Received March 16, 2022, accepted April 20, 2022, date of publication April 28, 2022, date of current version May 10, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3170888

A Sequence Mining-Based Novel Architecture for Detecting Fraudulent Transactions in Healthcare Systems

IRUM MATLOOB[1,2], SHOAB AHMED KHAN[2], RUKAIYA RUKAIYA[3], MUAZZAM A. KHAN KHATTAK[4], (Senior Member, IEEE), AND ARSLAN MUNIR[5], (Senior Member, IEEE)

[1]Department of Software Engineering, Fatima Jinnah Women University (FJWU), Rawalpindi 46000, Pakistan
[2]Department of Computer and Software Engineering, National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan
[3]Department of Computer Engineering, Sir Syed University of Engineering and Technology, Karachi 75300, Pakistan
[4]Department of Computer Science, Quaid-i-Azam University, Islamabad 45320, Pakistan
[5]Department of Computer Science, Kansas State University, Manhattan, KS 66506, USA

Corresponding authors: Irum Matloob (irummatloob@yahoo.com) and Muazzam A. Khan Khattak (khattakmuazzam@gmail.com)

This research was supported in part by the Pakistan Government Level Initiative and in part by Shifa International hospital, Islamabad, Pakistan. The publication of this article was financed with support from the Kansas State University Open Access Publishing Fund.

ABSTRACT With the exponential rise in government and private health-supported schemes, the number of fraudulent billing cases is also increasing. Detection of fraudulent transactions in healthcare systems is an exigent task due to intricate relationships among dynamic elements, including doctors, patients, and services. Hence, to introduce transparency in health support programs, there is a need to develop intelligent fraud detection models for tracing the loopholes in existing procedures, so that the fraudulent medical billing cases can be accurately identified. Moreover, there is also a need to optimize both the cost burden for the service provider and medical benefits for the client. This paper presents a novel process-based fraud detection methodology to detect insurance claim-related frauds in the healthcare system using sequence mining concepts. Recent literature focuses on the amount-based analysis or medication versus disease sequential analysis rather than detecting frauds using sequence generation of services within each specialty. The proposed methodology generates frequent sequences with different pattern lengths. The confidence values and confidence level are computed for each sequence. The sequence rule engine generates frequent sequences along with confidence values for each hospital's specialty and compares them with the actual patient values. This identifies anomalies as both sequences would not be compliant with the rule engine's sequences. The process-based fraud detection methodology is validated using last five years of a local hospital's transactional data that includes many reported cases of fraudulent activities.

INDEX TERMS Fraudsters, health insurance, healthcare, medical benefits, premium, sequence mining.

I. INTRODUCTION

The aim of healthcare ecosystems is to provide quality healthcare services which are business entities, such as set of medicines, tests, or procedures conducted during patients' treatment. In health informatics, information technology resources are applied to solve healthcare-related issues. Business Intelligence helps in making business-related decisions based on historical results. Worldwide, most governments are supporting their citizens by introducing medical support programs. Such medical support schemes give relief to their citizens. Moreover, many large enterprises and companies offer medical benefits coverage to their employees via insurance policies. In such cases, enterprises pay massive premium amounts, and employees get healthcare services. However, there is a long list of healthcare service quality issues among which the top of the list is healthcare fraud.

Healthcare fraud is the main obstacle in achieving medical benefit optimization. It is the broader term for three concepts: Fraud, Waste, Abuse (FWA) and an intentional act of getting unauthorized benefits. A doctor can prescribe unnecessary medicines and laboratory tests to increase hospital revenue; this is an example of waste. Similarly, a pharmacist who charges both patients and insurance companies for the same

The associate editor coordinating the review of this manuscript and approving it for publication was Tai-Hoon Kim.

prescription is commencing a fraud. The pharmacist commits healthcare abuse if he/she receives prescription for a particular brand, and enters that medicine brand into the computer system to charge the insurance company, while in reality, he/she gives cheaper brand medicine to the patient. Many such deceptive acts come under the generic description of healthcare FWA.

Healthcare has become a significant source of financial expenditure in most countries. According to National Healthcare Anti-Fraud Association (NHCAA), healthcare fraud causes loss in the tens of billions of dollars annually [1]. One of the reasons organizations face critical losses is that they pay a considerable amount as a premium to insurance companies. The employees are not availing their complete sum insured amount. We note that the *sum insured* in health insurance denotes the maximum amount for a particular year payable by the insurance company to the insuree (or employee) in case of hospitalisation. Moreover, big or small sum insured amount also results in healthcare frauds. It is critically important for organizations or enterprises to analyze their employees' needs before selecting insurance policies for their employees.

Rising health expenditures for patient treatment force hospitals to introduce new approaches for efficient healthcare services delivery. Medical benefit optimization is the best solution for this emerging problem. Providers like doctors, hospitals, or enterprises, perform medical benefit optimization with the help of data analytics to optimize their healthcare service quality and financial performance. During the optimization, the special focus is on anomaly or fraud detection. The overall scope of this research is described in Figure 1.

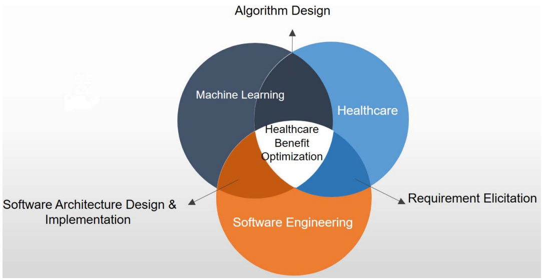


FIGURE 1. Overall scope of the research.

This paper provides a considerable understanding of recurring patient visits to each specialty (department) and the extraction of patients visit patterns. These patterns create awareness among doctors and hospital management regarding possible medical services on subsequent patient visits in each specialty {cardiology, dentistry, neurology, urology etc}. It also enables service providers to take preventive actions in the case of anomalous behavior. The proposed methodology can analyze more than 60 specialties of a hospital and can predict the separate set of sequences for each specialty. With the help of time series traces and prefixspan algorithm, the sequence rule engine is generated, based sequences of different pattern length and frequency, that helps in identifying misutilization of healthcare services. This research would be beneficial for improving the quality of healthcare services.

A. RESEARCH CONTRIBUTIONS

Our main contributions in this article are as follows:

- This paper proposes a sequence mining-based novel architecture for detecting fraudulent transactions in healthcare systems.
- This paper provides a considerable understanding of recurring patient visits to each specialty department) and the extraction of patients visit patterns.
- The patient visit patterns extracted in this research help create awareness among doctors and hospital management regarding possible medical services on subsequent patient visits in each specialty. This also enables service providers to take preventive actions in the case of anomalous behavior.
- The proposed fraud detection methodology can analyze more than 60 specialties of a hospital and can predict the separate set of sequences for each specialty.
- A significance of using patient time series traces is that we can get healthcare utilization details of each patient. With the help of time series traces and sequence mining algorithm, the sequence rule engine (knowledge base) is generated based on the sequences with different pattern length, which aids in determining the utilization of healthcare services.

The remainder of this paper is organized as follows. Section II discusses the research gaps in literature related to detecting fraudulent transactions in healthcare systems. Section III elucidates motivation for this work. Recent research works in detecting fraudulent transactions in healthcare systems are summarized in Section IV. Section V presents the proposed software design and architecture for fraud detection. Section VI elaborates the proposed methodology for detecting frauds in real-time. Section VII presents detailed results and analysis of these results. Conclusions are provided in Section VIII. Finally, Section IX presents limitation of current study, and provide directions for future research in this area.

II. GAP ANALYSIS

After conducting a detailed review of the related literature, we conclude that many authors have proposed solutions for fraud detection in healthcare and have also applied the data-mining, specifically machine-learning algorithms. Large numbers of fraud detection research projects are successfully conducted worldwide from local to national levels to control healthcare fraud. The research projects vary in data sets, healthcare frauds, analysis scale and techniques. Based on reviewed recent literature, fraud detection framework and methodology provides opportunities for exploring more efficient solutions for following issues while performing fraud detection:

1) Many research considers the clinical processes or disease correlations with medications to detect fraud for a particular disease.
2) Most of the existing studies use the domain knowledge to make the knowledge base, but there is a need for a system that can learn knowledge from the historical medical records using machine learning techniques. Frauds can be easily classified using such a self-learned knowledge base.
3) Most of the previous research is based on the financial analysis for detecting fraudulent activities, but there is a need for research that identifies anomalies using the association/linkages among healthcare ecosystem elements.
4) 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 research is service, either provided or availed.
5) In general medicine, many research projects are based on predicting patients' demands. Moreover, less focus is on the patient's behavior. There is a dire need to analyze the sequence of services that a patient avails from the hospital's specific specialty.
6) Healthcare Fraud cannot be effectively detected without studying patients' sequences of services availed from different specialties.

There are many research projects on prediction, analysis, and evaluation of healthcare services, healthcare operations, and resources. Many studies are based on machine learning algorithms for increasing healthcare services efficiency, and it seems that the trend of applying machine learning in the healthcare industry is increasing. The increasing cost in the healthcare industry can be handled by controlling frauds and preventing frauds. The fraud can only be controlled if it is detected at the right time. Once the fraud is detected, it is necessary to isolate it.

III. MOTIVATION

The detection of healthcare frauds is a non-trivial task and the designed system must be able to detect and isolate medical frauds on the spot. According to the Indiana Manufacturers Association and We Ask America's first public poll (January 2019), the reason for high hospital costs are healthcare frauds, as shown in Figure 2. The rising number of healthcare fraudulent cases has forced the healthcare insurance industry to incorporate data analytics for healthcare fraud detection and prevention. This has caused the medical data analytics market to grow rapidly [2]. According to the National Healthcare Anti-Fraud Association, the annual cost of health insurance fraud in the United States is approximately USD 80 billion.

Now a days, governments are spending massive amount of money in healthcare and a lot of capital is being allocated, especially in developing and low-income countries. According to the World Health Organization (WHO) 2016 report, the annual rise in the healthcare provision cost in developing countries is around 6%, which is greater than the 4% in developed countries. However, Pakistan is spending only 3% of its gross domestic product (GDP) on the education sector, health, and nutrition. According to the World Bank report (April 2017), this figure is significantly less than other countries. In Pakistan, unfortunately, the healthcare sector is one of the most neglected sector. Since the last decade, the government allocates 0.5pc to 0.8pc of its GDP for the health sector, which is less than the WHO benchmark (6pc of GDP). A few years back, Dr. Fateh M Khan, former director-general of Health Services in Sindh [3], suggested there should be some correlation between services and their corresponding specialist.

In most countries, including Pakistan, the government has just initiated health support programs through several national-level initiatives. One of these initiatives is the establishment of the Prime Minister Task Force on IT and Telecom in 2018 to lay down the foundation of the data standards and annotations for incorporating the state-of-the art plans in healthcare service delivery to the common person. A major area that is to be targeted is to reduce or prevent the chance of fraudulent activities in government-supported healthcare programs. We can achieve this by implementing different adaptive-modeling techniques to detect fraud activities from healthcare data. The challenges faced while establishing the authenticity of healthcare insurance claim data are addressed. There is a need to design a practical methodology that can handle all these issues and distinguish regular patients from fraudsters. Much research is carried out to identify insurance claim frauds [4]–[7]. However, most of the prior research focuses on either disease- and medication-related issues, or consider one or two specialties for fraud detection. Through extensive studies of medical support programs, we observe a dire need to analyze historical medical records and the sequence of services that patients avail from a specific specialty of the hospital. For example, if any false claim is made, the patient sequence can be analyzed to detect anomaly or fraud in the process.

IV. BACKGROUND AND RELATED WORK

The knowledge of outlier detection algorithms is important before the study of literature related to fraud detection. Outlier detection algorithms are generally classified into two broad categories: The first class of algorithms focuses on the identification of anomalies in individual data points and the second class of algorithms considers the data as the 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 the algorithm generates a model which can be used for real-time fraud detection. Proximity-based algorithms scan through the whole database for detecting fraud, but such approaches are not recommendable in a real-time environment [8].

Like all other industries, there is a critical need to monitor the clinical care processes and funds utilization in the

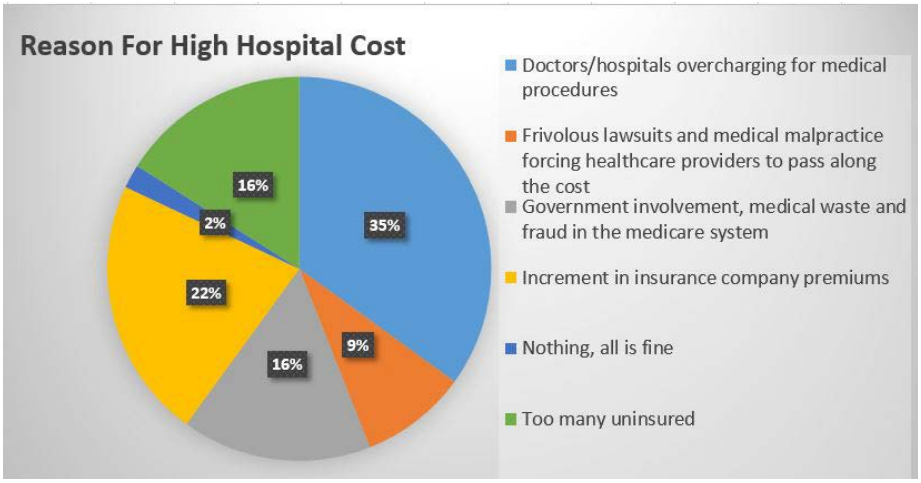


FIGURE 2. **Reasons for rising healthcare costs [1].**

healthcare ecosystem. The curse of fraud has covered all industries, but its damage to healthcare is inevitable. A lot of research have been conducted to detect fraud activities in healthcare. Fraud detection in healthcare can be divided into three levels. First level is fraud detection in hospital processes, second is fraud detection in disease diagnosis, and third is actor-level fraud detection.

A survey on recent literature shows that the trend of application of machine learning and statistical approaches for analyzing healthcare processes is increasing rapidly. In [9], a framework based on clinical pathways for automatic generation of fraud detection models is proposed. The analysis is performed on National Health Insurance (BNHI) in Taiwan. The model detected 69% frauds correctly but unable to detect cases with overdoses of medicines. In [10], graph theory-based analysis is conducted to identify fraud and waste cases in healthcare records. The Narcotics Relation Graph is created which is based on the three main entities doctor, patient, and pharmacy. The case of referral networks is also considered in which nodes are the providers. The links between nodes represent the number of referral between the two nodes. The heterogeneous graphs are created for understanding complex relationships among these entities. These relationships are used to detect millions of anomalous cases and results in recovery of funds. In [11], fraud detection via geographical analysis using clustering algorithms is performed. The goal of this research is to detect fraud cases by infusion therapy providers in medicare. The main clinical processes for performance improvement are studied in [12]. Patient care logs for analyzing patient care and treatment patterns. Then, density based clustering is applied to create model for detecting anomalies. From all these researches, it is clearly understood that fraud detection is important for the performance improvement of hospital processes. Data mining-based analysis technique analysis helps to detect fraud. In [13], detailed survey is provided on techniques used for fraud detection using medicare datasets.

In [14], the fraud detection model is proposed and disease-drug based relationship is used to identify outliers. The k-means and isolation forest algorithms are used on different datasets and it is identified that isolation forests perform better than k-means clustering for the model. In [15], the disease-based outliers are used to detect the fraud-related activities. The statistical rules are used to detect disease-based and period-based outlier which considered all the outliers as frauds. Most of the researches related to anomaly detection in healthcare have considered the clinical processes for a particular disease and utilized prior knowledge and applied the unsupervised models [16] [17]. In [18], [19], frameworks for fraud detection are proposed, and the focus of the authors is on the correlation of medicines, diseases, and patients. Frauds are detected by assigning weights to highly correlated data. Many authors have utilized graph theory to connect the entities, i-e, patients, diseases, and medicines. A correlation between the reference set (the actual knowledge) and the candidate set (the extracted knowledge) is proposed. The main problem is the availability of data. Most of the times, the studies are supplemented with the prior knowledge of the medicines that were being used for the various diseases.

Many researchers proposed actor-level fraud detection models for detecting frauds related to providers or patients. The provider-level fraud has a more critical effect on the healthcare expenditure as compared to patient-level fraud [20]. In [21], the unsupervised Bayesian hierarchical based methods are used for detecting frauds in medical claims. Information related to patients, doctors, and the bills are contained in the medical claims. The hierarchical model is used to detect hidden patterns existing between providers and medical procedures. In [22], provider's prescription patterns are used for detecting prescriber-related-drug frauds. Topic modeling is performed for detecting irrelevant or extra prescriptions. It is achieved by creating topic models which are used to group drugs with respect to the billing patterns and then covariation with the medical specialties is performed.

The output of this system enables the auditors to identify providers which are prescribing unnecessary medicines. Moreover, in [23], frauds are detected without considering the providers' and clients' roles. Machine learning-based system is developed that involves hierarchical processing along with assigning the weight to actors. The expectation-maximization clustering technique finds out the related groups of the actors. The main issue is that types of frauds are already defined using storyboards in this research. In [24], the rational treatment model for some diseases using graph mining and frequent pattern mining techniques is proposed. The copying prescription problem is also dealt for assessing the doctors' trustworthiness which is one of the critical metrics for detecting fraud at the provider level. In most of the previous research, not all the actors are considered in the analysis, whereas fraud is the joint act of multiple healthcare actors. Analysis based on the association among these actors remains neglected.

In [25], the detailed evaluation of machine learning algorithms in the health insurance industry is performed and it is suggested that random forest algorithm is more effective with detection rate of 23.8 as compared to other approaches. In [26] and [27], the role of statistical and data mining techniques is discussed; with these techniques, hidden information from the historical data can be extracted and analyzed. Machine learning techniques learn from the existing data and use the learned knowledge for future decision-making. These researches provided accurate fraud detection frameworks based on data mining techniques. The research is based on a detailed survey of the statistical approaches. These approaches are still being applied to identify and classify frauds in healthcare. Ortega *et al.* in [28] 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. In [29], the fraud detection model which is based on genetic support vector machines is developed. The model detects anomalies and also classifies fraudulent health insurance claims. The model achieved average accuracy of 87.91%. Liu and Vasarhelyi in [30] considered a clustering model which is based on the geographical location of Medicaid service providers and clients to identify fraudulent claims. In [31], the novel methodology is proposed in which medical specialties are encoded by applying procedure-level statistics. In [32], the technique based on primitive sub peer group analysis is proposed for detection of fraudulent behaviors from health insurance claims in [33]. The two step algorithm is proposed for detecting fraud cases. In first step, feature selection (principal component analysis) is performed followed by clustering and in the second step human decision support system is proposed.

In [34], a framework is proposed, which introduced an adaptable model using clinical ways for automatic fraud detection. The structured approach is used for fraud detection. This approach follows the clinical sequence of treatments for patients in the gynecology department using a graph mining algorithm. The expense feature and other features are selected as discriminating features for performing fraud detection. Therefore, adjustment of the model according to the site-specific cost policies is required. In [35], the framework for the fraud detection using unsupervised learning to detect the outliers in medicaid insurance-claimed data is proposed. Thorton *et al.* in [36] applied the multidimensional data models and approaches for predicting 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 in [37]–[43]. All the research has focused on 'PART- B' of this data (PUF). Statistical techniques are also used to generate decision rules, and k-means clustering is applied on time series-based insurance claim data to identify anomalies and outliers. Many researchers have focused on statistical, financial data and performed analytics using various tools. Fuzzy and Neuro-fuzzy analysis is performed in multiple types of researches for extracting interesting patterns [44]–[46]. In [47], the clustering technique is applied to identify the joint fraudsters and then the similarity adjacency graph is used along with group mining for distinguishing the normal behavior from abnormal behaviors.

In [48], the association rule mining is an essential technique that 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. The two most useful parameters to analyze the association rules' strength are namely: confidence and support [49]–[51]. The characteristics like uniqueness, understandability, applicability, and reliability for assessing the generated rules are discussed in [52]. Ou-Yang *et al.* in [53], have performed association rule mining on doctors prescriptions. Table 1 compares the existing state of the art designs that are proposed to detect actor-level frauds in healthcare systems.

After the review of the literature related to the three levels of fraud detection in healthcare, we are going to discuss fraud detection systems which are recently used in other industries. In [55], a detailed survey is performed on the supplementation of fraud detection systems in many other industries. In [56], fraud detection model using prudential multiple consensus model is developed, for detecting fraudulent transactions in E-commerce industry. The model is validated on real world dataset which contains high degree of imbalance data and results of validation shows that the ensemble model outperforms the other state of the art models. In [57], a novel LSTM based approach is proposed in which the learned sequential embedding, are used for the classifications of fraudulent behavior within telecommunication datasets. Extensive experiments on telecom data shows that this approach performs better than existing sequence embedding methods. In [58], Random forest, J48 and Naïve Bayes are applied to predict fake claims in automobile insurance and also simplifies the calculation of premium amounts based on previous

TABLE 1. Overview of existing actor-level fraud detection systems.

Frameworks and References	Data Mining approach	Type of Detected Fraud	Applied Data Mining Technique (s)
GSVMs [29]	hybrid	classifying insurance claims	Genetic support vector machines
Medical provider specialty predictions for anomaly detection [38]	Supervised	Physician related frauds	Multinomial Naïve Bayes
Fraud detection and frequent pattern matching [54]	Unsupervised	Disease based anomalies/frauds or period based claim related frauds	K means clustering
Fraud detection using outlier predictor in health insurance data [18]	Hybrid	Disease, medication related frauds	Discrimination rule based outlier analysis using clustering and graph theory
Healthcare fraud detection based on trustworthiness of doctors [24]	Hybrid	Provider (doctor) related fraud	Graph based mining Frequent mining algorithms
Fraudulent claims detection from expected payment deviations [41]	Supervised	Medicine payment related frauds	Regression models used
Predicting medical provider specialties to detect anomalous insurance claim [39]	Supervised	Fraudulent payments detected in dermatology and optometry	Bayesian inference, using probabilistic programming
Medical school training relate to practice evidence from big data [37]	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 [23]	Interactive machine learning	Prescription based abnormal behaviour	Pair wise comparison expectation maximization (EM)
Outlier-based Health Insurance Fraud Detection [36]	Unsupervised	Dental provider related frauds	Multi-dimensional data models Multivariate Clustering
A Survey: Healthcare fraud detection [30]	Hybrid	Rehabilitation, Septicaemia Pneumonia, payment related fraud detection	Geo-location Cluster analysis
Knowledge discovery from massive healthcare claims data [43]	Hybrid	Providers Related Frauds Social network	Social network analysis methods
Predicting healthcare fraud in Medicaid [35]	Hybrid	Patient related frauds Physician related frauds	Data models for patient claim and physicians.

financial details for different customers. The results show that random forest false claim prediction is more accurate as compared to other two classifiers. In [59], the fraud detection system is proposed which consists of two stages. In first stage, genetic algorithm based Fuzzy C-Means clustering is performed for identifying fraudulent cases. Then, identified anomalous samples are further verified by supervised learners namely Decision Tree (DT), Support Vector Machine (SVM), Group Method of Data Handling (GMDH) and Multi-Layer Perceptron (MLP). In [60], a similarity graph and page rank algorithm are applied for detecting provider level frauds. Similarity graph between the prescriptions of doctors of same specialty are created and then page rank algorithm is utilized to detect anomalies. The focus of these studies is to detect fraud mostly by performing financial analysis and only a few focused on actors connectivity details (telecommunication Industry). There is a need for research that focuses on actors associations details instead of financial analysis. It is observed that none of the system is able to detect all types of actor-level frauds.

Various researches have been conducted on health examination procedures. Such studies have attracted more attention as compared to studies that were conducted to predict patient behavior patterns. Data mining techniques and algorithms for predicting diseases and for analyzing medical records are presented in [61]–[63]. Sequential pattern mining is used in the medical field to identify frequent patterns or behaviors. In [64], Ohara *et al.* have utilized sequential pattern mining for prediction of serum anti-M*ü*llerian hormone (AMH) in women aged above 40. According to 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, whereas association rule mining generates association among features to predict future visits. Ou-Yang *et al.* in [65], have combined both these methods for predicting patients future visits. 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 is considered. Users' relationships are not possible as users can enter their claims on a single platform, and there is no possibility of interactions with other stakeholders (doctors, pharmacies, insurance companies, etc.). Camouflage behaviors are not easy to detect; using above mentioned approaches, temporal data mining can fulfill this purpose. Continuous time series data analysis are conducted for detecting frauds in [66]–[69]. Some recent studies are conducted to analyse discrete event-based sequential data, presented in [70]–[72]. The sequence of physician orders is used to perform temporal sequence analysis. In [73], Medicare database is used for fraud detection is proposed and train-test methodology performs perfectly in cross validation. The fraud detection in [74] is performed by adopting proactive strategies based on fourier transform and wavelet transform. The data is moved to the new domain where it can be analyzed easily and properly. In In [75], mobile healthcare claim frauds detection using SSIsomap activity clustering method is presented. Isomap is a method used to embed graph into Euclidean distance space. The algorithm named Semi-Supervised Isomap, is developed to further enhance the Isomap method. The behavioral pattern recognition is performed by using above mentioned algorithm. The results shows that this method is 50% more effective as compared to existing methods. Jurgovsky *et al.* in [76], have utilized the concept of sequence classification for detecting credit card fraud. The long and short term recurrent neural network is used to accumulate spending history of credit card holder. Later feature engineering is used for feature aggregation. The anomalous transactions are discriminated from genuine transactions by using the proposed methodology. The classification results are compared with random forest classifier results, and this methodology has proven to be more effective.

In [77], the unbalanced distribution of data is the major issue that decreases the performance of machine learning algorithms while detecting frauds is addressed. The three dimensional similarity metrics based approach namely transaction global similarity metric, feature local similarity metric and feature global similarity metric, is developed to handle the imbalance classification problem. In [78], 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.

It is observed that none of the research is considering patient sequences for each specialty, for detecting healthcare frauds. Most of the researches are performing payment based analysis.

V. SYSTEM ARCHITECTURE DESIGN

In this section we present the software design and architecture for the proposed healthcare fraud detection methodology. The architecture implements the idea of the designs. and presents the architecture in the form of two views: Use case view and logical view.

A. USE CASE VIEW

The use cases for the proposed solution are provided below:

- Login (analyst, doctor, etc.)
- Patient transaction
- Maintain patient information
- Maintain doctor information
- Select actor for analysis
- Monitoring ratings of actors
- View performance report
- Rule editor
- Close monitoring

These use cases are initiated by the analyst, doctors, or the investigators/management-related actors, and are depicted in Figure 3.

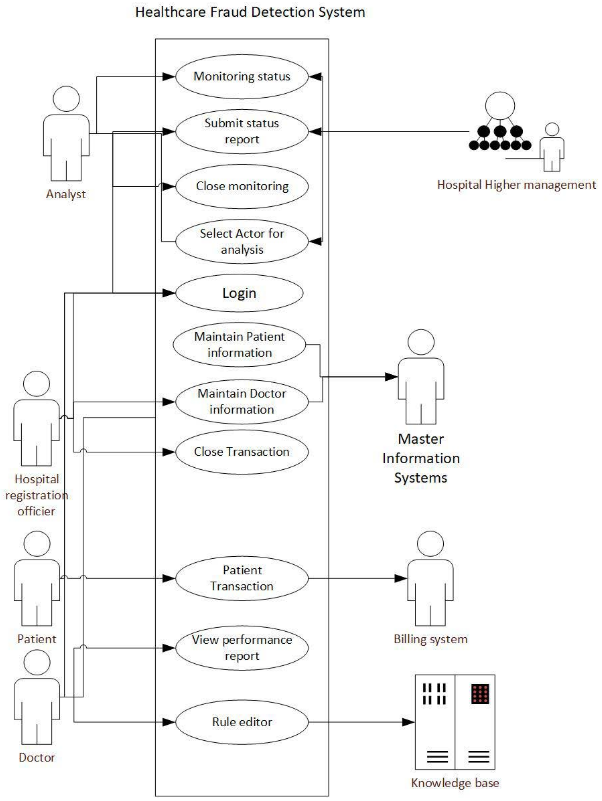
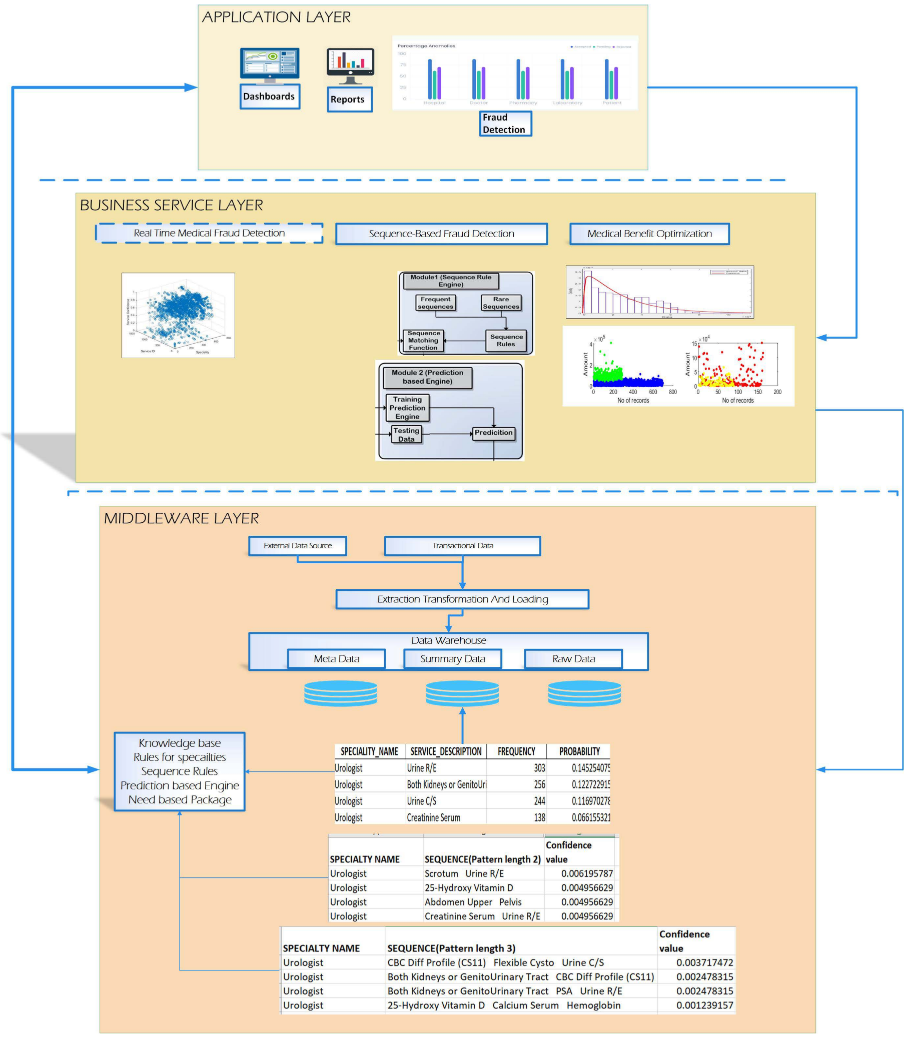

Healthcare Fraud Detection System
Monitoring status
Submit status report
Close monitoring
Select Actor for analysis
Login
Maintain Patient information
Maintain Doctor information
Close Transaction
Patient Transaction
View performance report
Rule editor
Analyst
Hospital Higher management
Hospital registration officier
Master Information Systems
Patient
Billing system
Doctor
Knowledge base

FIGURE 3. **Possible use cases of the designed architecture.**

- **Close monitoring**: This use case allows *analyst* to close the monitoring process. The main actor of this use case is the *analyst*. The *hospital management* is an actor involved within this use case.
- **Login**: This use case explains how a user logs into the healthcare benefit optimization system. The actors of this use case are *patients*, *doctors*, *registration officer* and *analyst*.
- **Maintain doctor information** : This use case allows the *registration officer* to maintain doctor's information in the hospital registration system. This includes updating or modifying doctors from the system.



SPECIALITY_NAME	SERVICE_DESCRIPTION	FREQUENCY	PROBABILITY
Urologist	Urine R/E	303	0.145254075
Urologist	Both Kidneys or GenitoUri	256	0.122722915
Urologist	Urine C/S	244	0.116970278
Urologist	Creatinine Serum	138	0.066155321

SPECIALTY NAME	SEQUENCE(Pattern length 2)	Confidence value
Urologist	Scrotum Urine R/E	0.006195787
Urologist	25-Hydroxy Vitamin D	0.004956629
Urologist	Abdomen Upper Pelvis	0.004956629
Urologist	Creatinine Serum Urine R/E	0.004956629

SPECIALTY NAME	SEQUENCE(Pattern length 3)	Confidence value
Urologist	CBC Diff Profile (CS11) Flexible Cysto Urine C/S	0.003717472
Urologist	Both Kidneys or GenitoUrinary Tract CBC Diff Profile (CS11)	0.002478315
Urologist	Both Kidneys or GenitoUrinary Tract PSA Urine R/E	0.002478315
Urologist	25-Hydroxy Vitamin D Calcium Serum Hemoglobin	0.001239157

FIGURE 4. Layering of subsystems in the proposed architecture.

The actors of this use case are *registration officer* and *hospital master information system*.

- **Select specialty for analysis**: This use case allows *analyst* to select the specialty from the MIS for the analysis. The actor starting this use case is the *analyst*. The *hospital Management system* and *MIS* are actors within the use case.
- **Monitoring ratings of actor**: This use case allows an *analyst* and *higher hospital management* to monitor transactions of actors. The *MIS* is an actor within the use case.
- **View performance report**: This use case allows a doctor to view his/her performance report for the previously completed transactions. The actor of this use case is doctor.
- **Submit report**: This use case allows an *analyst* to submit doctor, service and patient rating reports for the last few weeks. The actor in this use case is the *analyst*.
- **Rule editor**: This use case allows the *doctor* to add any new rule in his/her specialty. Knowledge base is also the actor within this use case.
- **Maintain patient information**: This use case allows the *registration officer* to maintain patient information in the hospital registration system. This includes adding or updating patients from the hospital master information system. The actor for this use case is the *registration officer*.

B. LOGICAL VIEW

1) SUBSYSTEM LAYERING

Layers provide a logical structure for the elements that are required to devise software solutions for the specified problem. In a layered architecture, components are organized into horizontal layers. Each layer performs a designated role. The three-layer architecture is used in our solution. Application layer, Business Service layer, and Middleware layer. Each

TABLE 2. Healthcare transaction detail.

Attribute name	Attribute type
Transaction ID	Float
MR_NO	nvarchar(255)
P_NAME	nvarchar(255)
GENDER	nvarchar(255)
DOB	datetime
AGE	float
EMP_ID	float
RELATION_	float
D_NAME	nvarchar(255)
MAJOR_ID	float
MAJOR_DESCRIPTION	nvarchar(255)
MINOR_DESCRIPTION	nvarchar(255)
SUBMINOR_DESCRIPTION	nvarchar(255)
SERVICE_ID	float
SERVICE_DESCRIPTION	nvarchar(255)
CATEGORY	nvarchar(255)
AMOUNT	float
DOCTOR	nvarchar(255)
SPECIALITY_NAME	nvarchar(255)
SERVICE_DATE	datetime

layer is only responsible for the tasks which are associated with their role. As shown in Figure 4, the middleware layer contains data storage for each element/ relational database of the healthcare ecosystem. Results from all these computations are stored in the knowledge base. Similarly, the Business service layer is responsible for the execution of rules, packages associated with the transaction. The algorithms are used in all methodologies and related computations are performed in this layer. It performs business logic against the data (e.g., computation using algorithms), and sends processed information to the Application layer. The layered logical structure is shown in Figure 4.

There are three main elements of the healthcare ecosystem namely patient, provider, services and each transaction in healthcare contains details about these three elements. The patient takes the service prescribed by the doctor. The main need is not only to detect the fraudulent transaction but to identify the main actor involved in performing this transaction. The proposed architecture facilitates the identification of fraudsters. Healthcare transactions involves following details as shown in Table 2.

In Table 2, patients perform transactions that are stored with a unique ID. The treatment/service taken by patient is described by *service_description, service_id, service_date* and *service_amount*. Service provider is the doctor whose speciality is also mentioned in the transaction details. The system context diagram is shown in Figure 5. When any transaction is performed, it is processed by the real time transaction processing component. The rules from the knowledge base are used for the evaluation of the transactions. Once the above evaluation is done, the status of the considered element is updated.

VI. PROPOSED METHODOLOGY

In the proposed design, knowledge base is created using the healthcare transactions data. We have generated sequences of the services. The frequent sequence algorithm *prefixspan* is applied to the sequence database. The pattern length 2, 3, 4, 5, and 6 is considered. Different values of minimum support are used according to the size of input records for each specialty. We have generated rare sequences with probability for each specialty. Based on all these results, we have generated a knowledge base for fraud detection. The knowledge base is composed of multiple files with services and service sequences of different pattern lengths for each specialty, along with their confidence values. These files are actually rules, and these rules are validated by the domain experts. For each specialty these files are generated using techniques, moreover, the number of files can be increased or decreased depending upon the considered transaction's sequence pattern length. The confidence values of each individual and the combination of services for the considered specialty are further processed to compute the confidence level for each rule. In each file, there is a different value of the threshold, and the threshold value decreases as pattern length increases. The weight is assigned to each type of rule file. The first file has a weight of 60, the second has a weight value of 30, the third has a weight value of 20, and the fourth file has a weight value of 10. The confidence values of rules are multiplied by the assigned weight. These weight values are configurable by the analyst/investigator. The confidence level for each rule is computed and it will be updated with time. The confidence value of each rule will get updated during real-time transaction processing. When any transaction is not found in the table then it is marked as dubious and the rating score of that element is updated, as shown in Figure 6. The rating score can be computed by using any other ranking algorithm can be used as per the application requirement.

A. REAL-TIME TRANSACTION PROCESSING

The transaction is analyzed against our knowledge base. When the transaction is taken place, the permutations of service sequences are computed for each transaction. There is a separate file for each type of pattern length. Each service is matched with the first table which contains individual services along with their confidence values. If the considered service is matched with the first table entries, then It will be evaluated against the second table. The second table contains two services patterns along with their confidence values within the specified specialty. If the pattern matches any of the entries of the second file, then the considered set of services are evaluated against the third file which contains sequences of three services along with their confidence values in the specified specialty. If the services matched with any of these entries, they will be evaluated against the fourth table which contains sequences with pattern length 4. If the sequence of services matches the sequences in our knowledge base, it will be considered a normal transaction otherwise will be considered dubious. All dubious transactions will be forwarded to the analyst. If the analyst finds the transaction normal, he can add this new sequence as a rule in the knowledge base. Rule editor is provided to both analyst and doctor. So that they can add the new rule to the knowledge base.

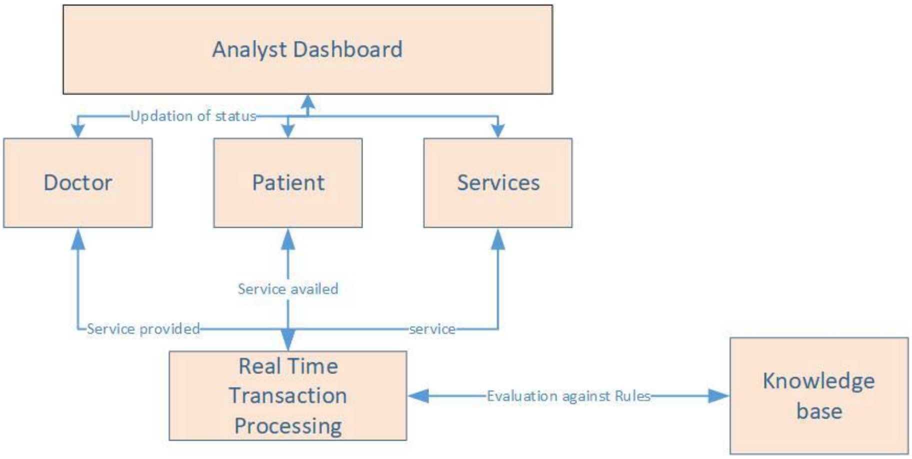


FIGURE 5. Generic representation of solution.

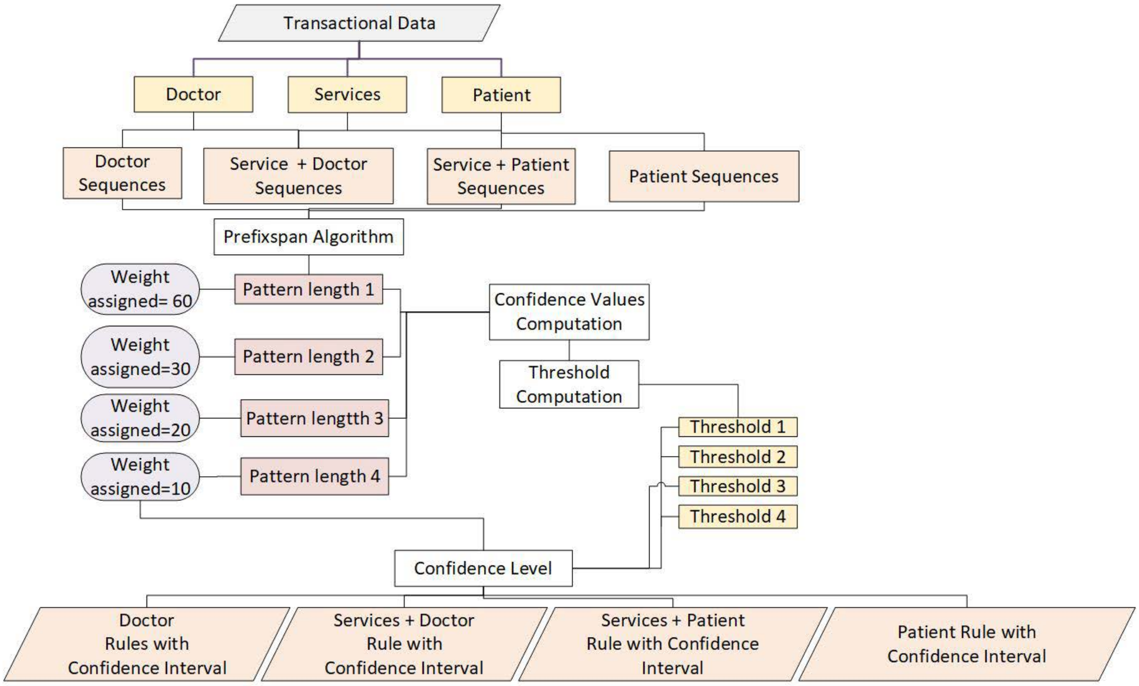


FIGURE 6. Proposed architecture for healthcare fraud detection solution.

VII. RESULTS AND ANALYSIS

This section presents the analysis to evaluate process based methodology for the extraction or detection of fraud cases. The transactional data is converted in to the database as shown in Figure 7.

The tables for each element of healthcare ecosystems are Patient, DOCTOR, SPECIALTY and SERVICE_TAKEN. Figure 7 depicts relations among these tables and Information about all transactions is contained in the FACT-TABLE. The rules, which are generated in the proposed framework, are stored in the DOCTOR_RULES table. The transactional data is transformed into sequence database. The relational diagram used for this methodology is depicted in Figure 8. Input sequences in the sequence mining and prediction methodology are stored in the input_sequence table. The results of prefixspan algorithms for frequent sequences are stored in the prefixspan_results table. The sequences' evaluation results with frequent sequences of a sequence rule engine are stored in prefixspan_status_result table. The results of rare sequences and their probabilities are stored in the rare_sequence table. The final results after detection of frauds are stored in the detected_sequence table.

The confidence value for each service in each specialty is computed as described in proposed methodology. Table 3 depicts confidence values for the sequences with pattern length 2. This section presents the analysis to evaluate the

TABLE 3. Sequences with pattern length 2.

Specialty name	Sequence with length 2	Frequency	Confidence value
Urologist	Creatinine Serum , Urine C/S	3	0.00375
Urologist	Abdomen + Pelvis Without Creatinine Serum	2	0.0025
Urologist	Both Kidneys or GenitoUrinary Tract Uroflow	2	0.0025
Urologist	Calcium Serum Uric Acid Serum	2	0.0025
Urologist	Abdomen + Pelvis Without Urine R/E	1	0.00125
Surgery - Plastic	Abdomen Upper Pelvis	3	0.017751
Surgery - Plastic	CBC Diff Profile (CS11) Chem 7	1	0.005917
Surgery - Plastic	Creatinine Serum Face With Contrast	1	0.005917
Surgery - Plastic	Creatinine Serum Hand without Contrast	1	0.005917
Surgery - Plastic	Elbow 3 views Hand 3 Views	1	0.005917
Surgery - Cardiac	Chest Xray 1 View Creatinine Serum	1	0.033333
Surgery - Cardiac	ECHO 2D & M Mode With Doppler Exercise Tolerance Tests	1	0.033333
Pediatrician	Blood C/S (Adult) Urine C/S	1	0.00063
Pediatrician	Blood C/S (Adult) Urine R/E	1	0.00063
Pediatrician	Blood C/S (Peads) CBC Diff Profile (CS11)	1	0.00063
Pediatrician	Brain/Head (3-D Imaging) Non Ionic Contrast Medium	1	0.00063
Pediatric Surgeon	Abdomen Upper Pelvis	1	0.015152
Pathologist	CBC Diff Profile (CS11) Urine R/E	2	0.035088
Pathologist	Helicobacter SGPT (ALT)	2	0.035088
Pathologist	25-Hydroxy Vitamin D TSH	1	0.017544
Orthopedic	Ankle 3 Views Foot 3 Views	6	0.004792
Orthopedic	Elbow 3 views Splint Long Arm	6	0.004792

TABLE 4. Sequences with pattern length 3.

Specialty name	Sequence with length 3	Frequency	Confidence value
Urologist	Both Kidneys or GenitoUrinary Tract Urine C/S Urine R/E	28	0.035
Urologist	Both Kidneys or GenitoUrinary Tract Creatinine Serum Urine R/E	5	0.00625
Urologist	Both Kidneys or GenitoUrinary Tract KUB Urine R/E	3	0.00375
Urologist	Both Kidneys or GenitoUrinary Tract CBC Diff Profile (CS11) Urine C/S	2	0.0025
Urologist	25-Hydroxy Vitamin D Calcium Serum Hemoglobin	1	0.00125
Urologist	25-Hydroxy Vitamin D Calcium Serum Phosphorous Serum	1	0.00125
Urologist	Abdomen + Pelvis With/Without Flexible Cysto Non Ionic Contrast Medium	1	0.00125
Urologist	C-Reactive Protein(CRP) High Sensitivity Urine C/S Urine R/E	1	0.00125
Urologist	FSH Scrotum Testosterone	1	0.00125
Surgery - Cardiac	25-Hydroxy Vitamin D Chem 7 Vitamin B 12	1	0.033333
Surgery - Cardiac	CBC Diff Profile (CS11) ESR SGPT (ALT)	1	0.033333
Orthopedic	25-Hydroxy Vitamin D Chem 7 LFT	1	0.000799
Orthopedic	25-Hydroxy Vitamin D Neck, 2 views Shoulder 3 views	1	0.000799
Orthopedic	25-Hydroxy Vitamin D Oscalcis Heel - AP/Lateral	1	0.000799
Orthopedic	Ankle 3 Views Ankle 3 Views Splint Sugar Tong	1	0.000799
Orthopedic	Fasting Chem 7 Glucose Fasting LIPID Profile(HDL,LDL,Chlstrl,Trig)	1	0.000799
Opthalmologist	A Scan Biometry for Cataract Surgery B Scan Ultrasonogrphy	1	0.009091
Opthalmologist	Brain without Contrast Humphrey Auto Mated Fields Humphrey Auto Mated Fields	1	0.009091
Opthalmologist	Creatinine Serum ESR HBA1C (HS16) TSH	1	0.009091
Opthalmologist	Fundus Flourescine Angiogram Intravetrial Injection "Avstrin" Laser Category D Laser Category D	1	0.009091
Oncologist	Chest,Abdomen and Pelvis Creatinine Serum Non Ionic Contrast Dye Charges	9	0.039474
Oncologist	CBC Diff Profile (CS11) Chemotherapy "5FU"1hour "w/o phar/supp" Consumable Supplies (OPD)	2	0.008772

TABLE 5. Sequences with pattern length 4.

Specialty name	Sequence with length 3	Frequency	Confidence value
Urologist	Both Kidneys or GenitoUrinary Tract Creatinine Serum Glucose Random Urine C/S Urine R/E	3	0.00375
Urologist	Both Kidneys or GenitoUrinary Tract CBC Diff Profile (CS11) Creatinine Serum Urine R/E	2	0.0025
Urologist	Both Kidneys or GenitoUrinary Tract Creatinine Serum Urine C/S Urine R/E	2	0.0025
Urologist	Calcium Serum Creatinine Serum Uric Acid Serum Urine C/S	2	0.0025
Urologist	24 Hours Urinary Oxalate Calcium Urine Creatinine Urine U. Citric Acid Uric Acid Urine	1	0.00125
Urologist	Abdomen + Pelvis With/Without Both Kidneys or GenitoUrinary Tract Non Ionic Contrast Medium Urine R/E	1	0.00125
Surgery - Plastic	C-Reactive Protein(CRP) High Sensitivity CBC Diff Profile (CS11) ESR Tissue C/S	1	0.005917
Surgery - Plastic	CBC Diff Profile (CS11) Chest Xray 1 View ESR Thoraco Lumber Spine 2view	1	0.005917
Surgery - Cardiac	CBC Diff Profile (CS11) Chem 7 Chest Xray 1 View ECG 12 Lead	2	0.066667
Surgery - Cardiac	Chest (3-D Imaging) Creatinine Serum ECG 12 Lead Non Ionic Contrast Medium	1	0.033333
Surgery - Cardiac	Chest Xray 1 View Creatinine Serum Potassium Serum Sodium Serum	1	0.033333
Pediatrician	Blood C/S (Peads) CBC Diff Profile (CS11) LFT Malarial Parasite Rapid Malaria Test By ICT Method	4	0.002519
Pediatric Cardiologist	24-Hour Ambulatory B.P. Monitor BUN Creatinine Serum LIPID Profile(HDL,LDL,Chlstrl,Trig) Uric Acid Serum	1	0.026316
Orthopedic	Chest Xray 1 View Scaphoid 4 View Splint Short Arm Wrist 3 Views	1	0.000799
Orthopedic	Ankle 3 Views Ankle 3 Views Splint Sugar Tong	1	0.000799
Orthopedic	Fasting Chem 7 Glucose Fasting LIPID Profile(HDL,LDL,Chlstrl,Trig)	1	0.000799
Opthalmologist	A Scan Biometry for Cataract Surgery B Scan Ultrasonogrphy	1	0.009091
Opthalmologist	Brain without Contrast Humphrey Auto Mated Fields Humphrey Auto Mated Fields	1	0.009091
Opthalmologist	Creatinine Serum ESR HBA1C (HS16) TSH	1	0.009091
Opthalmologist	Fundus Flourescine Angiogram Intravetrial Injection "Avstrin" Laser Category D Laser Category D	1	0.009091
Oncologist	Chest,Abdomen and Pelvis Creatinine Serum Non Ionic Contrast Dye Charges	9	0.039474
Oncologist	CBC Diff Profile (CS11) Chemotherapy "5FU"1hour "w/o phar/supp" Consumable Supplies (OPD)	2	0.008772

proposed methodology for the extraction or detection of fraud cases. The confidence value for each service in each specialty is computed as described in proposed methodology. Table 3 depicts confidence values for the sequences with pattern length 2. Table 4 depicts sequences with pattern length 3 along with their confidence values. The sequences with pattern length 4 are depicted in Table 5. The confidence level of rules is computed by using the following equation.

$$Confidence\ level = weight * confidence\ value \tag{1}$$

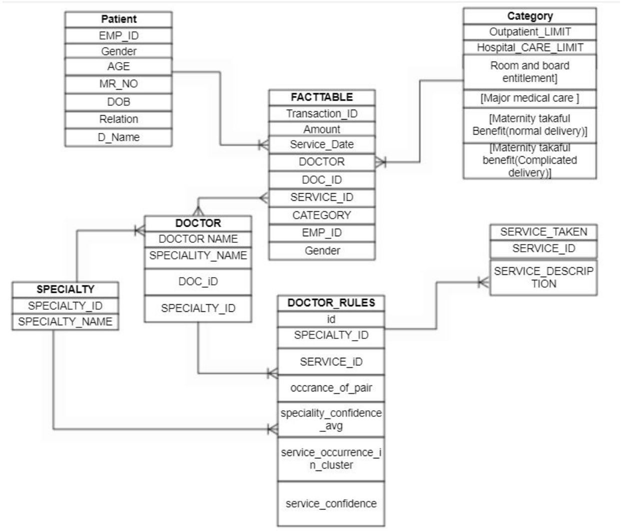


FIGURE 7. Relational database for medical fraud detection framework.

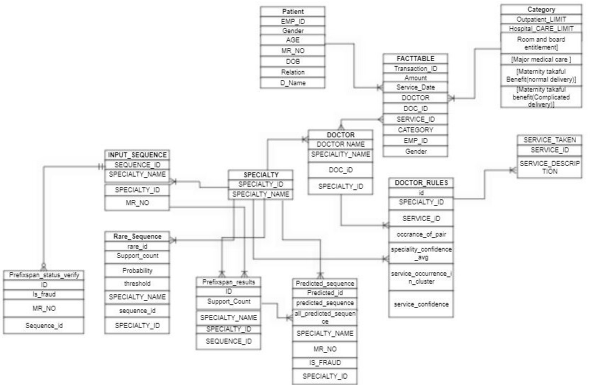


FIGURE 8. Relational diagram for sequence database.

TABLE 6. Confidence level for pattern length 2.

Confidence value	Confidence level
0.00375	0.11
0.0025	0.08
0.0025	0.08
0.0025	0.08
0.00125	0.04
0.017751	0.53
0.005917	0.18
0.005917	0.18
0.005917	0.18
0.005917	0.18
0.033333	1.00
0.033333	1.00
0.00063	0.02
0.00063	0.02
0.00063	0.02
0.00063	0.02
0.015152	0.45
0.035088	1.05
0.035088	1.05
0.017544	0.53
0.004792	0.14
0.004792	0.14

TABLE 7. Confidence level for pattern length 3.

Confidence value	Confidence level
0.035	0.70
0.00625	1.25
0.00375	0.075
0.0025	0.05
0.00125	0.025
0.00125	0.025
0.00125	0.025
0.00125	0.025
0.033333	0.66
0.033333	0.66
0.000799	0.01598
0.000799	0.01598
0.000799	0.01598
0.000799	0.01598
0.000799	0.01598
0.000799	0.01598
0.009091	0.18182
0.009091	0.18182
0.009091	0.18182
0.009091	0.18182
0.039474	0.78948
0.008772	0.17544

TABLE 8. Confidence level for pattern length 4.

Confidence value	Confidence level
0.00375	0.0375
0.0025	0.025
0.0025	0.025
0.0025	0.025
0.00125	0.0125
0.00125	1.025
0.005917	0.05917
0.005917	0.05917
0.066667	0.66667
0.033333	0.33333
0.033333	0.33333
0.002519	0.02519
0.026316	0.26316
0.000799	0.00799
0.000799	0.00799
0.000799	0.00799
0.009091	0.9091
0.009091	0.9091
0.009091	0.9091
0.009091	0.9091
0.039474	0.39474
0.008772	0.08772

TABLE 9. Services with confidence values in ophthalmologist specialty.

Specailty Name	Service	Confidence value
Opthalmologist	Galilei Scan	0.172222222
Opthalmologist	OCT Scan	0.094444444
Opthalmologist	Humphrey Auto Mated Fields	0.077777778
Opthalmologist	C-Reactive Protein(CRP) High Sensitivity	0.05
Opthalmologist	ESR	0.038888889
Opthalmologist	CBC Diff Profile (CS11)	0.033333333
Opthalmologist	Chalazion I & C	0.033333333
Opthalmologist	Intravetrial Injection "Avstrin"	0.027777778
Opthalmologist	25-Hydroxy Vitamin D	0.022222222
Opthalmologist	A Scan Biometry for Cataract Surgery	0.022222222
Opthalmologist	Brain with Contrast	0.016666667
Opthalmologist	CBC Profile (CS10)	0.016666667
Opthalmologist	Creatinine Serum	0.016666667
Opthalmologist	HBA1C (HS16)	0.016666667
Opthalmologist	LFT	0.016666667
Opthalmologist	APTT	0.011111111

Further, the subset of sequences with pattern length 2 are shown in Table 6. The confidence level for sequences with pattern lengths 3 and 4 are shown in Table 7 and Table 5 respectively.

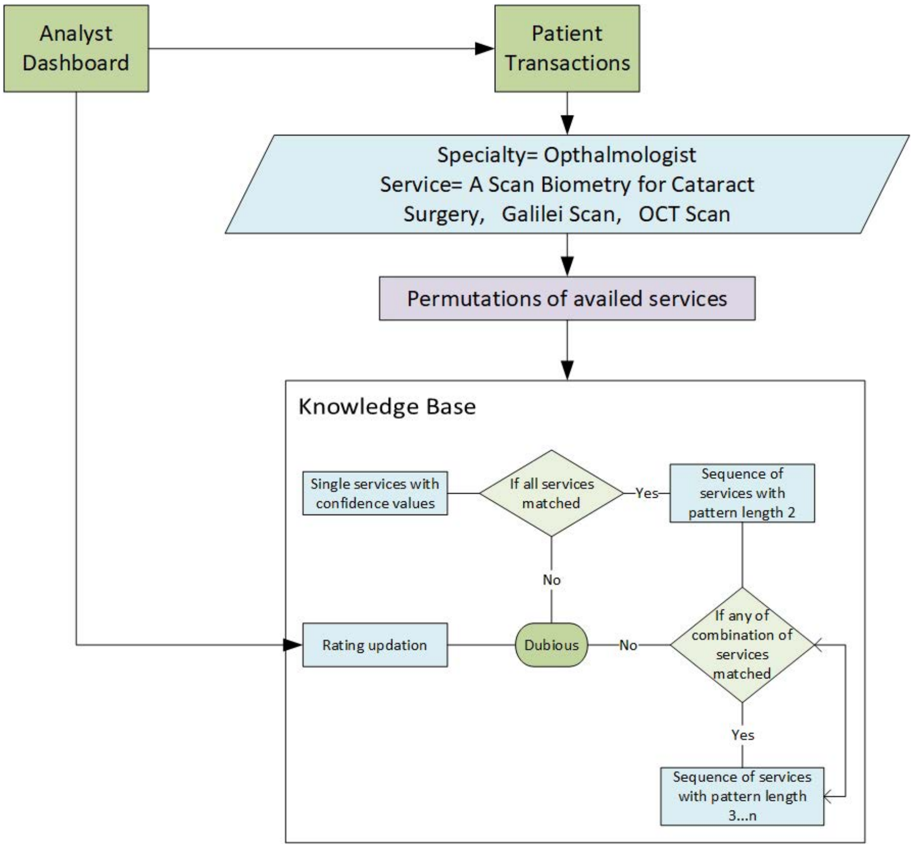

Analyst Dashboard
Patient Transactions
Specialty= Opthalmologist
Service= A Scan Biometry for Cataract Surgery, Galilei Scan, OCT Scan
Permutations of availed services
Knowledge Base
Single services with confidence values
If all services matched
Yes
Sequence of services with pattern length 2
No
Rating updation
Dubious
No
If any of combination of services matched
Yes
Sequence of services with pattern length 3…n

FIGURE 9. Real-time transaction processing for ophthalmologist specialty.

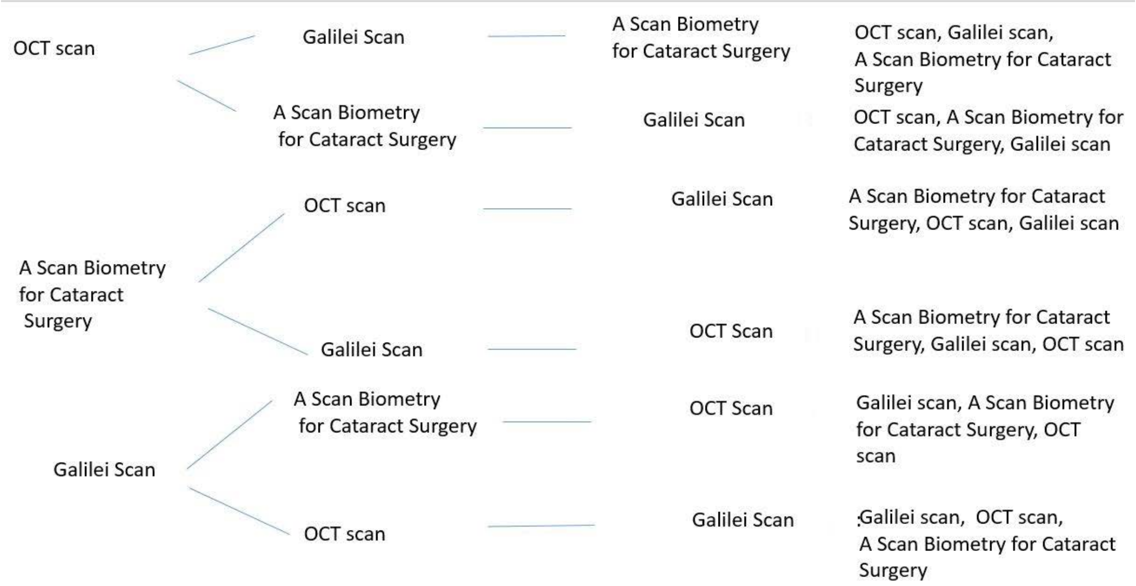

OCT scan
Galilei Scan
A Scan Biometry for Cataract Surgery
OCT scan, Galilei scan, A Scan Biometry for Cataract Surgery
A Scan Biometry for Cataract Surgery
Galilei Scan
OCT scan, A Scan Biometry for Cataract Surgery, Galilei scan
A Scan Biometry for Cataract Surgery
OCT scan
Galilei Scan
A Scan Biometry for Cataract Surgery, OCT scan, Galilei scan
Galilei Scan
OCT Scan
A Scan Biometry for Cataract Surgery, Galilei scan, OCT scan
Galilei Scan
A Scan Biometry for Cataract Surgery
OCT Scan
Galilei scan, A Scan Biometry for Cataract Surgery, OCT scan
OCT scan
Galilei Scan
;Galilei scan, OCT scan, A Scan Biometry for Cataract Surgery

FIGURE 10. Permutation of availed or provided services.

After a complete description of the knowledge base design, in the next subsection, we explain how the transaction will be processed in real-time.

1) CASE STUDY

The transaction performed in the Ophthalmologist specialty is considered in this section, to provide a clear understanding

TABLE 10. Sequences of services with pattern length 2 in ophthalmologist specialty.

Specialty name	Sequaence with pattern length 2	Confidence value
Opthalmologist	Humphrey Auto Mated Fields OCT Scan	0.018018018
Opthalmologist	25-Hydroxy Vitamin D CBC Diff Profile (CS11)	0.009009009
Opthalmologist	25-Hydroxy Vitamin D Calcium Serum	0.009009009
Opthalmologist	25-Hydroxy Vitamin D Hemoglobin	0.009009009
Opthalmologist	A Scan Biometry for Cataract Surgery B Scan Ultrasonogrphy	0.009009009
Opthalmologist	A Scan Biometry for Cataract Surgery Galilei Scan	0.009009009
Opthalmologist	Brain with Contrast Neck Soft Tissue with Contrast	0.009009009
Opthalmologist	CBC Profile (CS10) Chem 7	0.009009009
Opthalmologist	Fundus Flourescine Angiogram OCT Scan	0.009009009
Opthalmologist	Galilei Scan Galilei Scan	0.009009009

TABLE 11. Sequences of services with pattern length 3 in ophthalmologist specialty.

Specialty Name	Sequence with Pattern length 3	Confidence value
Opthalmologist	A Scan Biometry for Cataract Surgery Galilei Scan OCT Scan	0.009009009
Opthalmologist	Creatinine Serum ESR HBA1C (HS16) TSH	0.009009009
Opthalmologist	Creatinine Serum ESR HBA1C (HS16) TSH	0.009009009
Opthalmologist	C-Reactive Protein(CRP) High Sensitivity ESR Thyroid Profile (TT3, FT4, TSH)	0.009009009
Opthalmologist	Brain with Contrast Creatinine Serum Orbit with Contrast	0.009009009

of their proposed architecture. The real-time processing of transactions considered in this case is depicted in Figure 9.

As it is already mentioned that, permutations of services availed in transaction are computed. The services availed in considered case are Scan Biometry for Cataract Surgery, Galilei Scan and OCT scan. The permutations for the sequence mentioned are computed as shown in Figure 10.

The generated permutations are evaluated first against single service table as shown in Table 9 From Table 10, (A Scan Biometry for Cataract Surgery, Galilei Scan) sequence is matched with one of the permutation. Whenever any sequence is matched with the sequences of any table, the confidence value of that sequence will be increased. It can be seen from Table 9, A Scan Biometry for Cataract Surgery, Galilei Scan and OCT scan are present in the first table. The permutation of two services are evaluated against Table 10.

Now, the sequence with three services is matched with the third Table 11 which contains all the sequences with pattern length 3. It can be seen from the Table 11, sequence {A Scan Biometry for Cataract Surgery Galilei Scan, OCT Scan} is matched with one of the permutations which are generated earlier. The confidence value of this sequence will be updated in the knowledge base. In this case, we have considered only sequences with pattern length three because the sequence followed in the transaction is of pattern length 3. In this way, all transactions will be processed and evaluated against the knowledge base. Knowledge base rules confidence levels will be updated whenever any new transaction matches with the rule. With the help of this architecture, healthcare frauds can be detected and handled efficiently.

VIII. CONCLUSION

Healthcare fraud involves making false statement, claims, documentation or medical conditions to attain an illegal benefit. Healthcare frauds are wide spread and growing and have become a complex challenge that involve multiple entities and factors. Therefore, it is essential to provide solutions for controlling healthcare frauds. The absence of an effective fraud detection framework hinders the provision of healthcare services to genuine patients. This paper is an effort to address the following problems:

1) Our basic concern is the overburdened healthcare systems. The systems may choke and get overburdened when unnecessary treatments are prescribed by the doctors or practitioners. Such non-essential treatments may lead to wastage of healthcare resources and cause their shortage. As a result, genuine patients are deprived of the required healthcare services or facilities.
2) Consequences of medical frauds may range from financial frauds to other medical frauds, including identity theft. These days complete data of the patients' personal information is being obtained and captured in the systems that is used by various hospitals, clinics, and other health service providers. Consequently, confidentiality of electronic healthcare records is of utmost importance as these records may contain information like national identity card numbers of the patients and other confidential information that could provide a platform to the fraudsters for planting frauds. Usually, fraudsters may use victims' identities for availing medical services where there are similarities in fraudsters' medical-related information like blood type, weight, height, etc.
3) Over diagnosis of diseases may also cause physical harm to the patients, and in some cases, this situation may become life-threatening. For example, a patient may deliberately be diagnosed with cancer to gain attention and health care services more easily. However, it can cause physical harm to the patients.
4) Employer and enterprises are overburdened due to increased premium amounts.

Existing research provides significant work on detecting specific types of fraud, but are failed to provide us a standard approach for detecting all types of healthcare frauds.

The proposed methodology relies on the novel idea of analyzing patient sequences for detecting fraud at each speciality level and for fraud detection in clinical service processes. We have used prefix span sequence mining approach and Bayes rule for populating frequent and rare sequences in sequence rule engine based on sequence database of patient time series and particular patient traces for a specific speciality. Analysis of medical behaviors in clinical processes has led to the identification of anomalous sequences as patients deviate from sequences contained in sequence rule engine. In other words, we can detect sequences that deviate from frequent medical behaviors. Once anomalous sequences are identified, they are further analyzed to detect fraudulent cases.

In addition, various meetings have been arranged with medical domain experts to upgrade and evaluate 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 the proposed methodology is efficient and capable of identifying fraudulent cases. constructed detection model.

IX. LIMITATION AND FUTURE WORK

The dataset we used to validate the proposed methodology was difficult to obtain as it contains private and confidential information of patient's data. The dataset was in raw form and to handle the missing and redundant information was time-consuming. We used five years transactional dataset, but to check the effectiveness of the framework and methodology, larger datasets would be more useful and will show better visualization as well as strength of proposed work in larger perspective.

With the increase in data size, cloud computing will be required for processing. We used age, gender, marital status, relation, and number of visits during the design of fault detection process. The proposed methodology can be extended by adding more features or specialty that can contribute to the evaluation of frauds. It would reflect better and improved performance if applied for a wider perspective. In future, the proposed fraud detection methodology can be further improved with the design of computerized physician order entry system for each disease using machine learning techniques. Upon finding a set of sequences for every disease, provider level fraud detection would be more effective.

REFERENCES

[1] W. A. Ameerica, "Indiana statewide survey of registered voters," We ask America, Washington, DC, USA, Tech. Rep. Alabama 2nd Congressional District Republican Primary Survey, Jan. 2020.

[2] "Global healthcare fraud analytics and detection market forecast 2020–2028," Inkwood Research, New York, NY, USA, Tech. Rep., 2020. [Online]. Available: https://www.Reportlinker.com

[3] K. Amer, "The healthcare juggernaut," Dawn news, Karachi, Pakistan, Tech. Rep., 2018.

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

[5] 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. 2001.

[6] C. Antunes and A. Oliveira, "Sequential pattern mining algorithms: Trade-offs between speed and memory," Dept. Inf. Syst. Comput. Sci., Av. Rovisco Pais 1, Lisboa, Portugal, Tech. Rep. 1049-001, 2004.

[7] 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 (KDD)*, 2000, pp. 355–359.

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

[9] 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, 2006.

[10] 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.

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

[12] Z. Huang, X. Lu, and H. Duan, "Anomaly detection in clinical processes," in *Proc. AMIA Annu. Symp.*, 2012, p. 370.

[13] R. Bauder and T. Khoshgoftaar, "A survey of medicare data processing and integration for fraud detection," in *Proc. IEEE Int. Conf. Inf. Reuse Integr. (IRI)*, Jul. 2018, pp. 9–14.

[14] C. Zhang, X. Xiao, and C. Wu, "Medical fraud and abuse detection system based on machine learning," *Int. J. Environ. Res. Public Health*, vol. 17, no. 19, p. 7265, Oct. 2020.

[15] 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. (IC)*, Aug. 2017, pp. 1–7.

[16] A. Okita, M. Yamashita, K. Abe, C. Nagai, A. Matsumoto, M. Akehi, R. Yamashita, N. Ishida, M. Seike, S. Yokota, N. Umekawa, Y. Matsumoto, Y. Kishimoto, A. Okazaki, E. Komori, S. Sawada, and S. Takashima, "Variance analysis of a clinical pathway of video-assisted single lobectomy for lung cancer," *Surg. Today*, vol. 39, no. 2, pp. 104–109, Feb. 2009.

[17] J. van de Klundert, P. Gorissen, and S. Zeemering, "Measuring clinical pathway adherence," *J. Biomed. Informat.*, vol. 43, no. 6, pp. 861–872, Dec. 2010.

[18] J. Peng, Q. Li, H. Li, L. Liu, Z. Yan, and S. Zhang, "Fraud detection of medical insurance employing outlier analysis," in *Proc. IEEE 22nd Int. Conf. Comput. Supported Cooperat. Work Design (CSCWD)*, May 2018, pp. 341–346.

[19] M. S. Anbarasi and S. Dhivya, "Fraud detection using outlier predictor in health insurance data," in *Proc. Int. Conf. Inf. Commun. Embedded Syst. (ICICES)*, Feb. 2017, pp. 1–6.

[20] J. O. Savino and B. E. Turvey, "Medicaid/medicare fraud," in *False Allegations*. Amsterdam, The Netherlands: Elsevier, 2018, pp. 89–108.

[21] T. Ekin, G. Lakomski, and R. M. Musal, "An unsupervised Bayesian hierarchical method for medical fraud assessment," *Stat. Anal. Data Mining, ASA Data Sci. J.*, vol. 12, no. 2, pp. 116–124, Apr. 2019.

[22] B. Zafari and T. Ekin, "Topic modelling for medical prescription fraud and abuse detection," *J. Roy. Stat. Soc., C, Appl. Statist.*, vol. 68, no. 3, pp. 751–769, Apr. 2019.

[23] I. Kose, M. Gokturk, and K. Kilic, "An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance," *Appl. Soft Comput.*, vol. 36, pp. 283–299, Nov. 2015.

[24] H. Cui, Q. Li, H. Li, and Z. Yan, "Healthcare fraud detection based on trustworthiness of doctors," in *Proc. IEEE Trustcom/BigDataSE/ISPA*, Aug. 2016, pp. 74–81.

[25] B. Itri, Y. Mohamed, Q. Mohammed, and B. Omar, "Performance comparative study of machine learning algorithms for automobile insurance fraud detection," in *Proc. 3rd Int. Conf. Intell. Comput. Data Sci. (ICDS)*, Oct. 2019, pp. 1–4.

[26] J. Li, K.-Y. Huang, J. Jin, and J. Shi, "A survey on statistical methods for health care fraud detection," *Health Care Manage. Sci.*, vol. 11, no. 3, pp. 275–287, Sep. 2008.

[27] H. Joudaki, A. Rashidian, B. Minaei-Bidgoli, M. Mahmoodi, B. Geraili, M. Nasiri, and M. Arab, "Using data mining to detect health care fraud and abuse: A review of literature," *Global J. Health Sci.*, vol. 7, no. 1, p. 194, Aug. 2014.

[28] P. A. Ortega, C. J. Figueroa, and G. A. Ruz, "A medical claim fraud/abuse detection system based on data mining: A case study in Chile," in *Proc. DMIN*, vol. 6, 2006, pp. 26–29.

[29] R. A. Sowah, M. Kuuboore, A. Ofoli, S. Kwofie, L. Asiedu, K. M. Koumadi, and K. O. Apeadu, "Decision support system (DSS) for fraud detection in health insurance claims using genetic support vector machines (GSVMs)," *J. Eng.*, vol. 2019, pp. 1–19, Sep. 2019.

[30] Q. Liu and M. Vasarhelyi, "Healthcare fraud detection: A survey and a clustering model incorporating geo-location information," in *Proc. 29th World Continuous Auditing Reporting Symp. (WCARS)*, Brisbane, QLD, Australia, 2013, pp. 1–10.

[31] J. M. Johnson and T. M. Khoshgoftaar, "Medical provider embeddings for healthcare fraud detection," *Social Netw. Comput. Sci.*, vol. 2, no. 4, pp. 1–15, Jul. 2021.

[32] L. Settipalli and G. R. Gangadharan, "Healthcare fraud detection using primitive sub peer group analysis," *Concurrency Comput., Pract. Exper.*, vol. 33, no. 23, p. e6275, Dec. 2021.

[33] M. C. Massi, F. Ieva, and E. Lettieri, "Data mining application to healthcare fraud detection: A two-step unsupervised clustering method for outlier detection with administrative databases," *BMC Med. Informat. Decis. Making*, vol. 20, no. 1, pp. 1–11, Dec. 2020.

[34] 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, 2006.

[35] D. Thornton, G. van Capelleveen, M. Poel, J. van Hillegersberg, and R. M. Mueller, "Outlier-based health insurance fraud detection for us medicaid data," in *Proc. ICEIS*, 2014, pp. 684–694.

[36] D. Thornton, R. M. Mueller, P. Schoutsen, and J. van Hillegersberg, "Predicting healthcare fraud in medicaid: A multidimensional data model and analysis techniques for fraud detection," *Proc. Technol.*, vol. 9, pp. 1252–1264, Jan. 2013.

[37] K. Feldman and N. V. Chawla, "Does medical school training relate to practice? Evidence from big data," *Big Data*, vol. 3, no. 2, pp. 103–113, Jun. 2015.

[38] M. Herland, R. A. Bauder, and T. M. Khoshgoftaar, "Medical provider specialty predictions for the detection of anomalous medicare insurance claims," in *Proc. IEEE Int. Conf. Inf. Reuse Integr. (IRI)*, Aug. 2017, pp. 579–588.

[39] R. A. Bauder, T. M. Khoshgoftaar, A. Richter, and M. Herland, "Predicting medical provider specialties to detect anomalous insurance claims," in *Proc. IEEE 28th Int. Conf. Tools With Artif. Intell. (ICTAI)*, Nov. 2016, pp. 784–790.

[40] R. A. Bauder and T. M. Khoshgoftaar, "A probabilistic programming approach for outlier detection in healthcare claims," in *Proc. 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2016, pp. 347–354.

[41] R. A. Bauder and T. M. Khoshgoftaar, "A novel method for fraudulent medicare claims detection from expected payment deviations (application paper)," in *Proc. IEEE 17th Int. Conf. Inf. Reuse Integr. (IRI)*, Jul. 2016, pp. 11–19.

[42] R. A. Bauder and T. M. Khoshgoftaar, "The detection of medicare fraud using machine learning methods with excluded provider labels," in *Proc. 31st Int. Flairs Conf.*, pp. 404–409, 2018.

[43] V. Chandola, S. R. Sukumar, and J. C. Schryver, "Knowledge discovery from massive healthcare claims data," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2013, pp. 1312–1320.

[44] I. Gath and A. B. Geva, "Unsupervised optimal fuzzy clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 773–780, Jul. 1989.

[45] M. J. Lenard and P. Alam, "Application of fuzzy logic to fraud detection," in *Encyclopedia of Information Science and Technology*, 1st ed. Hershey, PA, USA: IGI Global, 2005, pp. 135–139.

[46] M. Köppen, N. Kasabov, and G. Coghill, "Advances in neuro-information processing," in *Proc. 15th Int. Conf. Adv. Neuro-Inf. Process. (ICONIP)*, Auckland, New Zealand, vol. 5507. Berlin, Germany: Springer Verlag, 2009.

[47] C. Sun, Z. Yan, Q. Li, Y. Zheng, X. Lu, and L. Cui, "Abnormal group-based joint medical fraud detection," *IEEE Access*, vol. 7, pp. 13589–13596, 2019.

[48] V. Hristidis, *Information Discovery on Electronic Health Records*. Boca Raton, FL, USA: CRC Press, 2009.

[49] W. Altaf, M. Shahbaz, and A. Guergachi, "Applications of association rule mining in health informatics: A survey," *Artif. Intell. Rev.*, vol. 47, no. 3, pp. 313–340, 2017.

[50] G. Toti, R. Vilalta, P. Lindner, B. Lefer, C. Macias, and D. Price, "Analysis of correlation between pediatric asthma exacerbation and exposure to pollutant mixtures with association rule mining," *Artif. Intell. Med.*, vol. 74, pp. 44–52, Nov. 2016.

[51] R. Cai, M. Liu, Y. Hu, B. L. Melton, M. E. Matheny, H. Xu, L. Duan, and L. R. Waitman, "Identification of adverse drug-drug interactions through causal association rule discovery from spontaneous adverse event reports," *Artif. Intell. Med.*, vol. 76, pp. 7–15, Feb. 2017.

[52] L. Zeng, B. Wang, L. Fan, and J. Wu, "Analyzing sustainability of Chinese mining cities using an association rule mining approach," *Resour. Policy*, vol. 49, pp. 394–404, Sep. 2016.

[53] 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, Oct. 2013.

[54] 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. (IC)*, Aug. 2017, pp. 1–7.

[55] P. Travaille, R. M. Müller, D. Thornton, and J. Van Hillegersberg, "Electronic fraud detection in the us medicaid healthcare program: Lessons learned from other industries," in *Proc. AMCIS*, 2011, pp. 1–11.

[56] S. Carta, G. Fenu, D. R. 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.

[57] G. Liu, J. Guo, Y. Zuo, J. Wu, and R.-Y. Guo, "Fraud detection via behavioral sequence embedding," *Knowl. Inf. Syst.*, vol. 62, no. 7, pp. 1–24, 2020.

[58] G. Kowshalya and M. Nandhini, "Predicting fraudulent claims in automobile insurance," in *Proc. 2nd Int. Conf. Inventive Commun. Comput. Technol. (ICICCT)*, Apr. 2018, pp. 1338–1343.

[59] S. Subudhi and S. Panigrahi, "Use of optimized fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 32, no. 5, pp. 568–575, Jun. 2020.

[60] 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.

[61] 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.

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

[63] 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, 2015.

[64] 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.

[65] 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.

[66] 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.

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

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] M. Herland, R. A. Bauder, and T. M. Khoshgoftaar, "The effects of class rarity on the evaluation of supervised healthcare fraud detection models," *J. Big Data*, vol. 6, no. 1, pp. 1–33, Dec. 2019.

[74] R. Saia and S. Carta, "Evaluating the benefits of using proactive transformed-domain-based techniques in fraud detection tasks," *Future Gener. Comput. Syst.*, vol. 93, pp. 18–32, Apr. 2019.

[75] Y. Gao, C. Sun, R. Li, Q. Li, L. Cui, and B. Gong, "An efficient fraud identification method combining manifold learning and outliers detection in mobile healthcare services," *IEEE Access*, vol. 6, pp. 60059–60068, 2018.

[76] 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.

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

[78] 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.

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 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, systems 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 was 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 the USA and Pakistan. 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, and Malaysia. He has served as a member for 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 has also served as the Chairman for the Pakistan Association of Software Houses (PASHA) and as a member for the Board of Governance of many entities in the Ministry of IT and Commerce.

RUKAIYA RUKAIYA is currently pursuing the Ph.D. degree with the Department of Computer and Software Engineering, College of Electrical and Mechanical Engineering, National University of Sciences and Technology, Pakistan. Her research interests include MAC and routing protocols, wireless *ad-hoc* networks, mission critical tactical networks, and networks security. Her awards and honors include a gold medal in the B.S. degree and an indigenous scholarship from the Higher Education Commission (HEC), Pakistan, for her Ph.D. studies.

MUAZZAM A. KHAN KHATTAK (Senior Member, IEEE) is working as Tenured Professor and Director ICESCO Chair for Data Analytics and Edge Computing, Quaid-i-Azam University, Islamabad Pakistan. He received the Ph.D. degree and Postdoc from IIUI and University of Missouri, Kansas City, MO, USA, in 2011 and 2016, respectively. He joined the National University of Sciences & Technology (NUST), Islamabad in 2013, and was promoted to Associate Dean in 2017. He has been at the School of Computer Science, University of Ulm, Germany, and at the Networking and Multimedia Lab, School of Computer and the Electrical Engineering, University of Missouri (UMKC), USA, as a Research Fellow. He is also a member of the Pakistan Academy of Sciences. His research interests include the Internet of Things, next generation intelligent networks, blockchain, information and network security, vehicular ad-hoc Networks and acoustic networks. He has published more than 160 publications and book chapters.

ARSLAN MUNIR (Senior Member, IEEE) received the M.A.Sc. degree in ECE from The University of British Columbia (UBC), Vancouver, Canada, in 2007, and the Ph.D. degree in ECE from the University of Florida (UF), Gainesville, FL, USA, in 2012. He was a Postdoctoral Research Associate with the Electrical and Computer Engineering (ECE) Department, Rice University, Houston, TX, USA, from May 2012 to June 2014. He is currently an Associate Professor with the Department of Computer Science (CS), Kansas State University (K-State). He is also an affiliated (ancillary) Faculty Member with the K-State Department of Electrical and Computer Engineering. He holds a Michelle Munson-Serban Simu Keystone Research Faculty Scholarship from the College of Engineering. He has published several scholarly peer-reviewed articles in prestigious journal and conferences, with three of his research papers receiving the best paper awards and two more being selected as the best paper finalists. He has published a book *Modeling and Optimization of Parallel and Distributed Embedded Systems* (Wiley–IEEE, 2016). He also holds three U.S. patents. His current research interests include embedded and cyber-physical systems, secure and trustworthy systems, computer architecture, artificial intelligence (AI), computer vision, parallel computing, and fault tolerance. He was the recipient of many academic awards, including the Doctoral Fellowship from the Natural Sciences and Engineering Research Council (NSERC) of Canada.