. He was an Attending Physician of adult cardiovascular disease with the Graham Hospital, Canton, IL, USA, and the Methodist Medical Center, Peoria, IL, USA, from August 1982 to September 1989.