DOI 10.29042/2018-4061-4070

Helix Vol. 8(5): 4061- 4070

Novel Healthcare Fraud Detection Approach to Identify Aberrant Medical Practitioners

[*1] Shivani S. Waghade, [2]Aarti M. Karandikar

[1, 2] Shri Ramdeobaba College of Engineering and Management, Nagpur - 440 013 (M.S.), India

**Email: waghadess@rknec.edu; karandikara@rknec.edu*

Received: 09th July 2018, Accepted: 14th August 2018, Published: 31st August 2018

Abstract

The Healthcare industry has been expanding enormously. Simultaneously, fraud in the healthcare industry is becoming a critical problem. One of the serious issues is the misuse of healthcare insurance systems. The misuse and fraud in the healthcare insurance systems significantly raise the costs of healthcare services for the providers and patients. One of the methods for healthcare fraud detection, through which these costs can be decreased, is to detect unusual medical practices which could possibly signify misuse or fraud. In this research paper, we have proposed a fraud detection approach which attempts to detect which medical practitioners show aberrant behaviour in their medical insurance claims. Based on the procedures performed by the medical practitioners, this algorithm tries to determine if, and when, medical practitioners function out of the standards of their own specialty, which may signify misuse of medical insurance systems or dearth of knowledge about billing policies. This research analyses the performance of the proposed approach against the Multinomial Naïve Bayes algorithm, used in the previous research to detect anomalies in the publicly available U.S. Medicare dataset. By analysing the results obtained by both the algorithms, we were able conclude that the proposed fraud detection approach yields better results.

Keywords: Data Mining; Healthcare; Fraud Detection; Multinomial Naïve Bayes Algorithm; Link Analysis Algorithm;

Introduction

Healthcare is a prominent concern for all the people in the world. Hence, for almost all the countries, healthcare has become a prime expenditure. As a result, healthcare industry has become one of the largest growing industries creating a greater impact on the countries' economies. Healthcare industry is a compound and complicated system. It involves various specialized medical practitioners qualified to diagnose several diseases and perform various types of medical treatment plans and procedures on the patients. But in every healthcare organization, there are physicians, pharmaceutical dealers and manufacturers, and medical staff that are to be paid along with expensive medical equipment used for the treatment. As a result, every medical treatment and procedure is associated with substantial cost which is often not affordable by many patients. Hence, to make the medical treatment plans and procedures affordable to more number of people, medical insurance schemes were introduced to dispense costs across the healthcare network and so that the patients and the medical equipment can be paid uniformly. But with the growing network of the medical insurance systems, there has been a rapid rise in the misuse and fraudulent activities.

The fraud in the healthcare industry has been a perpetual and critical problem for the government and taxpayers. The healthcare fraud detection has always been a matter of interest for the researchers. Traditionally, the healthcare frauds were detected manually by the auditors who used to manually assess the medical insurance data to identify unusual and suspicious claims. The traditional methods are quite expensive and time consuming. The manual detection of frauds or misuse in the healthcare insurance systems requires immense efforts and depends on the knowledge of the domain experts which could be specious. Hence, there was a need of automated systems to detect frauds in an efficient way. The modern advances of machine learning and data mining techniques paved the way for more efficient and automated detection methods. Thus, lately, in order to innovate more novel techniques, that can detect healthcare frauds automatically, mining of healthcare insurance data has gradually become a matter of interest.

There are several different types of healthcare frauds based on which party commits the fraud. The different parties involved in the misuse of healthcare policies are the service providers, which includes medical practitioners, hospitals, healthcare organizations and pathology laboratories, insurance subscribers, which includes medical insurance claimants comprising of both medical practitioners and patients, and insurance bearers, comprising healthcare systems run by government and private medical insurance companies [1]. There are numerous sources which provides the raw data, for detecting healthcare frauds, which are generally insurance claims. There are various types of healthcare data that are used by researchers to detect frauds, other than the medical insurance claims, including data about the medical practitioners, medication and drugs data, prescriptions, data of bills and transactions [2]. HCFA (Health Care Financing Administration), an eminent health department of U.S. government, provides two healthcare programs, namely Medicare and Medicaid. Many of the researchers, to detect frauds and abuse in the healthcare systems, uses the data released by Medicare or Medicaid which involves data of drugs, billing transactions and medical practitioners. To detect fraud by a person associated with the healthcare system, behavioural observation or profiling approaches based on data mining techniques are used by configuring behavioural patterns and inspecting it to find out existing deviations from the usual patterns [3]. Researchers have divided data mining approaches into two categories: supervised and unsupervised learning [4], [5]. However, some cases require another approach of data mining called as semi-supervised learning [1], [6]. Recently, many researchers use the healthcare data publicized by The Centers for Medicare

and Medicaid Services (CMS) for detecting frauds in the healthcare system [3], [7], [8]-[16]. Up till now, CMS has released the data only for the years 2012, 2013, 2014 and 2015. For that reason, there is still scope for addition of more future work to detect misuse in medical insurance systems as all the research done using this data is in the prelusive phases.

An anomaly detection approach, proffered by Arunasalam et al. [14], detects anomalies in the insurance claims data obtained from Medicare data by using Rule-based Data Mining. This approach uses unsupervised technique to develop applications that analyses medical insurance claims applying big data to detect anomalies and aid health insurers recognize hidden upcoding frauds which can't be detected by the transaction handling systems. Reeder et al. [9] proposed an approach that applies supervised techniques and graph algorithms, to the graph obtained from the datasets, released by Medicare and Medicaid to determine possible threats in healthcare system. Chawla et al. [15] used CMS data, released for 2012 and examined the past schooling of medical practitioners to find out the way a medical practitioner practices. They attempted to ascertain possible anomalies by comparing procedures, payments, and medical school fees predicated on a geo-location survey done with the overall dispensation of school procedure fees and payments nation-wide. By discovering the correlations between the medical practitioners' education backgrounds and the procedures performed by them, the proposed approach identify the medical practitioners who are possibly misusing medical insurance systems. Feng et al. [16] also used CMS 2012 data and while considering a single medical field, Urology, they analysed variability among the practitioners of Urology based on their usage of service and payment by using Linear Regression to predicted savings from a standard model of usage of service. Herland et al. [8] used multinomial Naïve Bayes algorithm in their proposed machine learning models which detects anomalies in the CMS 2013 data by predicting different fields of medical practitioners and classifying practitioners into their respective medical fields by analysing the medical procedures performed and billed by them. The machine learning model is evaluated with 5-fold cross validation by using F-Score computed for each field to identify when the medical practitioners shows unconventional behaviour in their medical insurance claims.

In this research paper, we proffered a fraud detection approach which uses Link Analysis algorithm to detect anomalies from 2014 CMS dataset [17]. The anomalies are the physicians who acts suspiciously different from the conventions of their own specialty, that is, who performs medical procedures other than their own specialty. We also executed the Multinomial Naïve Bayes algorithm, used in the previous study [8], to analyse and compare the performance of both the algorithms.

Material and Methods

In this research, we have proposed a fraud detection approach, which is based on Link Analysis algorithm, to detect the medical practitioners who have used medical procedures other than their own medical field and claimed insurance for the same procedures, which could possibly indicate fraud or misuse of medical insurance systems. Bauder et al. [8] used Multinomial Naïve Bayes to detect anomalous medical practitioners. Our aim is to take further the work done by Bauder et al. [8], by executing our fraud detection approach on a similar dataset, and analyse the performance of both the approaches. To achieve this, we classified the dataset into different classes of the specialties or the Provider Types. Then the F-score is computed for each of the Provider Types. The F-Score indicates the ability of the algorithm to detect the suspicious medical practitioners or providers in a selected Provider Type. It indicates that it is simpler to detect anomalous medical practitioners or providers in the Provider Types with F-Score over 90% than to detect them in the provider types with F-Score below 90%. The research done in [8] detects aberrant providers in the provider types used a simple learner, Multinomial Naïve Bayes classifier and deduced that the provider types with less F-Score might require more focused work. One of the reasons could be the overlapping between the procedures performed across several provider types with less F-Score. Another reason could be the high number of instances per provider in the provider types with F-Score between 0.5 and 0.9. Hence, more advanced data mining techniques can be used to get better results. We have executed our proposed fraud detection approach on the provider types of all the ranges of F-Score.

CMS Dataset:

We have used the publicly available CMS 2014 dataset [17] which is publicized by The Centers for Medicare and Medicaid Services (CMS). This dataset contains healthcare insurance claims data of medical practitioners or providers for the treatment they provided to their patients. In this dataset, for each medical practitioner or provider, there are multiple records where each record represents insurance claims for a specific medical procedure performed by the medical provider. Every record consists of medical provider's information including the medical provider's name, National Provider Identifier, Provider Type or specialty, and place of services, along with the Healthcare Common Procedure Coding System (HCPCS) code for the medical procedure performed, description of the HCPCS code, the No. of times the procedure was provided, No. of Medicare Beneficiaries, the Avg. Medicare Allowed Amount, Avg. Medicare Payment Amount, etc.

HCPCS code is a set of codes which denotes the medical procedures or treatments performed by the medical practitioners on their patients. National Provider Identifier is a unique identifier representing a medical practitioner or provider. For this research, we have extracted the following features from the dataset: providers' Provider Type, National Provider Identifier, Name, HCPCS code of the medical procedures performed, and the number of times the medical procedure was performed i.e. number of services. A sample of the records present in the dataset with the extracted attributes is shown in the following table:

National Provider Identifier	Name of Provider	Provider Type of the provider	HCPCS code	No. of services
1003000126	ENKESHAFI ARDALAN	Internal Medicine	99222	357

1003000126	ENKESHAFI ARDALAN	Internal Medicine	99223	98
1003000126	ENKESHAFI ARDALAN	Internal Medicine	99231	104
1003000126	ENKESHAFI ARDALAN	Internal Medicine	99232	1418
1003000126	ENKESHAFI ARDALAN	Internal Medicine	99233	175
1003000126	ENKESHAFI ARDALAN	Internal Medicine	99238	330
1003000126	ENKESHAFI ARDALAN	Internal Medicine	99239	223
1003000126	ENKESHAFI ARDALAN	Internal Medicine	99291	23

TABLE I SAMPLE OF RECORD FROM THE DATASET

Table I represents an extract of the records in the dataset indicating the information for the National Identifier Provider 1003000126. This extract of record shows that the provider 'ENKESHAFI ARDALAN' belongs to Provider Type 'Internal Medicine' and has performed multiple medical procedures signified by the HCPCS codes. The record also shows that how many times each of the procedure is performed by the provider indicated by No. of services.

Link Analysis Algorithm:

Link Analysis is a data mining technique which is used to analyse data by determining correlations between the nodes of a graph derived from the data, where nodes indicates objects in the data. PageRank algorithm [19], is one such Link Analysis algorithm, which was first used by Google to rank websites in their search-engine outcomes. In this algorithm, the World Wide Web is considered as a graph, where websites are nodes and links are edges, and the "importance" of the nodes in the graph is measured. This rank correlates to the probability that a "random walker" visits the node. The walker visits from node to node in the following way: with probability *d*, *called damping factor*, the walker selects a random outgoing edge and with probability *1 - d* the walker moves to a random node. The value of damping factor is usually taken as 0.85. The PageRank scores, or the probabilities, are measured by using the following equation 1:

$$p_{rank}(n_i) = \frac{1-d}{T} + d \sum_{n_j \, links\ to\ n_i} \frac{p_{rank}(n_j)}{c(n_j)} \qquad (1)$$

The above equation computes the PageRank score or probability p_{rank} for the random walker to visit node n_i by first considering the damping factor d, then by adding up the probabilities of being adjacent node n_i, and multiplied by the probability of the next edge (n_i, n_i). In the above equation, n_i represents the i^{th} node, n_j represents the j^{th} node, $c(n_i)$ represents the out-degree on n_i, and T is the count of all the nodes present in the graph. PageRank algorithm comes in various forms. Personalized PageRank algorithm [20] is a well-known and significant form of PageRank algorithm. Instead of starting from the random node, in Personalized PageRank algorithm, the walker starts from the selected nodes with higher priority. Personalized PageRank algorithm is a modified version of the PageRank algorithm which computes PageRank scores subjective to a specific area or matter of interest.

Proposed Approach

The proposed approach detects medical practitioners or providers who have performed one or more medical procedure other than their specialty or Provider Type. These medical providers who have used medical procedures other than the providers of the same Provider Type can be suspected as providers showing aberrant behaviour. These suspicious medical providers can be possibly termed as fraudsters who could be misusing medical insurance systems.

1. Classification of the Provider Types: In this research, the proposed algorithm attempts to determine which medical provider performs medical procedures other than his own Provider Type that signifies suspicious behaviour of the providers. Therefore, in the first step we classify the Provider Types of the providers listed in the dataset. From the records of every provider given in the dataset, we extract the Provider Type of the providers. There are 90 provider types in the dataset.

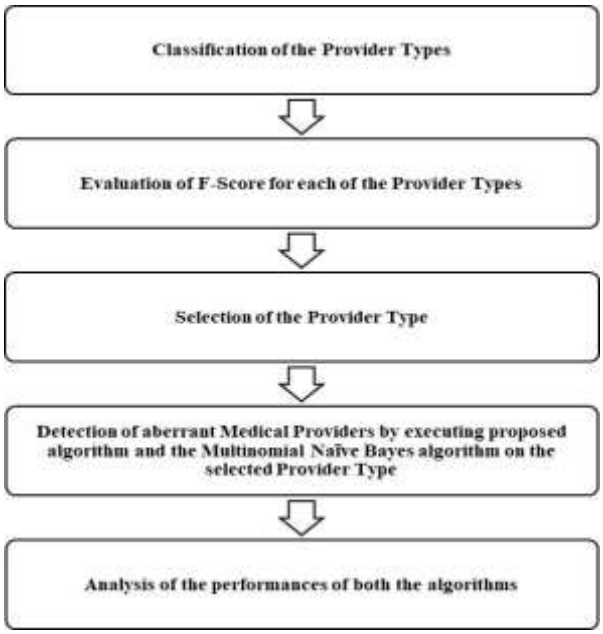


Fig. 1 Proposed Methodology

2. Evaluation of F-Score: After extracting the Provider Types, we find out two other fields which would be required to calculate F-Score for each Provider Type. Those two fields are number of instances i.e. the number of services performed under a Provider Type and number of procedure codes associated to that Provider Type. The F-Score indicates the ability of the algorithm or classifier to detect the suspicious medical practitioners or providers in a selected Provider Type. The F-Score is the weighted mean of Precision and Recall which results in numeric value between 0 and 1. The best value for F-Score is 1 and worst value is 0. Each Provider Type is assumed as a class, and the Provider Type in the consideration is accounted as a positive class while remaining classes or the Provider Types are considered as in the negative class. True positives (t_p) are the providers that are correctly identified to be acting within the norms of their Provider Type. False negatives (f_n) are the providers that could not be classified correctly to be acting within the norms of their

Provider Type. False positives (f_p) are the providers who are actually acting outside the norms of their Provider Type but are wrongly indicated as if they are acting within the norms of their Provider Type. Recall indicates that a given provider is identified correctly in his Provider Type and not in the remaining 89 Provider Types. The Recall is measured as follows:

$$Recall = \frac{(t_p)}{(t_p + f_n)} \quad (2)$$

Precision indicates that a given provider is identified correctly against total number of providers, from the remaining 89 Provider Types. The Precision is measured as follows:

$$Precision = \frac{(t_p)}{(t_p + f_p)} \quad (3)$$

The F-Score is computed as follows:

$$F - Score = \frac{2}{\left(\frac{1}{Recall}\right) + \left(\frac{1}{Precision}\right)} \quad (4)$$

3. Selection of the Provider Type: The provider types are then partitioned into three ranges of F-Score. The partitions of F-Score are in the ranges 0-0.5, 0.5-0.9 and 0.9-1. In the previous research [8], the aberrant providers were detected using a simple learner, Multinomial Naïve Bayes classifier. The Provider Types which has F-Score between 0.9 and 1 i.e. over 90% consists of less or no overlap in the procedures performed than the Provider Types with less F-Score and consists of more number of cases which represents that any simple classifier would produce better results under several circumstances. Hence, the provider types with less F-Score might require more focused work and better data mining techniques can be used to improved results. In this research, we have considered provider types of all the ranges and executed our proposed fraud detection approach on all the ranges. In this stage, we select a range of F-score is selected. Then in that range, we select a provider type in which we want to detect aberrant providers or outliers.
4. Detection of aberrant medical providers in the selected Provider Types: To detect suspicious providers in the selected Provider Type, the algorithm determines those providers who have performed medical procedures that belongs to different Provider Type. These providers can be flagged out as suspicious or fraud providers for further investigation. It can be assumed that there must be similarities in the procedure patterns of the providers belonging to the same Provider Type. The providers who infringe this assumption can be possibly considered as fraudulent providers. To achieve this, we have proposed a novel algorithm based on Personalized PageRank algorithm.

The proposed algorithm works as follows:

(a) Generate a graph with medical providers as nodes and the similar procedure codes as the edges between the nodes. If two providers use similar procedure code, then there is an edge between those two provider nodes.
(b) Then we select a Provider Type. For that specific Provider Type, we then set source nodes or providers belonging to that Provider Type.
(c) The algorithm starts with selecting a random node in the set of source nodes and follows an edge visiting the next node.
(d) Compute the score for that node by using Equation 1.
(e) The nodes whose score is high but with different provider type than the initially selected one are considered as the aberrant providers indicating that these providers could have claimed for medical procedures that are prescribed by providers of different Provider Type.
(f) These providers are flagged out as outliers for further investigation.

We also executed the Multinomial Naïve Bayes algorithm, which was used by Bauder et al. [8] in their research, on the same dataset, to evaluate and analyse the performance of the proposed algorithm against the Multinomial Naïve Bayes algorithm. By finding the posterior probabilities of class membership based on each feature value, which is learned from a set of labelled training events, the Multinomial Naïve Bayes learner classifies new events.

Analysis of performance of both the algorithms: After applying the algorithms, on a selected Provider Type, we identify providers who stands as outliers in their Provider Type showing suspicious and aberrant behaviour. We get two different list of providers including outliers determined by both the algorithms. We analyse the performance of both the algorithms based on the number of outliers detected by the algorithms, time required to detect outliers and ability to handle missing values in the procedure codes data.

Results and Discussion

This experiment attempts to identify those medical practitioners or providers who could be suspected of misusing medical insurance systems or involved in a healthcare fraud. To find out such fraudulent medical providers, we have attempted to take further the research done in [8] by implementing our proposed algorithm on a similar dataset and analyse the results and performance of our proposed fraud detection algorithm and the Multinomial Naïve Bayes algorithm used in the previous research. We have implemented our research on .NET framework using C#. The implemented model starts with loading and reading the CMS dataset. The screenshot of the dataset is shown in Fig. 2. Then the model classifies the provider types of the providers. There are 90 provider types in the dataset. For each provider type, the prerequisites for evaluating the F-Score, number of instances and the number of codes are computed. The screenshot of the classified provider types and the prerequisites of F-Score for each of the provider types is shown in Fig. 3.

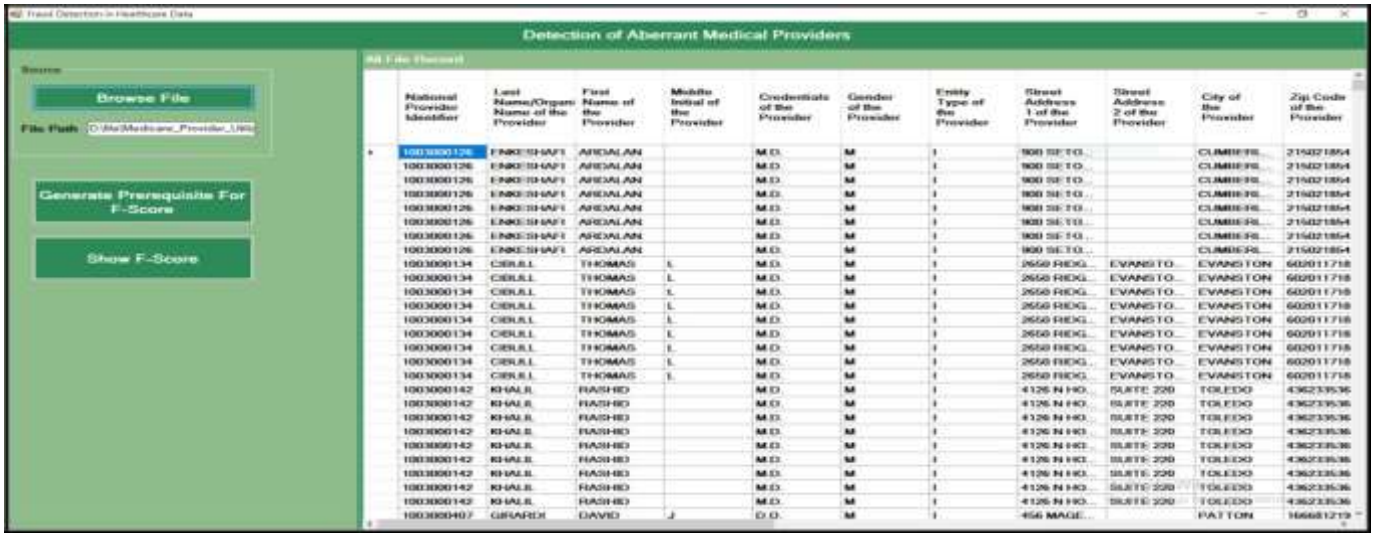
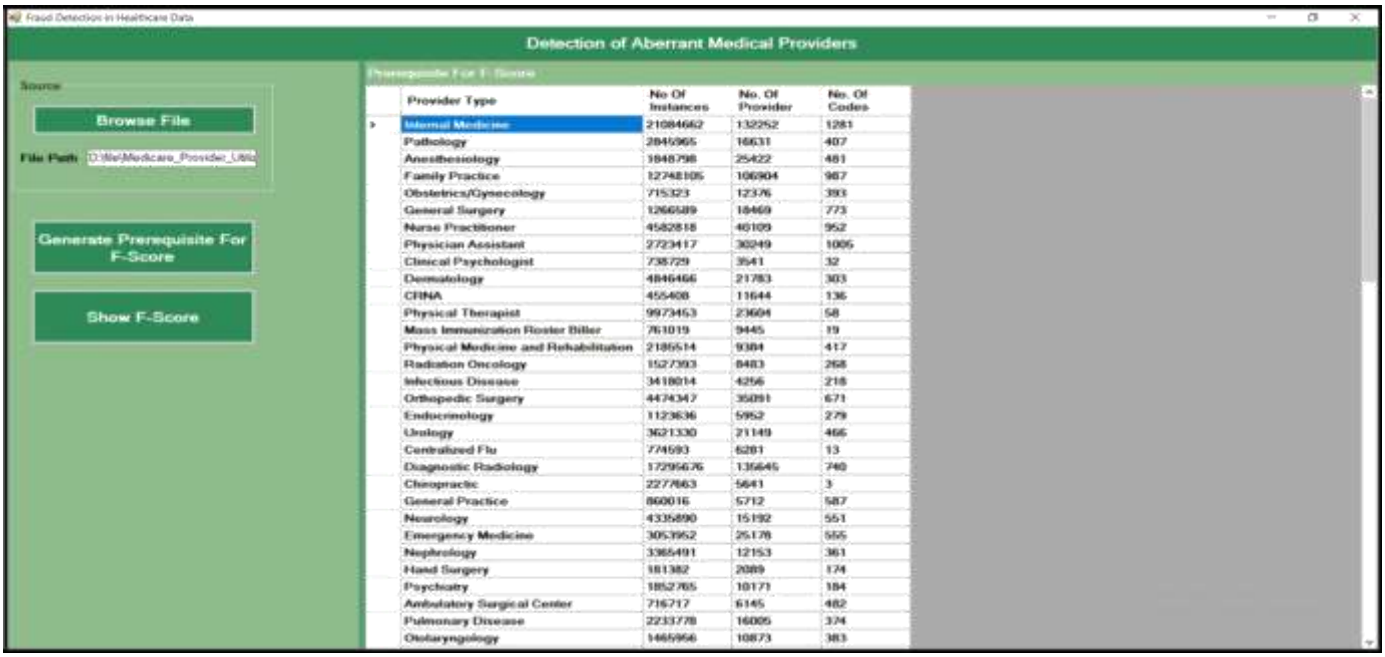
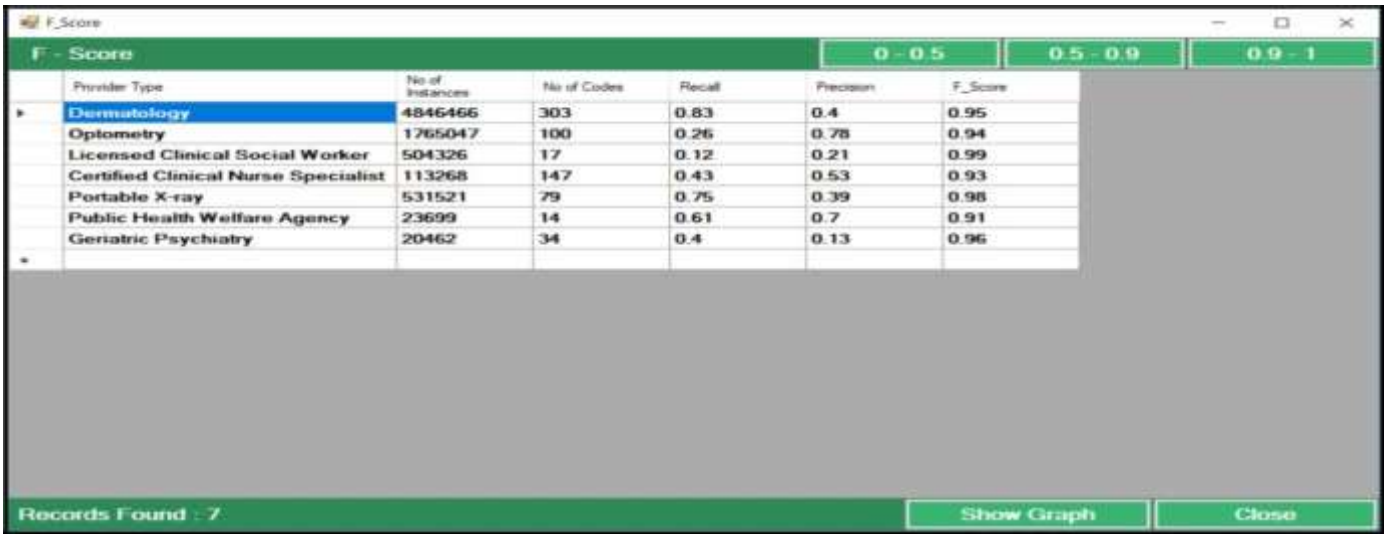


Fig 2: Screenshot of the CMS Dataset



Provider Type	No Of Instances	No. Of Provider	No. Of Codes
Internal Medicine	21084662	132252	1281
Pathology	2845965	16631	407
Anesthesiology	1848798	25422	481
Family Practice	12748105	106904	987
Obstetrics/Gynecology	715323	12376	393
General Surgery	1266589	18469	773
Nurse Practitioner	4582818	40109	952
Physician Assistant	2723417	30249	1005
Clinical Psychologist	738729	3541	32
Dermatology	4846466	21783	303
CRNA	455408	11644	136
Physical Therapist	9973453	23604	58
Mass Immunization Roster Biller	761019	9445	19
Physical Medicine and Rehabilitation	2186514	9384	417
Radiation Oncology	1527393	8483	268
Infectious Disease	3418014	4256	218
Orthopedic Surgery	4474347	35091	671
Endocrinology	1123636	5952	279
Urology	3621330	21149	466
Centralized Flu	774693	6281	13
Diagnostic Radiology	17295676	135645	740
Chiropractic	2277863	5641	3
General Practice	860016	5712	587
Neurology	4336890	15192	551
Emergency Medicine	3053952	25178	555
Nephrology	3365491	12153	361
Hand Surgery	181382	2089	174
Psychiatry	1852765	10171	184
Ambulatory Surgical Center	716717	6145	482
Pulmonary Disease	2233778	16005	374
Otolaryngology	1465956	10873	383

Fig. 3 Screenshot of the Classified Provider Types along with the Prerequites of F-Score



Provider Type	No of Instances	No of Codes	Recall	Precision	F_Score
Dermatology	4846466	303	0.83	0.4	0.95
Optometry	1765047	100	0.26	0.78	0.94
Licensed Clinical Social Worker	504326	17	0.12	0.21	0.99
Certified Clinical Nurse Specialist	113268	147	0.43	0.53	0.93
Portable X-ray	531521	79	0.75	0.39	0.98
Public Health Welfare Agency	23699	14	0.61	0.7	0.91
Geriatric Psychiatry	20462	34	0.4	0.13	0.96

Fig. 4: Screenshot of the Provider Types with F-Score Range 0.9-1

In the next stage, the F-Score for each of the provider type is computed. Then the provider types are partitioned into three ranges as discussed earlier. In this research, we are focussing on the provider types with F-Score of all the three ranges. The F-Score range 0.9-1 consists of 7, the range 0.5-0.9 consists of 42 and the range 0-0.5 consists of 41 provider types out of the 90 provider types. Fig. 4 shows the snapshot of the provider types with F-Score over 90% (i.e. between 0.9 and 1) along with the recall and precision values. We then selected provider types to find out outliers or aberrant providers in it from that range. The providers who have performed procedures of other provider types are considered as the aberrant providers acting outside the norms of their field indicating misuse or fraud in medical insurance systems. The outliers are those who have procedure codes of other field or provider type in their records. Since, originally, the dataset consisted of genuine medical providers, we made necessary changes in the dataset by altering the records and swapping the procedure codes of some providers to make them the outliers. For this experiment, we have named our proposed algorithm as 'Ranker' and Multinomial Naïve Bayes algorithm as 'Naïve Bayes'. The outliers in the result shown in Fig. 5 are highlighted in light green. For example, we selected the provider type 'Dermatology'. After applying both the algorithms, on the provider type 'dermatology', we got the list of providers in dermatology along with the outliers as shown in the Fig. 5. The outliers are those who have used procedure codes of other provider types. The results exhibit that the Naïve Bayes could generate 46 outliers while Ranker algorithm could correctly detect 56 outliers. After examining the records of the providers, we affirmed that there are 56 outliers or aberrant providers in the provider type 'Dermatology'. This indicates that our proposed algorithm (Ranker) can detect more correct outliers than the Naïve Bayes algorithm represented in the graph shown in Fig. 6.

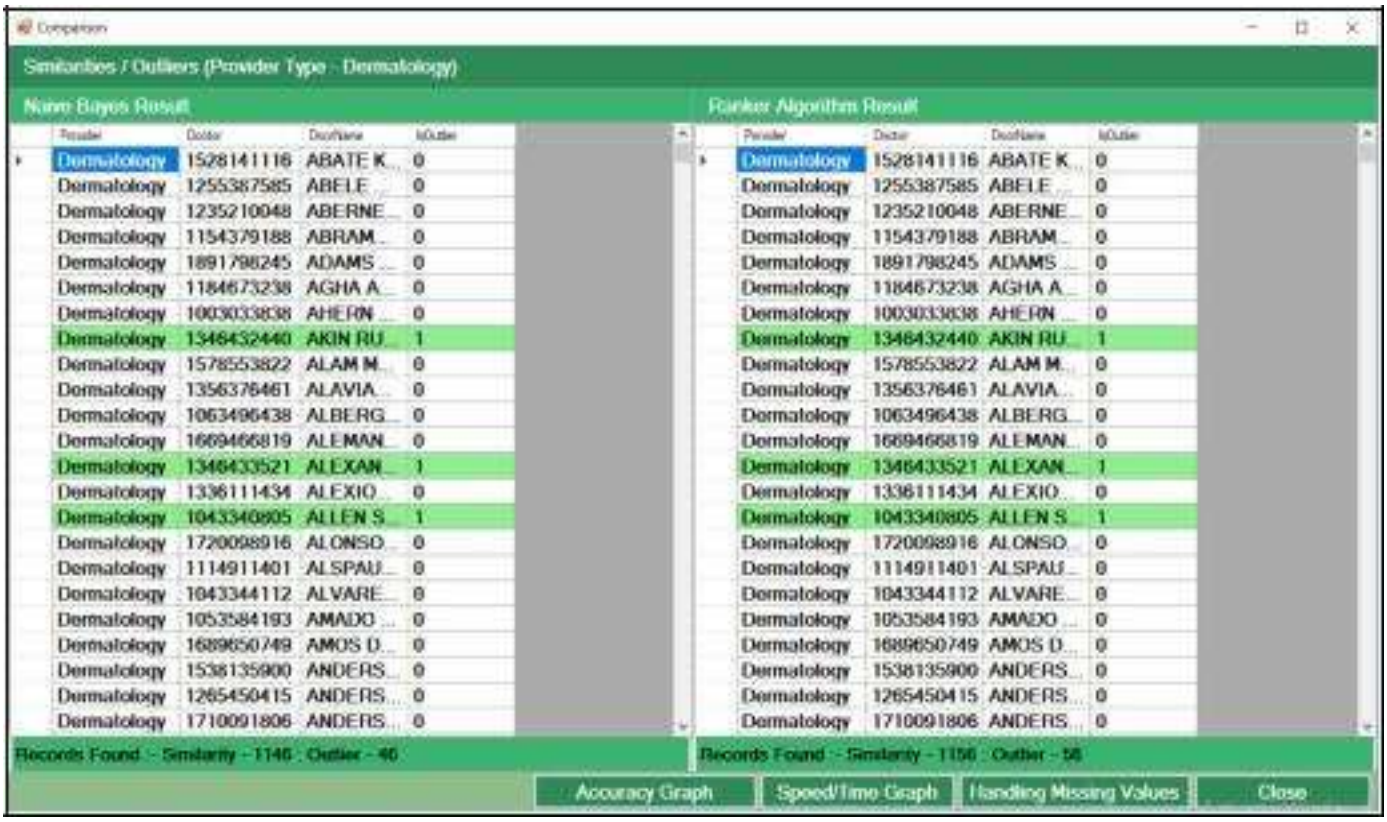


Fig. 5: Screenshot of the List of Providers of the Provider Type 'Dermatology' along with the Outliers (Highlighted)

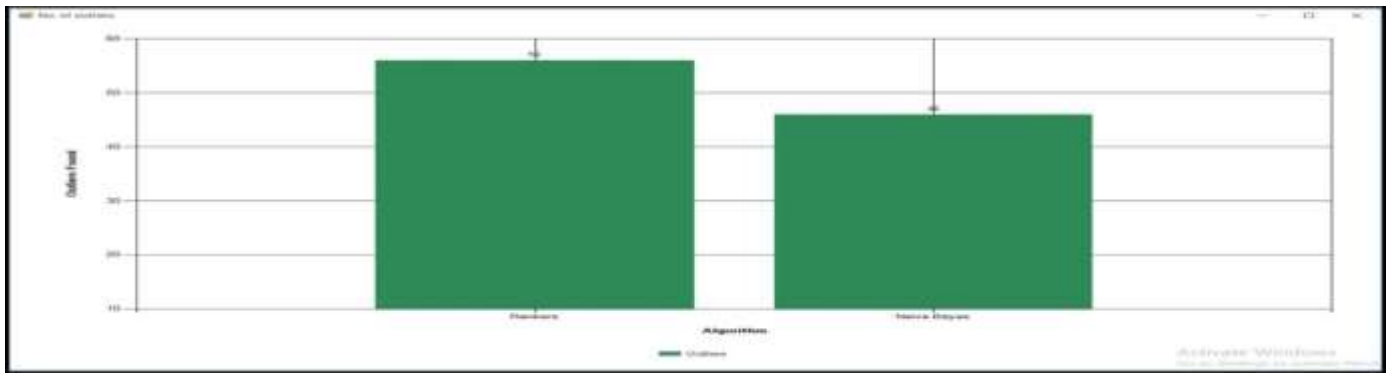


Fig. 6: Screenshot of the Graph Representing the No. of Outliers Detected by Ranker and Naïve Bayes Algorithm in the Provider Type 'Dermatology'

We also analysed that Naïve Bayes algorithm took more time to generate results than the proposed Ranker algorithm. Fig. 7 shows the representation of time comparison of both the algorithms for the provider type 'Dermatology'. Fig. 7 is the screenshot of the graph representing the amount of time taken (in milliseconds) by both the algorithms to detect outliers in the provider type 'Dermatology'. We also analysed the performance of both the algorithms based on their ability of handling missing values in the data. From this analysis, we learned that proposed algorithm (Ranker) handles missing values better than the naïve Bayes algorithm by deleting, ignoring, imputing or substituting data values wherever necessary. The comparison of both the algorithm based on ability of handling missing values in the provider type 'Dermatology' is represented in Fig. 8. Fig. 8 shows the graph comparing the number of missing values handled by both the algorithm.

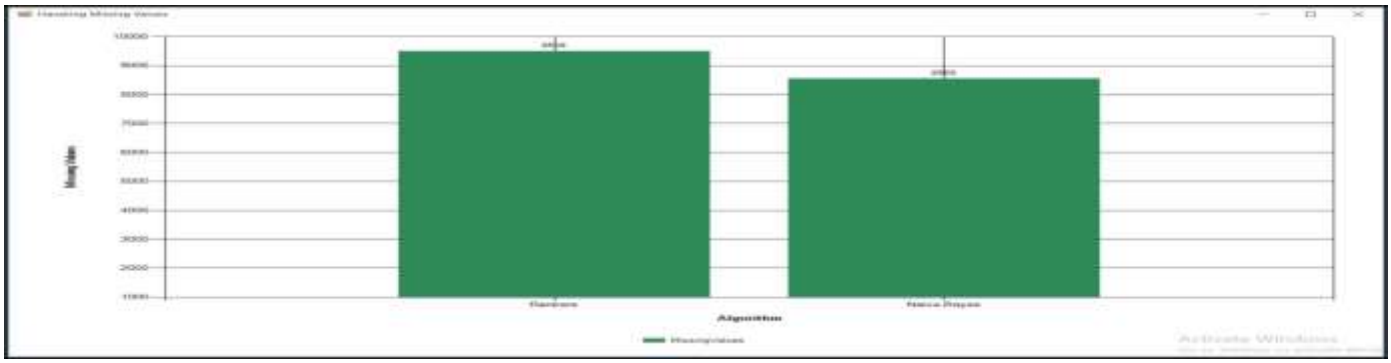


Fig. 7: Screenshot of the Graph Representing the Time Taken (in milliseconds) by the Ranker and Naïve Bayes Algorithm to Detect Outliers in the Provider Type 'Dermatology'

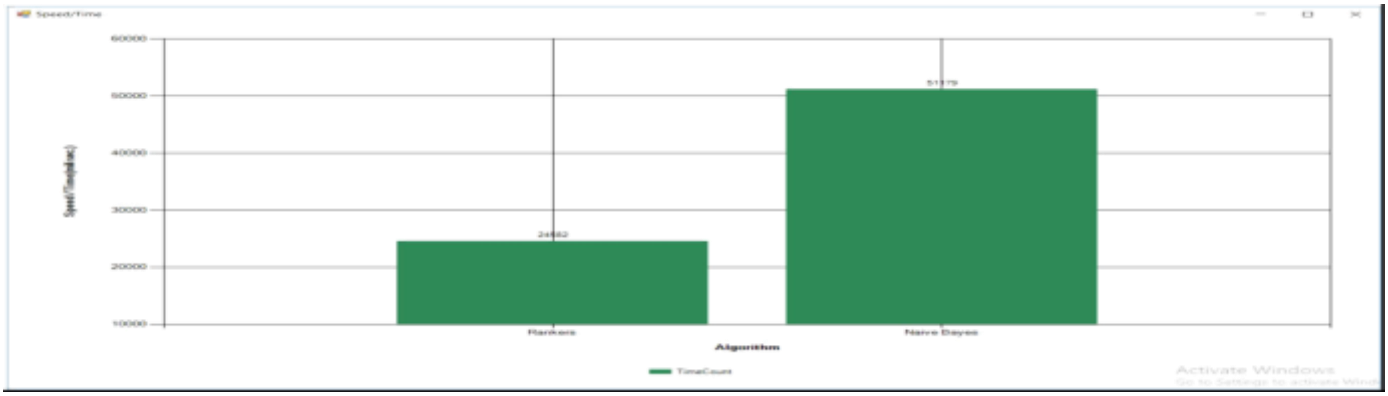


Fig. 8: Screenshot of the Graph Representing the No. of Missing Values Handled by the Ranker and Naïve Bayes Algorithm to Detect Outliers in the Provider Type 'Dermatology'

As discussed earlier, we also executed both the algorithms on the other two ranges of provider types, i.e., 0.5-0.9 and 0-0.5. For example, we selected the range 0.5-0.9 as shown in Fig. 9. We then selected the provider type 'Anesthesiology'.

Fig. 10 shows a screenshot of the results of both the algorithms identifying the outliers or aberrant providers. The results in Fig. 10 shows that the proposed algorithm could identify more outliers than the naïve bayes algorithm.

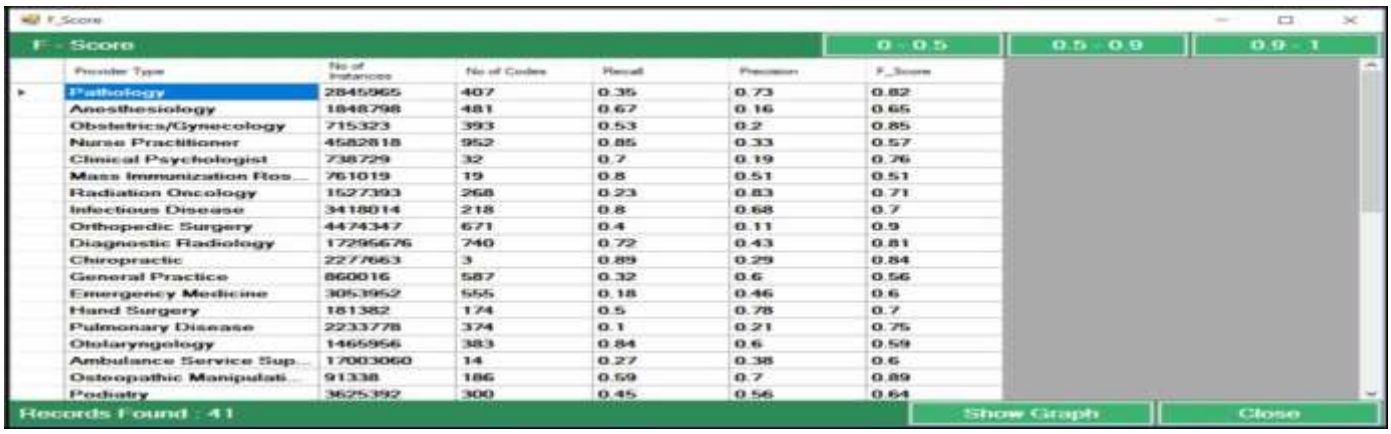


Fig. 9: Screenshot of the Provider Types with F-Score Range 0.5-0.9

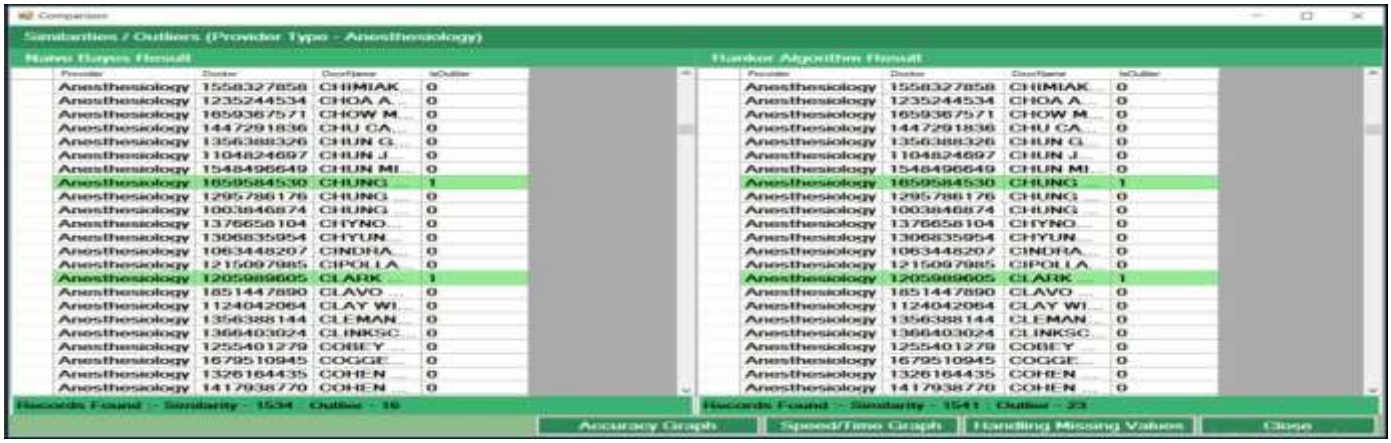


Fig. 10: Screenshot of the List of Providers of the Provider Type 'Anesthiology' along with the Outliers (Highlighted)

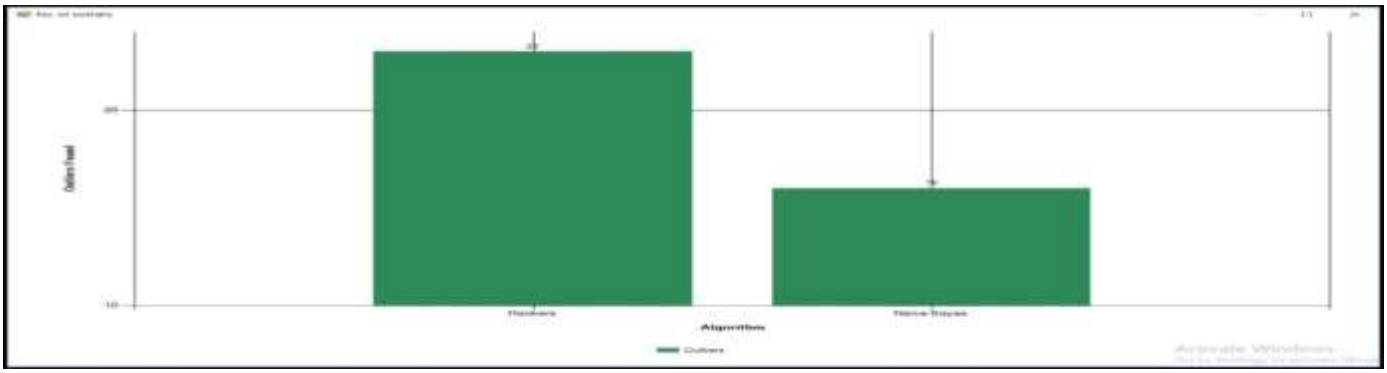


Fig. 11: Screenshot of the Graph Representing the No. of Outliers Detected by Ranker and Naïve Bayes Algorithm in the Provider Type 'Anesthiology'

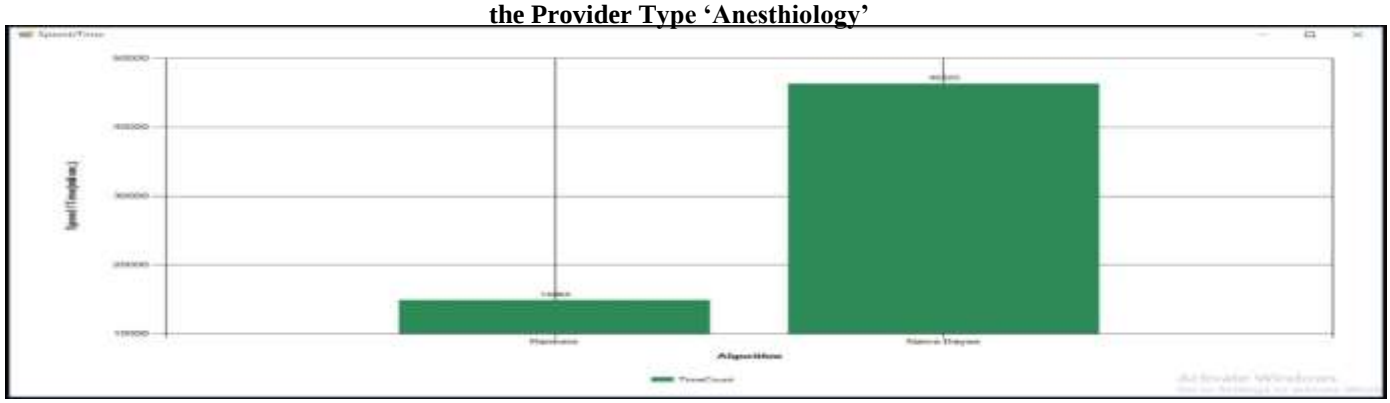


Fig. 12: Screenshot of the Graph Representing the Time Taken (in milliseconds) by the Ranker and Naïve Bayes Algorithm to Detect Outliers in the Provider Type 'Anesthiology'

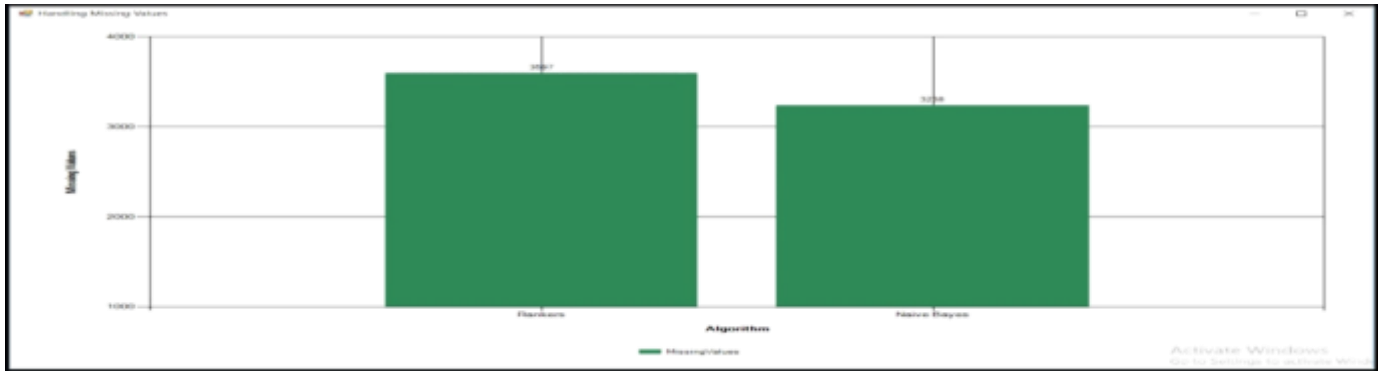


Fig. 13: Screenshot of the Graph Representing the No. of Missing Values Handled by the Ranker and Naïve Bayes Algorithm to Detect Outliers in the Provider Type 'Anesthiology'

The comparison of both the algorithms based on the number of outliers detected is represented by the graph shown in Fig.11. We analysed the performance of both the algorithms based on the speed of the algorithms and the ability to handle the missing values in the data. Fig. 12 and Fig. 13 shows the graphs representing the time taken by both the algorithms, to generate outlier results and their ability to handle missing values in the Provider type 'Anesthesiology' respectively. Fig. 12 and Fig. 13 clearly indicates that the Ranker takes less time to generate results and handles more missing values than the multinomial naïve bayes algorithm.

Similarly, we executed both the algorithms on all the provider types. Most of the results inferred that the proposed fraud detection algorithm tends to identify more outliers in less time as compared to the Multinomial Naïve Bayes algorithm while handling more missing values. Thus, the results and analysis deduced that the proposed fraud detection algorithm detects fraud or misuse of medical insurance systems by identifying the medical practitioners who have used medical procedures other than their own medical field and claimed insurance for the same procedures. It also deduced that the proposed fraud detection algorithm shows better performance and generate improved results than the Multinomial Naïve Bayes algorithm.

Conclusion

In this paper, we proposed a fraud detection approach that attempts to identify aberrant medical practitioners who have performed medical procedures that are different from the other practitioners of their field possibly indicating fraud or misuse of medical insurance systems. The proposed algorithm works on the concept of Link Analysis algorithm. We tried to analyse the performance of the proposed fraud detection algorithm against the Multinomial Naïve Bayes algorithm used in the previous study. By analysing the results of all the provider types, we conclude that, in terms of performance and accuracy, our proposed fraud detection algorithm generated better results than the Multinomial Naïve Bayes algorithm. The proposed fraud detection algorithm could detect more number of outliers or aberrant medical practitioners in less time while handling more missing values in the data than the Multinomial Naïve Bayes algorithm. A possible drawback of Naive-Bayes could be that the frequency-based probability estimate will be zero if there are no occurrences of a class label and an attribute value together. Thus, the posterior probability estimate gets affected when the probabilities are multiplied gives a zero, given a conditional independence assumption. Another drawback could be that it makes a firm assumption on the form of data distribution. It assumes that any two features are independent, which is not always the case. In this research, we executed and analysed the proposed fraud detection algorithm on the CMS dataset, which contains healthcare insurance claims data of medical practitioners for the treatment they provided to their patients. The proposed fraud detection approach considered only the attributes of medical providers' Name, National Provider Identifier, Provider Type, HCPCS code of the procedures performed and the number of services. The CMS dataset provides other attributes such as Number of Medicare Beneficiaries, the Average Medicare Allowed Amount, Average Medicare Payment Amount, etc. which can be focussed and worked upon to detect more suspicious providers based on their insurance claims' payment and benefits.

Acknowledgment

We would like to thank The Centers for Medicare and Medicaid Services (CMS) for permits to use and revise the dataset provided by CMS.gov. Any recommendations, results, conclusions, or views presented in this paper are those of the author(s) and do not express the views of the The Centers for Medicare and Medicaid Services.

References:

[2] D., Aydoğan, and Ş. Sağıroğlu, "*Health care fraud detection methods and new approaches",* Computer Science and Engineering (UBMK), 2017 International Conference on. IEEE, 2017.

[3] Liu, Q., Vasarhelyi, "*Healthcare fraud detection: A survey and a clustering model incorporating Geo-location information"*, In29th World Continuous Auditing and Reporting Symposium (29WCARS), Brisbane, Australia 2013.

[4] Jyothsna, V., VV Prasad, and K. Munivara Prasad, *"A review of anomaly based intrusion detection systems."*, International Journal of Computer Applications 28.7 (2011): 26-35.

[5] Li, Jing, Huang et al, "*A survey on statistical methods for health care fraud detection."*, Health care management science 11.3 (2008): 275-287.

[6] Joudaki, H, Rashidian et al., *"Using data mining to detect health care fraud and abuse: a review of literature."*, Global journal of health science 7.1 (2015): 194.

[7] A, Aisha, MA Maarof, and A Zainal, *"Fraud detection system: A survey."*, Journal of Network and Computer Applications 68 (2016): 90-113.

[8] Thornton D, RM. Mueller, P. Schoutsen, and JV., *"Predicting healthcare fraud in medicaid: a multidimensional data model and analysis techniques for fraud detection."*, Procedia technology 9 (2013): 1252-1264.

[9] Bauder Richard, A., Taghi M. K, Aaron R. et al., *"Predicting medical provider specialties to detect anomalous insurance claims."*, Tools with Artificial Intelligence (ICTAI), 2016 IEEE 28th International Conference on. IEEE, 2016.

[10] Branting, LK, Flo R., Jeffery G., T. Champney, et al. *"Graph analytics for healthcare fraud risk estimation."* Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on. IEEE, 2016.

[11] RM Musal, *"Two models to investigate Medicare fraud within unsupervised databases."* , Expert Systems with Applications 37.12 (2010): 8628-8633.

[12] Bauder RA., and Taghi M. K., R Aaron, *"A probabilistic programming approach for outlier detection in healthcare claims."* , Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on. IEEE, 2016.

[12] Bauder R.A., and Taghi M. K., Aaron.R, *"A novel method for fraudulent medicare claims detection from*

expected payment deviations (application paper)." , Information Reuse and Integration (IRI), 2016 IEEE 17th International Conference on. IEEE, 2016.

[13] G Capelleveen, P. Mannes, et al., *"Outlier detection in healthcare fraud: A case study in the Medicaid dental domain.",* International journal of accounting information systems 21 (2016): 18-31.

[14] Uma S., and B. Arunasalam., *"Leveraging big data analytics to reduce healthcare costs.",* IT professional15.6 (2013): 21-28.

[15] Keith F, and Nitesh C., *"Does medical school training relate to practice?* Evidence from big data.", Big data3.2 (2015): 103-113.

[16] JS Ko, Chalfin., Trock, Feng, et al., *"Variability in Medicare utilization and payment among urologists."* , Urology 85.5 (2015): 1045-1051.

[17] https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier.html;Medicare_Provider_Utilization_and_Payment_Data__
hysician_and_Other_Supplier_PUF_CY2014.csv

[18] M Herland, R. Bauder, and Taghi M. K., *"Medical provider specialty predictions for the detection of anomalous medicare insurance claims.",* Information Reuse and Integration (IRI), 2017 IEEE International Conference on. IEEE, 2017.

[19] Sergey B., Lawrence, Page., *"The anatomy of a large-scale hypertextual web search engine.",* Computer networks and ISDN systems 30.1-7 (1998): 107-117.

[20] Taher H., Hawelivala, *"Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search.",* IEEE transactions on knowledge and data engineering 15.4 (2003): 784-796.

