International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-10 Issue-9, July 2021

Fraud Detection in Healthcare System using Symbolic Data Analysis

Sahana Munavalli, Sanjeevakumar M. Hatture

Abstract: ***In the era of digitization the frauds are found in all categories of health insurance. It is finished next to deliberate trickiness or distortion for acquiring some pitiful advantage in the form of health expenditures. Bigdata analysis can be utilized to recognize fraud in large sets of insurance claim data. In light of a couple of cases that are known or suspected to be false, the anomaly detection technique computes the closeness of each record to be fake by investigating the previous insurance claims. The investigators would then be able to have a nearer examination for the cases that have been set apart by data mining programming. One of the issues is the abuse of the medical insurance systems. Manual detection of frauds in the healthcare industry is strenuous work. Fraud and Abuse in the Health care system have become a significant concern and that too inside health insurance organizations, from the most recent couple of years because of the expanding misfortunes in incomes, handling medical claims have become a debilitating manual assignment, which is done by a couple of clinical specialists who have the duty of endorsing, adjusting, or dismissing the appropriations mentioned inside a restricted period from their gathering. Standard data mining techniques at this point do not sufficiently address the intricacy of the world. In this way, utilizing Symbolic Data Analysis is another sort of data analysis that permits us to address the intricacy of the real world and to recognize misrepresentation in the dataset.***

Keywords: Data mining; Health insurance; Fraud detection; Anomaly detection; Symbolic data

I. INTRODUCTION

Healthcare systems like organizations or policies set up that are intended to design and give clinical consideration to individuals. Clinics, Hospitals, and community health agencies. Healthcare systems are unpredictable, and numerous things should be thought about kinds of emergency hospital systems, patient care, insurance, healthcare providers, and lawful issues. Medical services have become a significant consumption in the US. since 1980.

Manuscript received on June 20, 2021.
Revised Manuscript received on June 30, 2021.
Manuscript published on July 30, 2021.
* Correspondence Author
Sahana Munavalli*, Department of Computer Science and Engineering, Basaveshwar Engineering College (Autonomous), Bagalkot (Karnataka), India.
Sanjeevakumar M. Hatture, Department of Computer Science and Engineering, Basaveshwar Engineering College, Bagalkot Under Visvesvaraya Technological University, Belagavi (Karnataka), India.

Both the size of the healthcare sector and the gigantic volume of cash included make it an alluring fraud target. There are various kinds of healthcare care systems, for example, National Health Service, non-profit national health system, National Health Insurance System (NHIS), social health insurance, private health system, social health insurance system, etc. And so forth as per the Office of Management and Budget, in 2010, about 9%, or around $47.9 billion of the US'S Medicare expenditure was lost because of fraud. Consequently, effective fraud detection is significant for diminishing the expense of the healthcare system. Fraud is the crime of acquiring cash or monetary advantages by a trick or by lying. Fraud can be spread comprehensively, and it is very exorbitant to ensure the system. The medical services industry is a multifaceted framework with various moving segments. Simultaneously, fraud in this industry is transforming into a critical issue. Distinguishing medical services fraud and misuse, in any case, needs concentrated clinical information. As indicated by ACFE (Association Certified Fraud Examiner) Report Estimation, Organizations Worldwide Lose 5% of Revenues to Fraud i.e., Projected Losses Exceed $3.5 Trillion Worldwide. In medical services, fraud can happen in various circumstances, from superfluous and duplicate tests and strategies to hacking into a patient's clinical records to submit bogus claims. In clinical medical care systems, fraud might be forced like Billing for services not rendered, Billing for a non-covered assistance as a covered help, Misrepresenting locations of service, Misrepresenting areas of administration, Misrepresenting supplier of overhauling of deductibles or potentially co-installments, Incorrect revealing of diagnoses or procedures (incorporates unbundling), Overutilization of services, Corruption (payoffs and bribery), False or pointless issuance of doctor prescribed medications.

As the information turns out to be more perplexing than the standard ones since they contain internal variations and are organized. To sum up enormous arrangements of information, the need to stretch out standard data analysis strategies to symbolic data tables is expanding to get more exact data and sum up broad data sets.

Symbolic Data Analysis (SDA) can be characterized as the expansion of standard Data Analysis to such tables. With the assistance of approaches present in Symbolic Data Analysis, the huge dataset is addressed as symbolic objects, symbolic similarity implies some are performed for fraud analysis.

The information in health monitoring system is mind boggling as it contains the biochemistry report, ECG, blood pressure, pulse rate and other health related data. This sort of data might be quantitative and categorical Thus, it is better decision to address such sort of information into symbolic objects. The symbolic object can be sorted into assertion, Hoard and Synthetic.

Benefits of addressing information into representative articles:

1. They give a summary of the original symbolic data in an illustrative manner.
2. They can be easily changed in terms of the query of a Database.
3. By being autonomous of the independent data table they can recognize any coordinating with individual portrayed in any data table.
4. In the utilization of their elucidating part, they can give another symbolic data table of a more elevated level on which a symbolic data analysis of the subsequent level can be applied.
5. To portray an idea, they can effectively join a few properties dependent on different variables coming from different arrays and different underlying populations.

With the assistance of symbolic data, fraud in health care can be distinguished by utilizing symbolic similarity measures and symbolic clustering. The symbolic clustering methodology forms composite symbolic objects items utilizing a Cartesian join operator when two symbolic objects are merged. This composite object with the rest of the object is additionally utilized for similarity analysis.

The rest of the paper is organized into four sections: section 2 reviews the developments in fraud analysis in healthcare systems. Further the common types of frauds healthcare systems are enlisted in section 3. Symbolic data representation is described in section 4. The proposed model of fraud detection in healthcare systems using symbolic data analysis is described in section 5. The section 6 concludes the proposed work.

II. LTERATURE SUMMARY

The Centers for Medicare and Medicaid Services (CMS) releases health care information which is utilized by the greater part of the scientists for healthcare fraud detection. Lucy Fricker [1] proposed Enterprise Risk Management can assist with diminishing sorts of fraud. ERM is "a process, affected by an effected by an entity's board of directors the board and other work force, applied in methodology setting and across the enterprise, intended to recognize potential events that may influence the entity, and oversee hazard to be inside its danger craving, to give sensible confirmation with respect to the accomplishment of element goals of entity. Enterprise Risk Management can help reducing the fraud. Qi Liu[3] proposed a Geo-location clustering model. In this, preliminary knowledge of healthcare system and its fraudulent behaviors, and analyzes the characteristics of health care data and they compare currently proposed fraud detection approaches like clustering, decision making, etc. with Geolocation clustering model. Yi Peng [4] proposed SAS EM and CLUTO to a health insurance dataset to detect frauds. AS Enterprise Miner is an advanced analytics data mining tool intended to help users quickly develop descriptive and predictive models through a streamlined data mining process. CLUTO is a software package for clustering low- and high-dimensional datasets and for analyzing the characteristics of the various clusters. Shivani S. Waghade[5] presented machine learning and data mining are used for healthcare fraud detection and explained types of frauds, healthcare fraud and also conclude that the advanced machine learning techniques and newly acquired sources of the healthcare data would be forthcoming subjects of interest to make the healthcare affordable, to improve the effectiveness of healthcare fraud detection and to bestow a top-quality on healthcare systems. Ajith Abraham,[6] proposed a classification framework for the application of data mining techniques to insurance fraud detection and classified into three types of insurance fraud (automobile insurance, crop insurance, and healthcare insurance) and six classes of data mining techniques (classification, regression, clustering, prediction, outlier detection, and visualization).The main data mining techniques used for insurance fraud detection are logistic models, Decision trees, the Naïve Bayes, and support vector machine. Analysis. In this, proposed classification framework for the application of data mining techniques to insurance fraud Detection and the analysis results show that automobile insurance fraud in the most covered area of research (57%). The data mining application class that was used in most of the papers is the classification (57%). Matthew Herland [7] proposed a Supervised learning algorithm for class imbalanced data. Every day, there are a massive number of financial transactions generated by physicians administering healthcare services, such as hospital visits, drug prescriptions, and other medical procedures. Most of these financial transactions are conducted without any fraudulent intent, but there are a minority of physicians who maliciously defraud the system for personal gain. In machine learning, when a dataset portrays this discrepancy in class representation (i.e., a low number of actual fraud cases), it is known as a class imbalance. The main issue attributed to class imbalance is the difficulty in discriminating useful information between classes due to the over-representation of the majority class (non-fraud) and the limited amount of information available in the minority class (fraud).UC San Diego[8] proposed a health care fraud and abuse blockchain technical framework and prototype using key blockchain tools, for secure data storage and consensus mechanisms, which make the claims adjudication process more patient-centric to identify and prevent health care fraud and abuse. The primary aims of the blockchain solution are to (1) improve detection of potentially fraudulent and illegal health care transactions and reimbursements, (2) create a more inclusive process for validating claims deploying a patient-centric approach, and (3) enhance efficiency in the claims adjudication process through smart contract automation. Naga Jyothi [9] described various perspectives of data mining approaches for finding fraudulent activities. As there is massive data in health care systems, where it involves millions of people having different attributes for each tuple and data with various characteristics.

The characteristics of data describe the behaviors of both parties which involves fraud where each claim have their distinctive identifiers for insurance subscriber and service provider. With the help of these identifiers. It is possible to get behaviors over some time. Ms. Meena Kumari [10] proposed methods of identification, mitigation, and management of fraud are considered within the context of process improvements or modifications that can be implemented by the insurer. Renata M. C. R. de Souza [11] to manage ordered and non-ordered mixed feature type symbolic data. To be able to manage ordered and non-ordered mixed feature-type symbolic data, a previous pre-processing step was introduced to obtain a suitable homogenization of mixed symbolic data into modal symbolic data represented by weight distributions. The dynamic cluster algorithm with adaptive distances locally optimizes an adequacy criterion that measures the fitting between the classes and their representatives (prototypes). V. Ravi [12] a hierarchical agglomerative symbolic clustering methodology is Proposed. The procedure is based on the physical phenomenon in which a system of particles in space converges to the centroid of the system due to gravitational attraction between the particles are merged. The procedure terminates at some stage where there no available for merging. Anderson F.B.F. Costa [13] The main idea of this kernel k-means (KKM) consists of manipulating a kernel function for interval data to compute the distance between two vectors of interval data. kernel k-means(KKM) method is an extension of the classic K-means Kernel to symbolic interval data. A new clustering method is proposed for symbolic interval data based on kernel. This method is an extension of the classic K-means Kernel to symbolic interval data. The evaluation of this method can be done by comparing with a dynamic clustering method for interval data having adequacy criterion-based (on adaptive for each cluster) Euclidean distances carried out and result provided by clustering methods were assessed correctly using rand index by considering the synthetic and applications of real data sets. K. Chidananda Gowda [14] the clustering methodology forms composite symbolic objects using a Cartesian join operator when two symbolic objects are merged. Merging is the process of gathering, based on a similarity measure two samples and assigning them same cluster. The clustering methodology is proposed in a composite object when two selected objects are merged. This composite object, along with the rest of the objects of the set, is used in further similarity analysis. similarity components due to “position” which feature type is Quantitative and “span”,” content” which feature type is qualitative. A composite symbolic object is a new object resulting from merging two symbolic objects using a Cartesian join Operatorial. Vipin Kumar [16] explained a variety of similarity measures. similarity for continuous data is relatively well-understood, but for categorical data, the similarity computation is not straightforward, so for this data-driven similarity measures have been proposed for categorical data. In data-driven similarity measures for categorical data, a key task is to identify the characteristics of a categorical data set that affect the behavior of a similarity measure such as the size of data, number of attributes, number of values taken by each attribute, etc. This work used outlier detection as the underlying data mining task for the comparative evaluation. It will be useful to know if the relative performance of these similarity measures remains the same for the other data mining tasks. J. D. Kittoe [17] In this, the researcher focused on malaria cases data. In tackling the issue of malaria in a more cost-effective means, patterns were explored in finding the mean cost of drugs for the treatment of malaria. One of the major objectives is, to use data mining techniques to detect fraud and abuse in the NHIS concerning malaria-related cases[19]. Some of the issues and challenges are identified in the literature and enlisted in the following.

- Fetching irrelevant data from the data set.
- Categorizing large data set using classical data approach.
- Not easy to understand data when the dataset is more in classical data.
- Identification of genuine data.
- Using pattern recognition with the help of symbolic data.

Hence there is a scope to represent the complex healthcare data[20] containing the numeric, graphical, and video and signal data, by employing symbolic data representation. Using Symbolic Data, the healthcare data can be easily segregated in range which will help in detecting the fraud and Symbolic data analysis gives a new way of extending the standard input to a set of classes of individual entities. Such classes often represent the real units of interest, using the SDA more efficiently the output data can be retrieved.

III. FRAUD TYPES IN HEALTH CARE SYSTEM

A mind-boggling larger part of extortion occasions in the protection business follow a predetermined number of examples that are normally known to the protection specialists. Protection exchanges may have various kinds of fraud. In medical coverage can be explicit to every nation exploiting the insufficiency of the important enactment or being influenced by the nearby culture. For example, people in countries with a collectivist culture may have a higher inclination to mishandle the framework contrasted with the nations with individualistic societies. Individual and family ties are more grounded in the previous contrasted with the last mentioned and an uninsured individual may unlawfully get protection advantage masking himself as a guaranteed individual. This, obviously, requires the assent of the truly guaranteed individual. A specific model that is acknowledged to have some affinity to blackmail is a heuristic and ward on association experience. Yet every association has its own plan of such models, those models ordinarily cover. Regardless, the associations are ordinarily reluctant to uncover these models since they are stressed over fraudsters checking them (Morley et al., 2006).

Protection guarantees that match the realized examples can be effortlessly recognized by customary data set announcing instruments or programming languages. Be that as it may, this procedure gives just a harsh manual for protection specialists, on the grounds that solitary a little minority of such cases is to be sure fake. Henceforth, all cases that match the realized false examples should be firmly explored by specialists. This examination may target the guaranteed people as well as the colleagues like protection organizations, clinics, wellbeing focuses, and drug stores.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

Now and again the extortion may happen by the coordinated effort of various elements. It might even be submitted by the insurance agency representatives.

A portion of the misrepresentation types in the medical coverage area in India are as per the following:

- Charging over the top costs for treatment or medication.
- The surprisingly maximum number of solicitations for a specific insured in the brief period of time (3-4 days).
- Insurance transaction(s) where the insured has got some treatment or medication yet either has not paid any portions or has paid just the first installment.
- Cases where the insured purchasing medication without a clinical assessment.
- Claiming clinical solicitations with dates previously or after the start of the insurance time frame (this is allowed at times).
- An extreme number of clinical cases in a particular period.
- Bank account number changes of a colleague like office, health center, or drug store.
- Excessive quantities of manual receipt request whose sums are more modest than the standard assessment limit. claims whose payable sums are more noteworthy than the receipt sum that the insurance agency will pay.

Greater part of fraud in the clinical industry has a predetermined number of examples that are normally known to the insurance experts. For E.g., It may have various kinds of fraud. In medical coverage can be explicit to every nation exploiting the deficiency of the applicable enactment or being influenced by the neighborhood culture. For instance, individuals in nations with a collectivist culture may have a higher inclination to mishandle the framework contrasted with the nations with individualistic societies.

Fraud can occur in multiple situations, however, to handle the circumstance or to deal with the situation, the board or the associations should think about the intercession of innovation to keep away from the maltreatment of the assets.

IV. SYMBOLIC OBJECT

Different definitions and depictions of the symbolic objects are given. By Chidanda Gowda [6,18]. Symbolic objects are characterized by coherent combination of events preferring values and variables in which the variables can take one or more values need not be characterized on the same variables. We give under a nonformal portrayal of Symbolic objects of the sort of Assertion, Hoard and Synthetic.

As the information turns out to be more unpredictable than the standard ones since they contain inner variety and are organized. To sum up immense arrangements of data, the need to stretch out standard information examination techniques to symbolic data analysis table is expanding to get more exact data and sum up broad extensive dataset. Characterize "Symbolic Data Analysis" (SDA) as the extension of standard Data Analysis to such tables. Further with the help of approaches present in SDA, can analyze, and process the data. The purpose of SDA is to extend data mining techniques and traditional statistics to higher-level units. Symbolic data analysis gives a new way of thinking in Data Science by extending the standard input to a set of classes of individual entities. Such classes often represent the real units of interest. To take variability between the members of each class into account, classes are described by intervals, distributions, set of categories or numbers sometimes weighted, and the like. In that way, we obtain new kinds of data, called 'symbolic' as they cannot be reduced to numbers without losing much information. The symbolic data table is built where the rows are classes, and the variables can take symbolic values. Further extract new knowledge from these new kinds of data by at least an extension of Computer Statistics and Data Mining to symbolic data.

In Symbolic Data, the data is not restricted as classical data (standard data) and the data can be represented by lists, intervals, range, etc. as shown in Table 1.

Table 1: Symbolic Data Representation

Sl.No	Age Range	Blood Pressure (mm/Hg)	City	Type of Cancer	Gender
1	20,30	79/120	Boston	Brain tumor	Male
2	50,60	90/130	Boston	Lung, liver	Male
3	45,55	80/130	Chikago	Prostate	Male
4	47,47	86/121	EI Paso	Breast p, lung(1-p)	Female

V. METHODOLOGY

The block diagram proposed of fraud detection in healthcare system is depicted in Figure 1. Firstly, input needs to be given as a dataset, which consists of complex data, to this data cleaning and preprocessing steps to be carried out because to think about how exactly the missing data to be filled. If scaling of feature is considered and how it needs to be done. Here there will be a need for Dummy variables or a real dataset? Is data going to be encoded? Whether encoding of dummy variables is done? after pre-processing step, conversion of the input dataset into symbolic data to be done and analysis of it using the SDA technique should also be done. Finally, by using machine learning algorithms like clustering/decision-making categorization of data to be done where the dataset will be genuine or not.

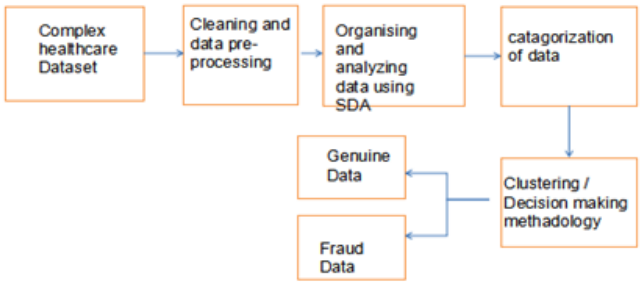

Figure 1:Fraud Analysis Using SDA Workflow

The block diagram of the proposed fraud detection in healthcare system using symbolic data analysis is depicted in Figure 1. The different processing stages of the proposed methodologies are acquisition of healthcare data (Dataset), data cleaning and pre-processing,

organizing, and representing the data into symbolic object and analyzing data using symbolic data analysis, categorization of the symbolic data and finally making the decision of data associated with fraud or genuine using pattern matching technique. The functionality of individual block is explored in the following:

A. Complex Health Care Data Set:

The patient data is collected from the patient data collected may contain the classical data stored in a database. In order to represent the classical data is passed through the pre-processing stage for further feature extraction. viz. data acquisition and construction of the healthcare Data Set.

B. Cleaning and Preprocessing:

The collected data into symbolic object the data are collected, often the data will not be in a form that is suitable for processing. To make the data suitable for processing and represented into symbolic object it is essential to transform them into a format the input such as multidimensional, time series, or semi-structured format. The feature extraction phase is often performed in parallel with data cleaning in which noisy data and irrelevant data are removed from the collection i.e., in collected data some data may be missed, or some may be irrelevant, for that purpose data cleaning is to be done. The result of this phase is a well-structured data set, which can be effectively used for representation as symbolic object.

C. Organizing and Analyze data:

Once the dataset is cleaned and pre-processed, the next phase is to analyze the data. With the help of Symbolic Data Analysis, using the mentioned. Technique large and complex data can be processed and visualized by converting the complex data into the range of data (Symbolic Object).

D. Pattern Extraction and Matching:

After analyzing the data through symbolic data analysis, the output viz. symbolic object will be categorized in the form of Genuine or Fraud based on the patterns recognized (hierarchical clustering algorithm). To detect, whether the given dataset is genuine or not, with the help of suspicious patterns with machine learning algorithms using symbolic data is performed. The experimental results are promising and detecting with an accuracy of 99%. The symbolic data representation of the healthcare data and symbolic data analysis are well acquainted for the healthcare systems.

VI. RESULTS AND DISCUSSION

The experimental results are added below for the both the tests carried for Skin Cancer Fraud Detection and Heart Disease Fraud Detection. For Skin Cancer, content similarity is being considered and similarly for heart disease, span similarity is considered.

For Skin Cancer Fraud Detection K-means algorithm is being used. It is an iterative algorithm divides dataset into separate clusters based on the similarity, by this way that each dataset will be segregated with similar groups that has similar properties.

Attributes considered for Skin Care Fraud Detection:

1. SO_{skin_cancer}=Melanoma{one half is unlike other half, irregular, 6mm or larger, varied from area to other, Changing color, size, shape}
2. SO_{skin_cancer}=Cancer{Asymmetry, Border, Diameter, Color, Evolution}
3. SO_{skin_cancer}=Squamous_cell_carcinoma{one half is like other half, scalloped, 6mm or larger, no variation, Changing color}
4. SO_{skin_cancer}=Bascal_cell_carcinoma{one half is unlike other half, poorly defined, below 6mm,Varied,Changing size}
5. SO_{skin_cancer}=Merkel_cell_Cancer{one half is unlike other half, Poorly Defined,6mm or larger, varied from other area, Changing shape}

For Heart Disease Fraud Detection, Naive Bayes algorithm is used which is a supervised learning algorithm, it is based on Bayes theorem and used for solving classification problems. It is mainly used in classification of text that has a high-dimensional training dataset. It is a probabilistic classifier, i.e., it predicts on the basis of the probability of an object.

Considering the below table 2, the abbreviations of the terminologies are added with respect to the respective Cancer types, and the various attributes which are being considered during the execution. The input data provided by the user through Excel file as a dataset will be passed and the result will be obtained will be based on the respective algorithm used.

Attributes considered for Heart Disease Fraud Detection:

SOHeart_Disease=Heart{Age, Gender, Cp, restbp, chol, fbps, restecg, thalac ,exang, slope, oldpeak, ca, thal}

The above-mentioned respective attributes are considered while comparing the fraud detection.

Table 2: Abbreviations

S=Skin Cancer Types:	S1=Melanoma S2=Squamous cell carcinoma, S3=Bascal cell carcinoma S4=Markel cell Cancer
A=Asymmetry:	A1= One half is unlike another half A2= One half is like another half
B=Border:	B1= Irregular B2= Not clear B3= Scalloped B4=Poorly Defined
C=Color:	C1= Varied from Area to other C2=No Variation C3= Other
D=Diameter:	D1=6mm or larger D2=Below 6mm D3=Other
E=Evolution:	E1= Change in color, size, shape E2= Change in color E3=Change in Shape
S=Status	G= Seems to be Genuine F= Seems to be Fraud

In the below mentioned Table 3, w.r.t Table 2 abbreviations are used, and the result is obtained comparing the user input with the processed data

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

Table 3: Output Result table

Sl. No	S	A	B	C	D	E	S
1)	S1	A1	B1	C1	D1	E1	G
2)	S1	A2	B2	C1	D1	E1	F
3)	S2	A2	B3	C2	D1	E2	G
4)	S2	A2	B2	C2	D1	E2	F
5)	S3	A1	B1	C1	D1	E1	G
6)	S3	A2	B1	C1	D1	E1	F
7)	S4	A1	B4	C1	D1	E3	G
8)	S4	A1	B4	C1	D1	E2	F

Table 4: Abbreviations

A=Age:	A1=Above 15 and below 25 A2= Above 35 and below 45 A3= Above 45 and below 65 A4= Above 65 and below 100
G=Gender:	G1= 1 for Male G2= 0 for Female
CP=Chest Pain:	1= Typical Angina 2= A-Typical Angina 3= non-Angina 4=A-symptomatic pore
R=Rest BP:	R1= Normal (<120/80 mm Hg) R2=Prehypertension(120-129 and 80-89-mm Hg) R3= Hypertension(>130/90 mm Hg)
C=Cholesterol:	C1= Less than 200 mg/dL C2=200-239 mg/dL C3=>240 mg/dL
E=ECG:	E1= 0 for Normal E2= 1 for Having ST-T E3=2 for Hyper Therapy
S=Status	G= Seems to be Genuine F= Seems to be Fraud

As mentioned below i.e., in Table 5, w.r.t Table 4 abbreviations are used, and the result is obtained comparing the user input with the processed data

Table 5: Output Result table

Sl. No	A	G	CP	R	C	E	S
1)	A1	G1	2	R2	C2	E1	G
2)	A1	G1	2	R3	C1	E1	F
3)	A2	G1	2	R3	C2	E3	G
4)	A2	G1	2	R1	C1	E1	F
5)	A3	G2	3	R3	C3	E2	G
6)	A3	G2	1	R1	C1	E1	F
7)	A4	G2	1	R3	C2	E3	G
8)	A4	G2	4	R1	C1	E1	F

VII. CONCLUSION

Anomaly detection, clustering, and classification can successfully detect anomalies or outliers in large sets of data. This can be very useful for the insurance industry which has problems with fraudulent claims. Once the anomalous claims are detected, several analyses must be made on them to conduct a thorough investigation. The main task in these analyses is to narrow the target for detecting frauds.

REFERENCES

1. Lucy Fricker, "Causes and Challenges of Healthcare Fraud in the US", North Carolina State University Raleigh, NC, USA,2013
2. https://www.researchgate.net/publication/220924701
3. Qi Liu," Healthcare fraud detection: A survey and a clustering model incorporating Geo-location information", Rutgers University Newark, New Jersey, United States,2013.
4. Yi Peng," Application of Clustering Methods to Health Insurance Fraud Detection", Institute of Information Science, Technology & Engineering, USA,2006.
5. Shivani S. Waghade," A Comprehensive Study of Healthcare Fraud Detection based on Machine Learning", Shri Ramdeobaba College of Engineering and Management, Nagpur,2018.
6. Ajith Abraham," Computational Intelligence Models for Insurance Fraud Detection", Machine Intelligence Research Lab, WA, USA,2013.
7. Matthew Herland, "The effects of a class rarity on the evaluation of supervised healthcare fraud detection models", Florida Atlantic University,777 Glades Road, Boca Raton, FL, USA,2019.
8. UC San Diego," Combating Health Care Fraud and Abuse: Conceptualization and Prototyping Study of a Blockchain Antifraud Framework", School of Medicine, Department of Anesthesiology and Division of Infectious Diseases and Global Public Health, La Jolla, CA, United States,2020.
9. P. Naga Jyothi," Performance on Fraud Detection in Medical Claims of Healthcare Data",2019.
10. Ms. Meena Kumari," FICCI Working Paper on Health Insurance Fraud", Joint Director, IRDA,2019.
11. Renata M. C. R. de Souza, "A Clustering Method for Mixed Feature-Type Symbolic Data using Adaptive Squared Euclidean Distances", Centro de Informatica - CIn / UFPEAv. Prof. Luiz Freire, s/n - Cidade Universitaria,2007.
12. T. V. Ravi, "Clustering of Symbolic Objects Using Gravitational Approach", IBM Solutions Research Center, Indian Institute of Technology, New Delhi, India, 1999.
13. Anderson F.B.F. Costa," A Kernel K-means Clustering Method for Symbolic Interval Data", Federal University,2010.
14. K. Chidananda Gowda," Symbolic Clustering Using a New ////99Similarity Measure", S.J. Coll. of Eng., Mysore, India,1992.
15. Miin-Shen Yang, "Fuzzy clustering algorithms for mixed feature variables", Department of Applied Mathematics,2003.
16. Vipin Kumar "Similarity Measures for Categorical Data: A Comparative Evaluation", Proceedings of the SIAM International Conference on Department of Computer Science and Engineering University of Minnesota,2008
17. J. D. Kittoe, "Exploring fraud and abuse in National Health Insurance Scheme (NHIS) using data mining technique as a statistical model", Department of Computer Science and Engineering,2017.
18. Shanmukhappa A Angadi, Sanjeevakumar M Hatture, Face Recognition Through Symbolic Modeling of Face Graphs and Texture, International Journal of Pattern Recognition and Artificial Intelligence, 33(12), 2019.
19. Rashmi. P.Karchi Nagarajan Munusamy, Exploration of Unmixing and Classification of Hyperspectral Imagery, International Journal of Innovative Technology and Exploring Engineering(IJITEE), 8(7), pp. 723-733, 2019.
20. Sanjeevakumar M Hatture, Nagaveni Kadakol, Clinical diagnostic systems based on machine learning and deep learning, Demystifying Big Data, Machine Learning, and Deep Learning for Healthcare Analytics, pp. 159-183, 2021.

AUTHORS PROFILE

Sahana Munavalli, is pursuing M.Tech in Computer Science and Engineering. She is working with fraud analysis in healthcare system in fourth Semester MTech, in the Department of Computer Science and Engineering, Basaveshwar Engineering College(Autonomous), Bagalkot - 587103, Karnataka State, India. Her area of interests includes network and information security.

Sanjeevakumar M. Hatture, received the bachelor's degree in Electronics and Communication Engineering from Karnataka University, Dharwad, Karnataka, India, and the master's degree in computer science and Engineering from Visvesvaraya Technological University, Belagavi, Karnataka, India, and Ph.D. Degree in the Department of Computer Science and Engineering at Basaveshwar Engineering College, Bagalkot under Visvesvaraya Technological University, Belagavi, Karnataka, India. His research interests include biometrics, machine learning, image processing, pattern recognition, soft-computing, Cyber security Internet of Things and network security. He is life member of professional bodies like IEEE, IEI, ISTE, IAENG and IRED.