Fraud Detection in Healthcare Based on Machine Learning

Aryan Singh
aryansingh57109@gmail.com
SCSET
Bennett University (Times Of India Group) ,
Greater Noida – 201310 (U.P),
India

Harshita Jain
harshitaa.jen@gmail.com
SCSET
Bennett University (Times Of India Group),
Greater Noida – 201310 (U.P),
India

Abstract

Healthcare is one of the most important aspects in peoples life and we need to make it affordable. Healthcare is a complex system of interlocking parts. It is growing at a rapid pace. At the same time, this has become a really big realeaytion of fraud in this industry. The main part of it is abuse of the medical insurance systems. Detecting healthcare fraud manually is tedious work. Automatic detection of healthcare fraud using machine learning and data mining techniques has recently been explored. Abstract In this paper, we try to provide a review of fraud in healthcare and methods that help for detecting such fraud. Based on the techniques used, we will identify the major sources and characteristics of healthcare data for different research studies conducted in the literature.

Based on this review it can be inferred that the novel machine-learning methods and newly obtained forms of healthcare data would be promising topics of research to make health care more affordable, enhance the performance of healthcare fraud detection, and provide high-level quality to health systems. This paper provides a review of a lot of recent research which implements machine learning and data mining to identify the fraud in healthcare sector. Further, it is proposed that several unique features of health insurance claims systems can be identified to study the uncommon behavior of claims submissions and more advanced ML algorithms can be used for better performance.

Keywords: Machine-Learning, Fraud Detection and Healthcare

INTRODUCTION

Healthcare plays a vital role in people'ss lives and needs to be within financial reach. The healthcare sector is a complex network with numerous interconnected elements. It is growing rapidly, but at same time, fraud has become a significant issue. Manually identifying fraud within the healthcare domain is a challenging task. Recently, techniques involving machine learning and data mining have emerged for the automatic detection of healthcare fraud.

In this paper, we aim to review the occurrence of fraud in the healthcare sector and the methods employed to identify such fraudulent activities. By focusing on the techniques utilized and identifying key sources and features of healthcare data, we studied various research works available in the literature.

Healthcare fraud is gradually regarded as one of the most significant social concerns. The ewe: Healthcare fraud is an evident issue. They were skittish of the government and detection also just needs to be better methods. Unsurprisingly, however, it takes a lot of glitter to be able to detect 0% individual healthcare fraud. Traditionally, fraud detection in health care relies heavily on the expertise of domain experts (which is wrong, costly and time-consuming). At current, detecting fraud in the healthcare domain is still a manual process, free from having few auditors review and determine suspicious medical insurance claims, which takes countless hours of work. However, these recent advances in machine learning and data mining techniques have brought efficient and automated detection of healthcare frauds. Over the past few years, there has been an increasing interest in utilizing data mining to detect fraudulent behaviour in healthcare.

RELATED WORK

Fraud in the healthcare sector has significantly enhanced losses for both people and organisations, and nation states [1]. Fight against healthcare fraud has become utmost important. Therefore, many researchers implemented fraud detection systems in healthcare. Fraud Detection systems are asked to discover, detect, and report frauds being played on the system they watch for. Typically, there are two modes in which Fraud detection is executed. Previously, to identify the fraud, manual fraud audit regulations were enforced [3]. Auditing requires a wealth of expertise and skill within that domain. These procedures result from complex transactions, are time-intensive, and The process is repetitive and manual, requiring a significant amount of time to execute. So it was automatic systems that were developed to detect fraud. These computerized systems are Integrated.

Given our focus on data mining, we must consider various fraud types, particularly in healthcare, and explore effective detection methods.

PROBLEM STATEMENT

The sophistication and prevalence of healthcare fraud represent a staggering financial loss of billions per year, Maryann has said in press release. Fraud detection methods that rely on predefined rules are often not sophisticated enough to identify complex patterns of fraudulent behavior, and can generate a lot of false positives. The Proposed research builds an efficient machine learning-based fraud detection system based on LOF and Random Forest algorithms to:

1. Detect unusual patterns in healthcare claims data that may be fraud
2. Lower rates of false positives in fraud detection
3. Real-time Enable Fraud Detection of Potential Cases
4. Enhance the precision of fraud classification in healthcare systems

The research questions specifically addressed by the study are:

a) How can we boost the detection ability of fraud healthcare claims by integrating LOF and Random Forest algorithms?

b) What are the specific characteristics and parameters that really impact their performance towards fraud detection?

c) c)Comparison of the systematic literature review and the proposed hybrid approach with traditional fraud detection methods including accuracy, efficiency, and false-positive rates.

The aim of this research is to design a more reliable and effective fraud detection process which would be able to deal with changing patterns of fraud while achieving high rates of detection against false positives in the context healthcare claims processing.

Key Components of the Problem Statement:

1. Challenge Identification:
    - Increased sophistication of healthcare fraud
    - Challenges with conventional detection techniques
    - The demand for detecting in real-time
2. Scope:
    - Check out the ML based approach
    - LOF and Random Forest algorithms Hybridization
    - Use case – healthcare claims processing
3. Objectives:
    - Minimize Wrong Alarms
    - Enable real-time detection
    - **Enhance pattern recognition**
4. Expected Outcomes:
    - Development of a hybrid detection system
    - Improved fraud detection rates
    - More efficient classification of fraudulent claims
    - Better resource allocation for fraud investigation

A. Types of Frauds in Healthcare

Healthcare frauds have several fraudulent behaviors that change to the occasion. It is specific topic for every country. Several types of fraud occur in the healthcare industry. The types of fraud can be classified on the basis of which group or individuals are engaging/engaged in the fraud :

- Fraud by Services Providers
    - Services providers may bill for the medical's services that are not actually performed;
    - Services providers can bill for the stages of a medical procedure as through as though they were separate treatment;
    - Services providers might charge for expensive medical services than one actually performed;
    - Just to generate insurance payments, service providers may perform unnecessary medical services;
    - Providers misrepresent the covered treatment as medically necessary & some other may not cover treatments just for getting it insured overall in to account of obtaining insurance;
    - Validation of unnecessary medical practices, a service provider may decide to alter the patient diagnosis and treatment history;

- Fraud by Insurance subscribers:
    - Employment / eligibility records may be falsified in order to get a cheaper premium rate;
    - Subscribers can submit reimbursement claims for medical services which are not actually provided;
    - Subscribers may use someone else's coverage or insurance card to fraudulently obtain the benefits;

- Frauds by Insurance carriers:
    - Fake reimbursements;
    - Misrepresenting ofbenefit / service statements;

- Conspiracy frauds:
    - In such frauds there are two or more parties involved for example patient and doctor/insurance company etc;

B. Data for Healthcare Frauds

The raw data for healthcare fraud detection is generally insurance claims which comes from many different sources. Except for insurance claims data, other types of data used in the healthcare fraud detection are physicians' data, prescriptions data given to these physicians, medication or drugs derived from these prescriptions and bills or transactions. Healthcare system data characteristics are unique to each country. Consequently, the retrieval of work in fraud detection gets assessed factoring the narrative of government health data. U.S. Health Care Financing Administration (HCFA)[3] is one of the major governmental departments in health field. Medicare and Medicaid are health care programs in the Medicare and Medicaid data, which cover medication and drugs performed; bills & transactions made; health care providers are mostly used by the U.S to identify frauds and abuse realised in the health systems.

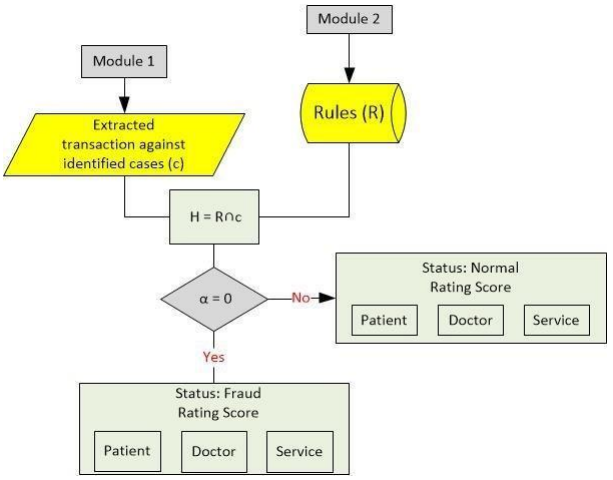


Fig1: a two-module approach for healthcare fraud detection

C. Methods for Healthcare Frauds Detection

Frauds tends to have many little minor specification included in different complex various types of patterns and their data is accumulated over long time span. Finding such patterns is incredibly difficult in a world with so much data but few ways to evaluate it. The industry standard has been thousands of heath care claims by parade of auditors. So traditionally fraud detection was a job for the veterans. But this process becomes so much slower and tedious as we have larger dataset. More focus on automated/automated fraud detection systems With improved data mining and machine learning tools. Machine-learning based techniques for detection of anomaly and detection of fraud are behavioural profiling methods that are used.

The three different behaviourral profiling methods based on machine learning techniques are used for the detection of anomaly and detection of fraud, and for this purpose, behavior pattern if each person who is involved in healthcare system is configured to observe to check deduction from standards. Many of the researchers categorize Data mining methods into two classes which are called supervised and unsupervised learning. But, in somecases, along with these two approaches, semi- supervilearning is also involved in these classifications.

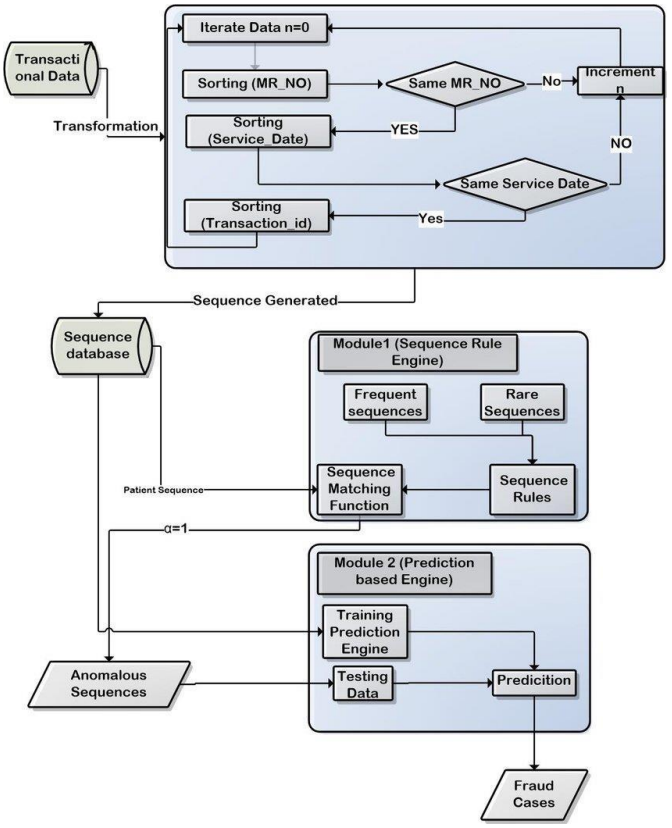


Fig2: a data-driven approach for healthcare fraud detection

Algorithm1 (Local Outlier Factor (LOF))

Methodology

A. Local Density Estimation

LOF computes the local density of all data points relative to their closest neighbors. Specifically, it is calculated through the minimum number of data points (k) that lie within a specified distance (epsilon) from the point in question[3][4].

B. Comparison with Neighbors

Each data point's local density is compared with its neighbors' local densities. Points in sparse regions (i.e., points with lower local density compared to their neighbors) are referred as potential outliers[3][4].

C. LOF Score Calculation

LOF score quantifies how much of an outlier each individual

data point is. The larger the LOF score, the greater likelihood of being an outlier[3].

Advantages

A. Invariance to Local Density Variations

LOF is robust as it can also detect outliers in the situation where the local density varies, making it well suitable for complex and dynamic data[3].

B. Scalability

It is efficient on larger datasets, since it works locally with respect to the neighbourhood only (offering lower computational complexity than global methods)[3].

C. Flexibility

It is applicable on all kinds of data such as numerical, categorical and mixed which makes it versatile to be used in different domains[3].

D. Interpretability

The LOF scores offer interpretation of the quality of an anomaly as a relative score for how much anomaly is contributed by each data point, useful to discriminate anomalous data points from normal ones and will help in making decisions about which data points are more anomalous than other[3].

Algorithm2 (Random Forest)

Methodology

A. Ensemble Learning

Random Forest is an ensemble learning method that combines several decisions trees to improve the accuracy and robustness of predictions. Each decision tree is trained on a random subset of the features and a random subset of the data points[5].

B. Classification

In fraud detection, Random Forest can be used to classify healthcare claims as either legitimate or fraudulent based on a number of features such as patient demographics, claim amounts and provider information.

C. Feature Importance

One advantage of using Random Forest is that it can identify the features that were most important to the classification, which provides insights into the patterns and relationships within the data associated with fraud.

D. Handling Imbalanced Data

In cases where the data is heavily imbalanced between two classes e.g. because fraudulent claims are many times more rare than legitimate claims, Random Forest works well when coupled with for example SMOTE (Synthetic Minority Over-sampling Technique)5].

Advantages

A. High Accuracy

Due to ensemble nature, Random Forest generally gets good accuracy because it is very less prone to overfitting and more generalized [5].

B. Handling High-Dimensional Data

Random Forest can effectively handle high-dimensional data, making it suitable for complex healthcare datasets that involve numerous features [2].

C. Computational Efficiency

The Random Forest model is computationally efficient, and it can scale to large datasets relatively quickly despite being complex [2].

D. Feature Selection

Random Forest also gives the feature importance score which helps to use only the relevant features in the model thus reducing dimensionality and improving model interpretability [5].

By combining these two algorithms, you can leverage the strengths of both density-based anomaly detection (LOF) and ensemble learning (Random Forest) to create a robust and accurate fraud detection system in healthcare.

CONCLUSION

This study was conducted on healthcare frauds, types of healthcare frauds, types and also sources of healthcare data and methods for healthcare frauds. Several studies are considered in the literature. Thus, it is inferable that under the view of the healthcare domain Data' is an overarching concern. Most of the data comes from government sources and private insurers. Abstract Mostly Healthcare fraud detection is used by using machine learning and data mining. Machine learning approaches are of three types, supervised, unsupervised and semi-supervised. Many researchers use semi-supervised learning approaches in most of the scenarios. However, in some instances new semi- supervised learning methods can be suggested for exception that augment the ability of identifying frauds in health care system. However, to hide every single Healthcare instance there doesn't exist any standard approach or patterns. This review concludes that sophisticated machine learning methods and novel sources of health data will be emerging topics because the focus on making healthcare cost effective, means to enhance efficiency of medical fraud detection and ensuring high quality health care systems.

FUTURE SCOPES

After reviewing different studies/thesis on healthcare frauds detection, we realize that the very nature of frauds or abuse occurring in health insurance systems can be of different

unusual patterns. More research work must be done on advanced data mining and machine learning techniques to detect such suspicious patterns. It is important to suggest new methods while keeping specific aspects of healthcare data in mind. In order to do this, the correlation of various entities of health care data can be considered.

LITERATURE REVIEW

The Center/Centers for Medicare and Medicaid Services (**CMS**) releases healthcaree data which is used by most researchers for healthcare frauds detection.

Srinivasan et al. [1] proposed an anomly detection technique using unsupervised rule-based data mining based on insurance claims data obtained from medicare data Big data applications for healthcare fraud, abuse, waste and error detection and analysis of health insurance claims were developed. Once lost in the noise, now these applications are used to avail private health insurers identify hidden cost overruns that transaction processing systems can't detect-specifically hybrids of medical insurance claim anomalies.

Branting et al. extracted supervised Healthcare data through leveraging existing Medicare and Medicaid data, as well as employing graph analyzation techniques for decision tree expansions. They suggested the method to estimate risk of healthcare fraud by utilizing network algorithms on network graph from open-source dataset.

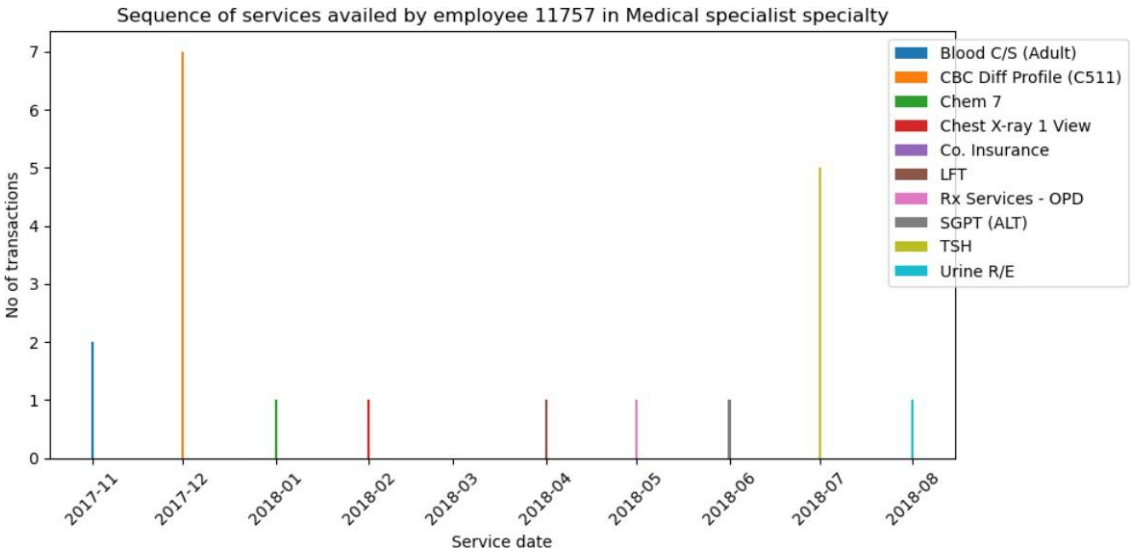
It was then able to conclude how a physician practices in terms of its analysis on the physicians past education by using the CMS 2012 data for research. They then compared medical school charges, procedures, and payments as well as find possible anomalies in the data by presenting a geographical analysis with the national distribution of school procedure payments and charges. The authors seek to find associations between training environments and the types of procedures that practicing physicians perform in order to detect abuse or inadequacy in medical insurance systems.

Ko et al.[1] also used 2012 CMS data, it limited its focus to a single field, Urology. Estimate National Savings Based on Urologist Service Utilization — The authors relate their estimated savings to a consistent service utilization against variability in the field of Urology by utilizing within specialty estimates for both service utilization and payment.

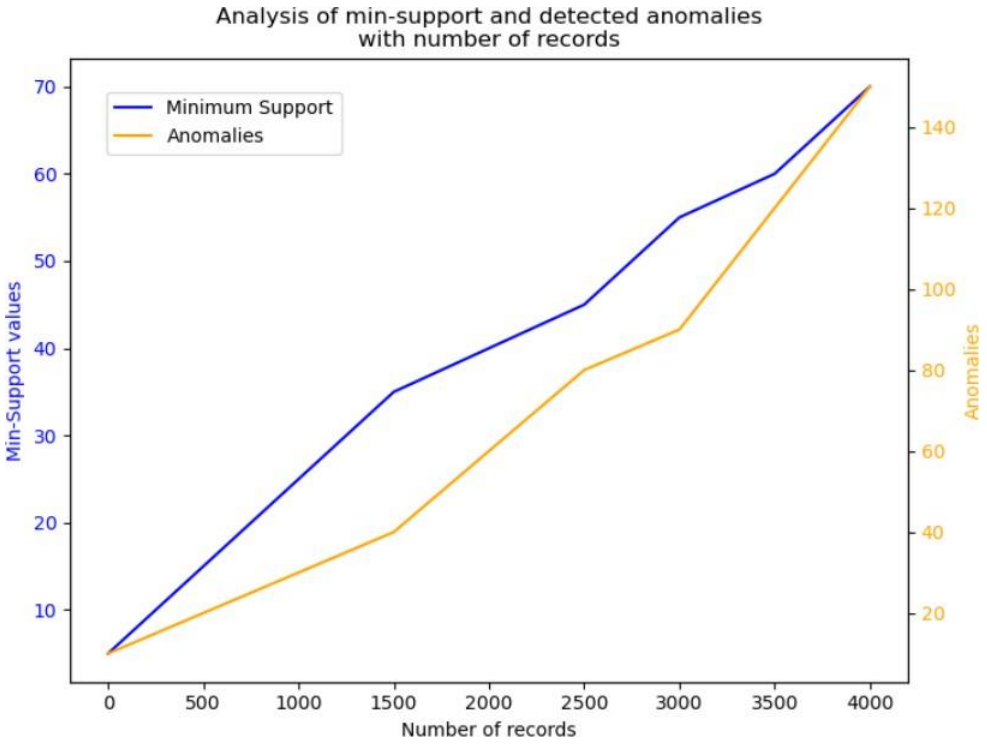
Using a 2013 CMS dataset[3], one study constructed a ML model to identify deviaton behavior by physicians in claiming medical insurance bounds. It attempts to ascertain whether, and when, doctors are practicing outside the norms of their specialty or otherwise practicing in a way that suggests mis-usage, fraudulent behaviour by defaulting billing codes used, or basic ignorance over billing procedures.

Using 5-fold cross-validation, precision, recall, and Fscore, are computed to evaluate the model. It leverages the multinomial Naïve Bayes technique. The F-score of over 0.90 for predicting multiple classes of physicians shows great potential to leverage ML in a novel way

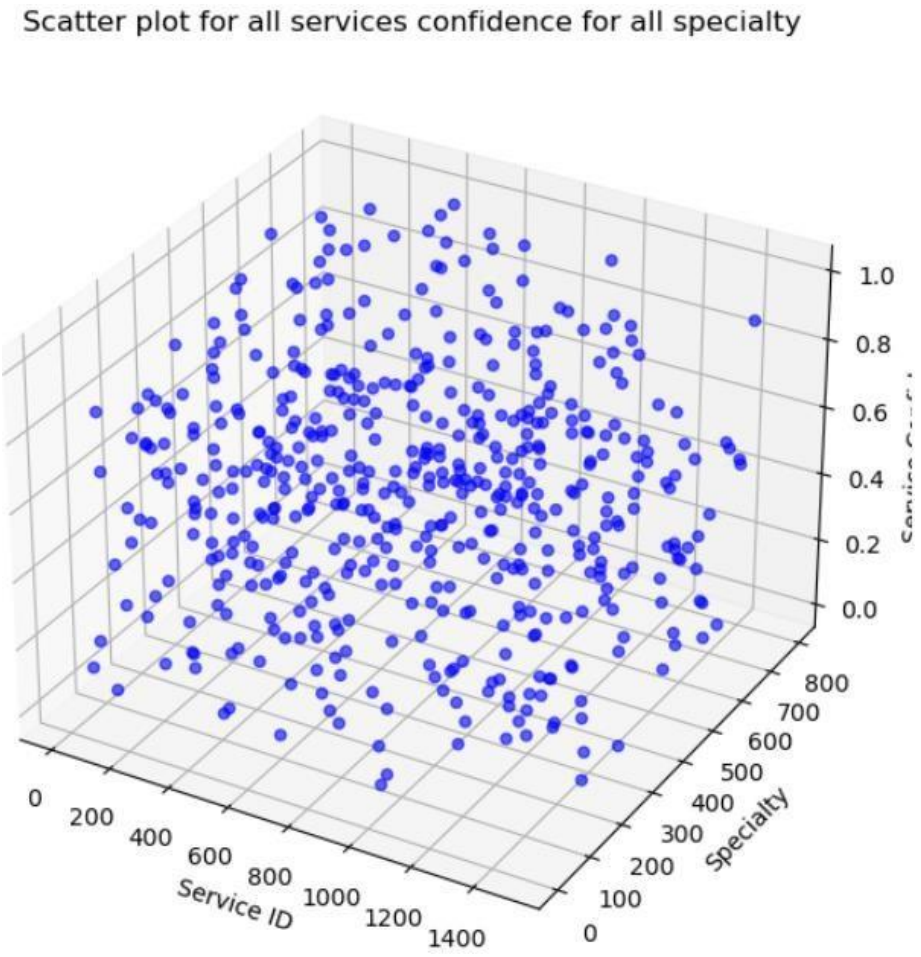
THE OUTPUTS OF OUR REPORT



Output1: Sequence of services availed by employee in medical specialist specialty



Output2: all the inputs and outputs parameters



Output3 : Scatter plot of all the services

Table1. Attributes in Dataset.

Attributes	Value
Unique number of service	1206
Unique number of doctors	486
Unique number of specialty	62
Total number of transactions	441,506

Table 2. Attributes of each transaction in the Dataset.

Attributes	Data Types
MRNO	Varchar (255)
Gender	Char
Date of birth	Date
Employee ID	Varchar (255)
Department Name	Char
Relation	Char
Service ID	Varchar (255)
Service Description	Char
Doctor	Varchar
Specialty	Varchar
Amount	Money
Category	Char

Table3. Differences [1]

Frameworks and References	Data Mining Approach	Type of Detected Fraud	Applied Data Mining Technique (s)
Predicting medical provider specialties to detect anomalous insurance claim	Supervised	Fraudulent payments in dermatology and optometry	Bayesian inference, using probabilistic programming
Medical school training relate to practice evidence from big data	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	Interactive machine learning	Prescription-based abnormal behaviour	Pair wise comparison expectation maximization (EM)
Outlier-based Health Insurance Fraud Detection	Unsupervised	Dental provider-related frauds	Multi-dimensional data models Multivariate Clustering
A Survey: Healthcare fraud detection	Hybrid	Rehabilitation, Septicemia Pneumonia, payment-related fraud detection	Geo-location Cluster analysis
Knowledge discovery from massive healthcare claims data	Hybrid	Providers Related Frauds Social network	Geo-loaction, social network analysis methods
Predicting healthcare fraud in Medicaid	Hybrid	Patient-related frauds, Physician-related frauds	Data models for patient claim and physician