Applied GIS

ISSN: 1832-5505
Vol-10, Issue-01, 2022

IDENTIFYING HEALTH INSURANCE CLAIM FRAUDS USING MIXTURE OF CLINICAL CONCEPTS

Dr. J. Thilagavathi, D. Raja, Mrs. M. Indra Priya, Ms. R. Latha Priyadharshini
Professor [1,2], Associate Professor [3], Assistant Professor [4]
thilagavathi@actechnology.in, raja.d@actechnology.in, indrapriya.m@actechnology.in,
lathapriyadharshini.r@actechnology.in
Department of CSE, Arjun College of Technology, Thamaraikulam, Coimbatore-Pollachi Highway,
Coimbatore, Tamilnadu-642 120

ABSTRACT

In order to pay for the expensive medical treatment, patients rely on health insurance offered by either the public or private sectors, or both. Some medical professionals perpetrate insurance fraud because their patients rely on them. Even if there aren't many of them, insurance companies reportedly lose billions of dollars annually as a result of fraud. This work presents a formulation of the fraud detection issue over definite claim data consisting of medical operation and diagnostic codes. The data set is modest in size. By converting procedure and diagnostic codes into Mixtures of Clinical Codes (MCC), our innovative representation learning technique allows us to identify fraudulent claims. We further explore potential expansions of MCC using Long Short Term Memory networks and Robust Principal Component Analysis. Our experimental results demonstrate promising outcomes in identifying fraudulent records.

INTRODUCTION

The use of data analytics has grown in importance across all sectors of economic growth. Health records, clinical data, medications, insurance claims, provider information, and patient information "potentially" provide data analysts tremendous opportunity, given that healthcare is one of the biggest financial sectors in the US economy. In the US, healthcare costs more than $3 trillion per year, and insurance companies handle billions of claims annually [1]. A typical healthcare reconciliation procedure is shown in Figure 1 as a short flowchart by by using several entities. Prior to providing any treatment, the provider's office checks the patient's insurance or other funding sources to make sure there is sufficient coverage. After then, the supplier of the service uses the results of the first exams to make a diagnosis. After then, the doctor or other medical professional will do testing on the patient, which may include more diagnostics or perhaps surgery. Typically, the patient's report will include these diagnoses and procedures as well as additional data such as demographics, personal information, and information about previous and current visits. Usually, the patient pays the copay that is specified in their insurance plan and then checks out. After then, a medical coder receives the patient's report, uses it to compile all of the provider's information into a "superbill," The healthcare sector is a major moneymaker, so it's not surprising that some people file false or misleading claims to insurance. "An intentional deception or misrepresentation made by a person, or an entity, with the intent to defraud or mislead in healthcare transactions" is how the National Health Care Anti-Fraud Association (NHCAA) describes healthcare fraud. being aware that he or other parties might gain an unauthorised advantage from the deceit [3]. Even if they only make up a tiny proportion, such false assertions have a hefty price tag. Losses due to fraud in the US are estimated by NHCAA to be in the tens of billions of dollars [3]. Studies reveal that only a tiny fraction of the losses is recovered each year, despite the fact that healthcare businesses have stringent procedures pertaining to fraud and abuse control [4]. Most frequent fraudulent practices perpetrated by dishonest practitioners in the healthcare arena include the following. _ Using erroneous diagnosis to justify unnecessary medical treatments. _ "Upcoding" refers to the practice of billing for expensive services or procedures rather than the actual procedures themselves. creating assertions for processes that were never carried out. Claiming insurance funds by doing operations that are not medically required. _ "Unbundling" refers to the practice of billing for individual steps in a process rather than as a whole. Falsely claiming that services that aren't provided are essential to be reimbursed by insurance, particularly for operations that enhance one's appearance.

If you want to fix any or all of the problems mentioned above, you can't only rely on your domain expertise. In order to effectively control fraudulent activity, domain experts may greatly benefit from automated data analytics that can spot fraudulent claims early on. Looking at the issue of healthcare fraud detection from the perspective of health insurance providers is the main topic of this study. When faced with a claim and limited data (i.e., procedure and diagnostic codes), we find a way to determine if a procedure is valid or fraudulent. Various techniques, including data mining [5], classification [6], [7], Bayesian analysis [8], statistical surveys [9], non-parametric methods [10], and expert analysis, have been used to identify the challenge of fraud detection in the medical sector. To construct models for claim status prediction, existing techniques incorporate parameters from a claim database such as doctors' profiles, background history, claim amount, service quality, services provided per provider, and related variables. Despite the effectiveness of these strategies, datasets that are not in the public domain. The answers are also notoriously hard to transfer since the variables used in those datasets are so varied and often incompatible. Because of regulations such as the General Data Protection Regulation (GDPR) in Europe and the Health Insurance Portability and Accountability Act (HIPAA) in the United States, we were only able to access diagnosis and procedure codes for this study. Not to mention that other industries are less reluctant to exchange data than the healthcare business. As an added complication, it is not possible to transfer solutions across software systems since each system reports patient variables differently. Consequently, we limit our issue formulation to procedure and diagnostic codes, which are universally applicable regardless of their place of origin. Based on the assumption that the claim data is a heterogeneous collection of medical ideas, our solution technique uses the International Classification of Diseases (ICD) coding scheme for clinical diagnoses and treatments. Not only that, but the suggested method is compatible with other code styles, such as Current Procedural

Standard Medical Terminology (CPT) and the Healthcare Common Procedure Coding System (HCPCS), or both systems independently, unchanged. Using probabilistic topic modelling, we portray an insurance claim as a Mixture of latent Clinical Concepts (MCC). We are unaware of any previous work that has represented insurance claims as latent space combinations of clinical notions. Every claim is interpreted as a manifestation of some combination of clinical ideas, whether it be pain, mental illness, or infectious disease. Additionally, clinical codes (i.e., codes for diagnoses and procedures) comprise each clinical notion. Clinics, hospitals, and doctor's offices offer the intuitive basis for our concept. Services are often provided to patients in response to particular concerns that arise from one or more diseases. The next step in patient care is the practitioner carrying out the actual operations. Pain, mental illness, infectious illnesses, and their treatments are all examples of clinical concepts that might be used to describe the procedures and diagnosis in a claim. Please take note that we refrain from giving these ideas any specific names or explanations since they are not always clear, complicated, or requiring expertise in the field. We use Robust Principal Component Analysis and Long-Short Term Memory networks to expand the MCC model. Our objective in expanding MCC is to identify and categorise statements as either fraudulent or non-fraudulent based on the important ideas they include. We build upon MCC by training an LSTM network using claim concept weights as a sequence representation. We may use this network to express the claims as LSTM-classifiable sequences of dependent notions. By breaking claims down into sparse vector representations with low ranks, we may use Robust Principal Component Analysis (RPCA) to filter out concepts with substantial weights. Ideal weights devoid of noise are captured by the low-rank matrix. We may summarise our unique contributions to this work as follows. Using a basic set of definite claim data that includes procedure and diagnostic codes, we define the issue of detecting fraudulent claims. As a novel representation learning strategy, we provide clinical ideas as an alternative to procedure and diagnostic codes.

_ Using LSTM and RPCA for classification, we expand the combinations of clinical ideas. When compared to a baseline technique and the Multivariate Outlier Detection (MOD) [11], we find that our methods perform better. To find out-of-the-ordinary provider payments in Medicare claims data, the two-step Multivariate Outlier Detection approach is utilised. The first stage involves creating appropriate residuals by constructing a multivariate regression model using thirteen carefully selected characteristics. After that, a generalised univariate probability model is fed the residuals. In order to find potential outliers in the claim data, they used probabilistic programming approaches in Stan [12]. Our studies, which employ a different issue formulation, make use of the identical CMS (Centres for Medicare and Medicaid Services) dataset as the authors. We use MOD on MCC traits, while their research covers provider and beneficiary data pertaining to Medicare beneficiaries in the state of Florida. In contrast, the baseline classifier uses the training claim data to categorise a test claim as the majority. On the inpatient dataset collected from CMS, our experimental findings demonstrate that MCC + LSTM achieves recall scores of 50%, precision scores of 61%, and accuracy values of 59%. Furthermore, on the outpatient dataset, it shows recall scores of 72 percent, accuracy of 78 percent, and precision of 83 percent. Our hope is that this new study on detecting false claims with little but conclusive evidence will be sparked by the suggested issue formulation, representation learning, and solution. Here is how the remainder of the paper is structured.

EXISTING SYSTEM

Yang and Hwang developed a fraud detection model using the clinical pathways concept and process-mining framework that can detect frauds in the healthcare domain [13]. The method uses a module that works by discovering structural patterns from input positive and negative clinical instances. The most frequent patterns are extracted from every clinical instance using the module. Next, a feature-selection module is used to create a filtered dataset with labeled features. Finally, an inductive model is built on the feature set for evaluating new claims. Their method uses clustering, association analysis, and principal component

ISSN: 1832-5505
Vol-10, Issue-01, 2022

analysis. The technique was applied on a real-world data set collected from National Health Insurance (NHI) program in Taiwan. Although the authors constructed different features to generate patterns for both normal and abusive claims, the significance of those features is not discussed. Bayerstadler et al. [14] presented a predictive model to detect fraud and abuse using manually labeled claims as training data. The method is designed to predict the fraud and abuse score using a probability distribution for new claim invoices. Specifically, the authors proposed a Bayesian network to summarize medical claims' representation patterns using latent variables. In the prediction step, a multinomial variable modeling predicts the probability scores for various fraud events. Additionally, they estimated the model parameters using Markov Chain Monte Carlo (MCMC) [15]. Zhang et al. [16] proposed a Medicare fraud detection framework using the concept of anomaly detection [17]. First part of the proposed method consists of a spatial density based algorithm which is claimed to be more suitable compared to local outlier factors in medical insurance data. The second part of the method uses regression analysis to identify the linear dependencies among different variables. Additionally, the authors mentioned that the method has limited application on new incoming data.

Kose et al. [18] used interactive unsupervised machine learning where expert knowledge is used as an input to the system to identify fraud and abuse related legal cases in healthcare. The authors used a pairwise comparison method of analytic hierarchical process (AHP) to incorporate weights between actors (patients) and attributes. Expectation maximization (EM) is used to cluster similar actors. They had domain experts involved at different levels of the study and produced storyboard based abnormal behavior traits. The proposed framework is evaluated based on the behavior traits found using the storyboard and later used for prescriptions by including all related persons and commodities such as drugs. Bauder and Khoshgoftaar [19] proposed a general outlier detection model using Bayesian inference to screen healthcare claims. In their trials, they used the Stan model, which is comparable to [20]. It should be noted that their focus is only on provider level fraud detection, disregarding any relationships based on clinical codes. Many of those techniques rely on proprietary datasets or combine datasets that do not share common features. This makes direct comparisons between the research very challenging. Healthcare providers and insurance firms are understandably wary about sharing extensive datasets, if they do so at all, due to the severe penalties for breaches of healthcare information privacy and security imposed by laws like HIPAA and GDPR. Because of this, we frame the issue in terms of definite claim data, which consists of minimally detailed codes for diagnoses and procedures. In this context, we address the issue of identifying fraudulent or valid procedures by combining clinical codes with encodings based on RNNs and RPCAs.

Disadvantages

Making false diagnoses to justify procedures that are not medically necessary.Fabricating claims for unperformed procedures. Performing medically unnecessary procedures to claiminsurance payments. Billing for each step of a procedure as if it is a separateprocedure, also called “unbundling”.Misrepresenting non-covered treatments as medicallynecessary to receive insurance payments, especially forcosmetic procedures.

PROPOSED SYSTEM

We use Robust Principal Component Analysis and Long-Short-Term Memory networks to expand the MCC model. Our objective in expanding MCC is to identify and categorise statements as either fraudulent or non-fraudulent based on the important ideas they include. We build upon MCC by training an LSTM network using claim concept weights as a sequence representation. In order to train the LSTM to classify the claims, we may use this network to express them as sequences of dependent ideas. By breaking claims down into sparse vector representations with low ranks, we may use Robust Principal Component Analysis (RPCA) to filter out concepts with substantial weights. Ideal weights devoid of noise are captured by the low-rank matrix. We may summarise our unique contributions to this work as follows. In order to define the issue of detecting fraudulent claims, the system uses a simple

ISSN: 1832-5505
Vol-10, Issue-01, 2022

definitive claim data consisting of procedure and diagnosis codes.

The system introduces clinical concepts over procedure and diagnosis codes as a new representation learning approach.

The system extends the mixtures of clinical concepts using LSTM and RPCA for classification.

Advantages

➢ The proposed system uses Support Vector Machine (SVM) for classification with MCC.

➢ Multivariate Outlier Detection method is an effective method which is used to detect anomalous provider payments within Medicare claims data.

IMPLEMENTATION

Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as

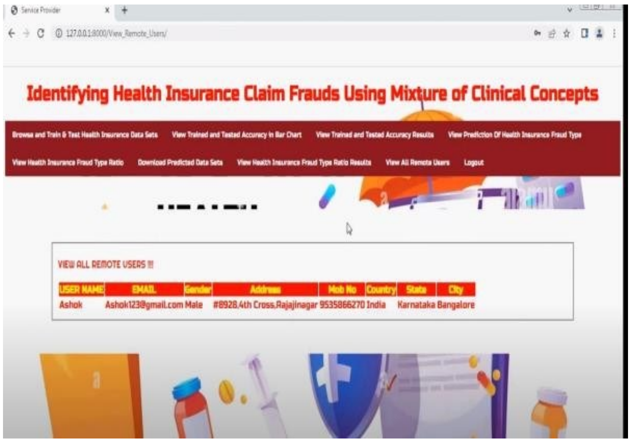
Login, Browse and Train & Test Health Insurance Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View Prediction Of Health Insurance Fraud Type, View Health Insurance Fraud Type Ratio, Download Predicted Data Sets, View Health Insurance Fraud Type Ratio Results, View All Remote Users

View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

Remote User

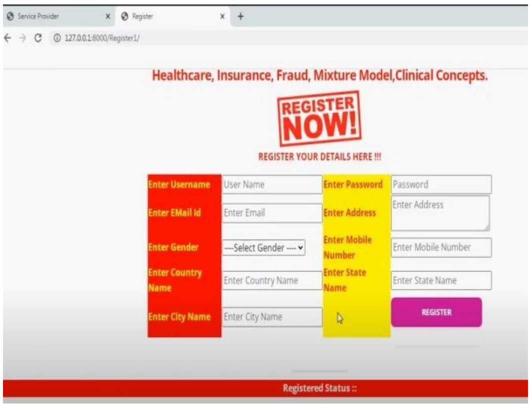
In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, PREDICT HEALTH INSURANCE CLAIM FRAUD TYPE, VIEW YOUR PROFILE.


Service Provider
127.0.0.1:8000/View_Remote_Users/
Identifying Health Insurance Claim Frauds Using Mixture of Clinical Concepts
Browse and Train & Test Health Insurance Data Sets
View Trained and Tested Accuracy in Bar Chart
View Trained and Tested Accuracy Results
View Prediction Of Health Insurance Fraud Type
View Health Insurance Fraud Type Ratio
Download Predicted Data Sets
View Health Insurance Fraud Type Ratio Results
View All Remote Users
Logout

VIEW ALL REMOTE USERS !!!

USER NAME	EMAIL	Gender	Address	Mob No	Country	State	City
Ashok	Ashok123@gmail.com	Male	#8928,4th Cross,Rajajinagar	9535866270	India	Karnataka	Bangalore

Fig.1. Home page.



Service Provider × Register × +

127.0.0.1:8000/Register1/

Healthcare, Insurance, Fraud, Mixture Model,Clinical Concepts.

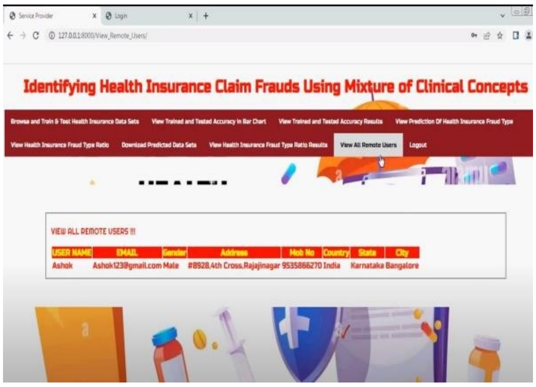
REGISTER NOW!

REGISTER YOUR DETAILS HERE !!!

Enter Username	User Name	Enter Password	Password
Enter EMail Id	Enter Email	Enter Address	Enter Address
Enter Gender	----Select Gender ----	Enter Mobile Number	Enter Mobile Number
Enter Country Name	Enter Country Name	Enter State Name	Enter State Name
Enter City Name	Enter City Name		REGISTER

Registered Status ::

Fig.2. Register page.



Service Provider × Login × +

127.0.0.1:8000/View_Remote_Users/

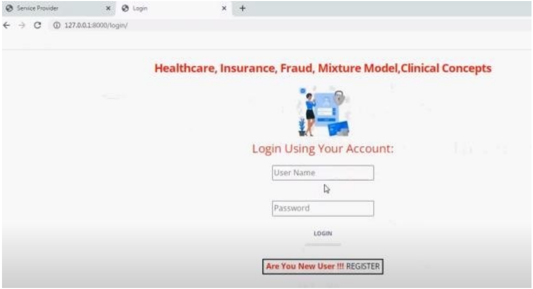
Identifying Health Insurance Claim Frauds Using Mixture of Clinical Concepts

Browse and Train & Test Health Insurance Data Sets | View Trained and Tested Accuracy in Bar Chart | View Trained and Tested Accuracy Results | View Prediction Of Health Insurance Fraud Type | View Health Insurance Fraud Type Ratio | Download Predicted Data Sets | View Health Insurance Fraud Type Ratio Results | View All Remote Users | Logout

VIEW ALL REMOTE USERS !!!

USER NAME	EMAIL	Gender	Address	Mob No	Country	State	City
Ashok	Ashok123@gmail.com	Male	#8928,4th Cross,Rajajinagar	9535866270	India	Karnataka	Bangalore

Fig.3. User details.



Service Provider × Login × +

127.0.0.1:8000/login/

Healthcare, Insurance, Fraud, Mixture Model,Clinical Concepts

Login Using Your Account:

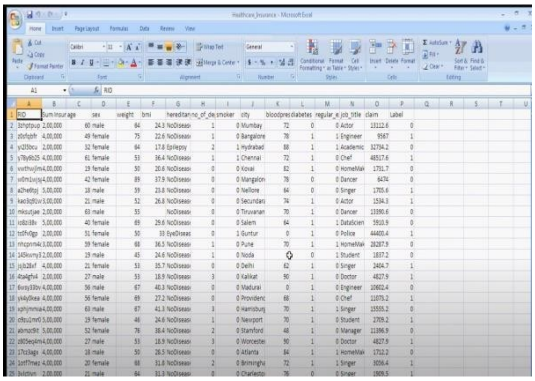
User Name

Password

LOGIN

Are You New User !!! REGISTER

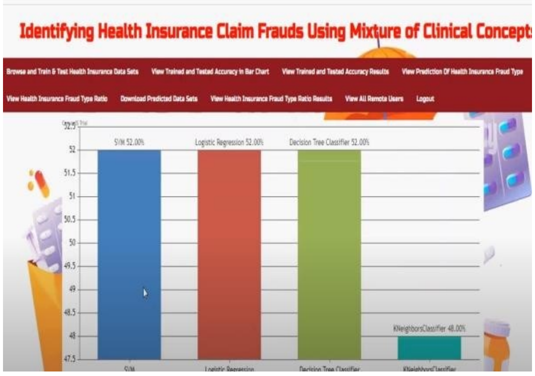
Fig.4. Login details.



Healthcare_Insurance - Microsoft Excel

A1 RID

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	RID	Sum Insur	age	sex	weight	bmi	hereditar	no_of_de	smoker	city	bloodpres	diabetes	regular_e	job_title	claim	Label
2	3zhptpup	2,00,000	60	male	64	24.3	NoDiseas	1	0	Mumbay	72	0	0	Actor	13112.6	0
3	z0sfqbfr	4,00,000	49	female	75	22.6	NoDiseas	1	0	Bangalore	78	1	1	Engineer	9567	1
4	yi2l5bcu	2,00,000	32	female	64	17.8	Epilepsy	2	1	Hydrabad	88	1	1	Academic	32734.2	0
5	y78y6b25	4,00,000	61	female	53	36.4	NoDiseas	1	1	Chennai	72	1	0	Chef	48517.6	1
6	vwthwjlm	4,00,000	19	female	50	20.6	NoDiseas	0	0	Kovai	82	1	0	HomeMak	1731.7	0
7	w0m1wjsj	4,00,000	42	female	89	37.9	NoDiseas	0	0	Mangalor	78	0	0	Dancer	6474	0
8	a2he6tpj	5,00,000	18	male	59	23.8	NoDiseas	0	0	Nellore	64	0	0	Singer	1705.6	1
9	kao3q91w	3,00,000	21	male	52	26.8	NoDiseas	0	0	Secundara	74	1	0	Actor	1534.3	1
10	mksutjae	2,00,000	63	male	55		NoDiseas	0	0	Tiruvanan	70	1	0	Dancer	13390.6	0
11	io8zi38v	5,00,000	40	female	69	29.6	NoDiseas	0	0	Salem	64	1	1	DataScien	5910.9	0
12	tc0fv0gp	2,00,000	51	female	50	33	EyeDiseas	0	1	Guntur	0	1	0	Police	44400.4	1
13	nhcpnm4c	3,00,000	59	female	68	36.5	NoDiseas	1	0	Pune	70	1	1	HomeMak	28287.9	0
14	145kwny3	2,00,000	19	male	45	24.6	NoDiseas	1	0	Noda	[illegible]	0	1	Student	1837.2	0
15	jsjb28xf	4,00,000	21	female	53	35.7	NoDiseas	0	0	Delhi	62	1	0	Singer	2404.7	1
16	4ta4gfv4	2,00,000	27	male	53	18.9	NoDiseas	3	0	Kalikat	90	1	0	Doctor	4827.9	1
17	6wsy33bv	4,00,000	56	male	67	40.3	NoDiseas	0	0	Madurai	0	1	0	Engineer	10602.4	0
18	yk4y0kea	4,00,000	56	female	69	27.2	NoDiseas	0	0	Providenc	68	1	0	Chef	11073.2	1
19	xphjmmia	4,00,000	63	male	67	41.3	NoDiseas	3	0	Harrisburg	70	1	1	Singer	15555.2	0
20	o9zu2mr0	5,00,000	19	female	46	24.6	NoDiseas	1	0	Newport	70	1	0	Student	2709.2	1
21	abmzc9it	5,00,000	52	female	76	38.4	NoDiseas	2	0	Stamford	48	1	0	Manager	11396.9	0
22	z805eq4m	4,00,000	27	male	53	18.9	NoDiseas	3	0	Worcester	90	1	0	Doctor	4827.9	1
23	17cz3ags	4,00,000	18	male	50	28.5	NoDiseas	0	0	Atlanta	84	1	1	HomeMak	1712.2	0
24	1otf7mez	4,00,000	20	female	68	31.8	NoDiseas	2	0	Brimingha	72	1	1	Singer	3036.4	1
25	3vlctivn	2,00,000	21	male	64	31.3	NoDiseas	0	0	Charlesto	76	0	0	Singer	1909.5	1

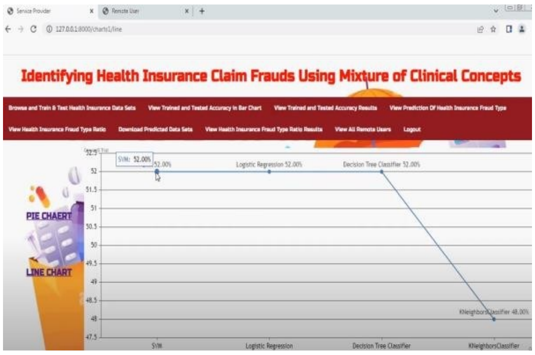


Identifying Health Insurance Claim Frauds Using Mixture of Clinical Concepts

Browse and Train & Test Health Insurance Data Sets | View Trained and Tested Accuracy in Bar Chart | View Trained and Tested Accuracy Results | View Prediction Of Health Insurance Fraud Type | View Health Insurance Fraud Type Ratio | Download Predicted Data Sets | View Health Insurance Fraud Type Ratio Results | View All Remote Users | Logout

Fig.5. Dataset details.

Fig.6. Output results.



Service Provider × Remote User × +

127.0.0.1:8000/charts1/line

Identifying Health Insurance Claim Frauds Using Mixture of Clinical Concepts

Browse and Train & Test Health Insurance Data Sets | View Trained and Tested Accuracy in Bar Chart | View Trained and Tested Accuracy Results | View Prediction Of Health Insurance Fraud Type | View Health Insurance Fraud Type Ratio | Download Predicted Data Sets | View Health Insurance Fraud Type Ratio Results | View All Remote Users | Logout

Fig.7. Output results.

CONCLUSION

In this paper, we pose the problem of fraudulent insurance claim identification as a feature generation and classification process. We formulate the problem over a minimal, definitive claim data consisting of procedure and diagnosis codes, because accessing richer datasets are often prohibited by law and present inconsistencies among different software systems. We introduce clinical concepts over procedure and diagnosis codes as a new representation learning approach. We assume that every claim is a representation of latent or obvious Mixtures of Clinical Concepts which in turn are mixtures of diagnosis and procedure codes. We extend the MCC model using Long-Short Term Memory network (MCC + LSTM) and Robust Principal Component Analysis (MCC + RPCA) to filter the significant concepts from claims and classify them as fraudulent or non fraudulent. Our results demonstrate an improvement scope to find fraudulent healthcare claims with minimal information. Both MCC and MCC + RPCA exhibit consistent behavior for varying concept sizes and replacement probabilities in the negative claim generation process. MCC + LSTM reaches an accuracy, precision, and recall scores of 59%, 61%, and 50%, respectively on the inpatient dataset. Besides, it presents 78%, 83%, and 72% accuracy, precision, and recall scores, respectively on the outpatient dataset. We notice similarity between the results of MCC and MCC + RPCA, as both use an SVM classifier. We believe that the proposed problem formulation, representation learning and solution will initiate new research on fraudulent insurance claim detection using minimal, but definitive data.

REFERENCES

[1] National Health Care Anti-Fraud Association, "The challenge of health care fraud," https://www.nhcaa.org/resources/health-care-antifraud- resources/the-challenge-of-health-care-fraud.aspx, 2020, accessed January, 2020.

[2] Font Awesome, "Image generated by free icons," https://fontawesome.com/license/free, 2020, online. [3] National Health Care Anti-Fraud Association, "Consumer info and action," https://www.nhcaa.org/resources/health-care-anti-fraudresources/ consumer-info-action.aspx, 2020, accessed January, 2020.

[4] W. J. Rudman, J. S. Eberhardt, W. Pierce, and S. Hart-Hester, "Healthcare fraud and abuse," Perspectives in Health Information Management/ AHIMA, American Health Information Management Association, vol. 6, no. Fall, 2009.

[5] M. Kirlidog and C. Asuk, "A fraud detection approach with data mining in health insurance," Procedia-Social and Behavioral Sciences, vol. 62, pp. 989–994, 2012.

[6] V. Rawte and G. Anuradha, "Fraud detection in health insurance using data

mining techniques," in 2015 International Conference on Communication, Information & Computing Technology (ICCICT). IEEE, 2015, pp. 1–5.

[7] C. Phua, D. Alahakoon, and V. Lee, "Minority report in fraud detection: classification of skewed data," Acm sigkdd explorations newsletter, vol. 6, no. 1, pp. 50–59, 2004.

[8] T. Ekina, F. Leva, F. Ruggeri, and R. Soyer, "Application of bayesian methods in detection of healthcare fraud," chemical engineering Transaction, vol. 33, 2013.

[9] J. Li, K.-Y. Huang, J. Jin, and J. Shi, "A survey on statistical methods for health care fraud detection," Health care management science, vol. 11, no. 3, pp. 275–287, 2008.

[10] R. J. Freese, A. P. Jost, B. K. Schulte, W. A. Klindworth, and S. T. Parente, "Healthcare claims fraud, waste and abuse detection system using non-parametric statistics and probability based scores," Jan. 19 2017, uS Patent App. 15/216,133.