ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 2, 2024

Identifying Health Insurance Frauds Claim Frauds Using Mixture Of Clinical Concepts

[1]K. LAKSHMAN REDDY SIR, [2]GUDALA BHANUPRAKASH

[1](Assistant Professor), MCA, S.V.K.P & Dr K.S. Raju Arts & Science College

[2]MCA, scholar, S.V.K.P & Dr K.S. Raju Arts & Science College

ABSTRACT

Patients depend on health insurance provided by the government systems, private systems, or both to utilize the high-priced healthcare expenses. This dependency on health insurance draws some healthcare service providers to commit insurance frauds. Although the number of such service providers is small, it is reported that the insurance providers lose billions of dollars every year due to frauds. In this paper, we formulate the fraud detection problem over a minimal, definitive claim data consisting of medical diagnosis and procedure codes. We present a solution to the fraudulent claim detection problem using a novel representation learning approach, which translates diagnosis and procedure codes into Mixtures of Clinical Codes (MCC). We also investigate extensions of MCC using Long Short Term Memory networks and Robust Principal Component Analysis. Our experimental results demonstrate promising outcomes in identifying fraudulent records.

1.INTRODUCTION

DATA analytics has progressively become crucial to almostany economic development area. Since healthcareis one of the largest financial sectors in the US economy,the massive amount of data, including health records, clinicaldata, prescriptions, insurance claims, provider information, andpatient information “potentially” presents incredible opportunitiesfor data analysts. Health insurance agencies process billionsof claims every year and healthcare expenses is over threetrillion dollars in the United States [1]. Figure 1 presents a conciseflow of a typical healthcare reconciliation process by usingdifferent entities involved. First, the service provider's officeensures that the patient has adequate coverage through his/herinsurance plan or other funds before getting any service. Next,the service provider identifies relevant diagnoses based on theinitial examinations performed on the patient. The serviceprovider then runs tests on the patient using one or moremedical interventions such as further diagnostics and surgical procedures. These diagnoses and procedures are usually tagged with the patient's report along with other information such as personal, demographic, and past/present visit information. At this point, the patient typically pays a co pay defined in his/her insurance plan and checks out. Then, the patient's report is sent to a medical coder who abstracts the information and createsa "superbill" containing all information about the provider, Given the economic volume of the healthcare industry, it is natural to observe fraudulent and fabricated claims submitted to insurance companies. The National Health Care Anti-Fraud Association (NHCAA) defines healthcare fraud as "Anintentional deception or misrepresentation made by a person,or an entity, with the knowledge that the deception couldresult in some unauthorized benefit to him or some otherentities" [3]. Those fabricated claims bear a very high cost,albeit they constitute a small fraction. According to NHCAA

the fraud related financial loss is in the orders of tens ofbillions of dollars in the United States [3]. Although there arestrict policies regarding fraud and abuse control in healthcareindustries, studies show that a very small portion of the lossesare recovered annually [4]. Most typical fraudulent activities committed by dishonest providers in the healthcare domain include the following.

_ Making false diagnoses to justify procedures that are notmedically necessary.

_ Billing for high priced procedures or services instead ofthe actual procedures, also called "upcoding".

_ 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 medically necessary to receive insurance

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 2, 2024

payments, especially forcosmetic procedures.

It is not feasible or practical to apply only domain knowled get o solve all or a subset of the issues listed above. Automated data analytics can be employed to detect fraudulent claims at an early stage and immensely help domain expertsto manage the fraudulent activities much better.

In this paper, we focus on the problem of health care fraud detection from health insurance providers' viewpoint.We answer the question of how to classify a procedure as legitimate or fraudulent from a claim when we only have limited data available, i.e. diagnosis and procedure codes. The problem of fraud detection in medical domain has been identified using different approaches such as data mining [5],classification methods [6], [7], Bayesian analysis [8], statistical surveys [9], non-parametric approaches [10], and expert analysis. Existing methods use physicians profile, background history, claim amount, service quality, services performed per provider, and related metrics from a claim database tocreate models for claim status prediction. Although thesemethods are successful, they often employ datasets that arenot publicly available. Furthermore, the variables featured inthose datasets are diverse and generally incompatible, whichmakes the solutions very difficult to transfer. In this studywe limit our available data to diagnosis and procedure codes,because obtaining third-party access to richer datasets is often prohibited by Health Insurance Portability and Accountability Act (HIPAA) in the US, General Data Protection Regulation(GDPR) in Europe or similar law in other regions. Besides, the healthcare industry is more apprehensive to share data compared to other sectors. Moreover, different software systems report different patient variables, which prohibits transferring solutions from one system to another. As a result, we confine our problem formulation to diagnosis and procedure codes which can always be handled in the same way whether they are country-specific or international. Our solution approach assumes the claim data as a mixture of medical concept swith respect to clinical codes of diagnoses and procedures in International Classification of Diseases (ICD) coding format. Moreover,

ISSN2454-9940
www.ijasem.org
Vol 18, Issue 2, 2024

the proposed approach works on other coding formats, e.g., Current Procedural Terminology (CPT) andHealthcare Common Procedure Coding System (HCPCS), ortheir combinations without any modification.

We represent an insurance claim as a Mixture of latentClinical Concepts (MCC) using probabilistic topic modeling.To the best of our knowledge this is the first work representing insurance claims as mixtures of clinical concepts in a latentspace. We assume that every claim is a representation oflatent or obvious mixtures of clinical concepts such as pain,mental or infectious diseases. Moreover, each clinical concept is a mixture of clinical codes, i.e., diagnosis and procedu recodes. The intuition behind our model comes from the services provided by doctor's offices, clinics, and hospitals. In general, a patient gets services based on specific issues consisting ofone or more diagnoses. Next, the service provider performs necessary procedures to treat the patient. Therefore, the diagnoses and procedures in a claim can be represented as a mixture of clinical concepts such as pain, mental, infectious diseases and/or their treatments. Note that, we do not explicitlylabel or interpret these concepts, as they are often not obvious, complex or require domain knowledge.

We extend the MCC model using Long-Short Term Memorynetworks and Robust Principal Component Analysis. Our goalin extending MCC is to filter the significant concepts fromclaims and classify them as fraudulent or non-fraudulent. Weextend MCC by using the concept weights of a claim as asequence representation within a Long-Short Term Memory(LSTM) network. This network allows us to represent theclaims as sequences of dependent concepts to be classified bythe LSTM. Similarly, we apply Robust Principal Component Analysis (RPCA) to filter significant concept weights by decomposing claims into a low-rank and sparse vector representations. The low-rank matrix ideally captures the noise-free weights.

Our unique contributions in this study can be summarized as follows.

We formulate the fraudulent claim detection problem overa minimal, definitive claim data consisting of procedureand diagnosis codes.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 2, 2024

_ We introduce clinical concepts over procedure and diagnosiscodes as a new representation learning approach.

_ We extend the mixtures of clinical concepts using LSTMand RPCA for classification.

We compare our approaches to the Multivariate Outlier Detection (MOD) [11] and a baseline method and report improved performance. Multivariate Outlier Detection method consists of two steps which are used to detect anomalous provider payments within Medicare claims data. In the firststep, a multivariate regression model is built on 13 handpicked features to generate corresponding residuals. Next, theresiduals are used as inputs to a generalized univariate probability model. Specifically, they used probabilistic programming methods in Stan [12] to identify possible outliers in the claimdata. The authors use the same CMS (Centers for Medicareand Medicaid Services) dataset that we use in our experiments with a different problem formulation. Their study incorporates providers and beneficiary data that was related to Medicare beneficiaries within the state of Florida, while we employ MOD on MCC features. On the other hand, the base line classifier assigns a test claim as the majority label present in the training claim data. Our experimental results show that MCC + LSTM reaches an accuracy, precision, and recall scores of 59%, 61%, and50%, respectively on the inpatient dataset obtained from CMS. In addition, it demonstrates 78%, 83%, and 72% accuracy, precision, and recall scores, respectively on the outpatient dataset We believe that the proposed problem formulation, representation learning and solution will initiate new research on fraudulent claim detection using minimal, but definitive data.

2.LITERATURE SURVEY

The following section discusses healthcare insurance fraud, its impact and its types. Also, this section provides a brief discussion on electronic claims management system and various data mining methods being deployed for healthcare insurance fraud detection. Healthcare Insurance Fraud Rebecca S. Busch in her book, Healthcare Fraud: Auditing and Detection Guide, describes Healthcare Fraud as “a knowing and intentional execution of a scheme to defraud a healthcare benefit program”.

Healthcare Fraud has been defined by the National Healthcare Anti-Fraud Association (NHCAA) as “an intentional deception or misrepresentation that an individual or entity makes knowing that the misrepresentation could result in some unauthorized benefit to the individual, or to the entity or to some other party” (BlueCross, 2016). Thornton et al. (2015) characterizes Healthcare Fraud as a crime which is ever evolving, where new schemes emerge on a regular basis (Thornton, Brinkhuis, Amrit, & Aly, 2015). Fraudsters, as the technology is advancing, are also becoming increasingly innovative in their methods for perpetrating fraudulent schemes (West & Bhattacharya, 2016). In order to make a continual progress in the improvement of the healthcare industry, it is very important to understand the working of the fraudsters and the methods that are used by the fraudsters (Thornton, Brinkhuis, Amrit, & Aly, 2015). FBI (2006) states that it is very difficult to place an exact value on the theft that is done through the means of insurance fraud. Fraud is supposed to be deliberately undetectable in nature (FBI, 2006). The complexity and confusing nature of the healthcare system makes it very hard to uncover fraudulent activities (Abdallah, Maarof, & Zainal, 2016). So, the number of instances of fraud that are discovered is much lower than the actual number of fraud instances (FBI, 2006). Fraudulent transactions are masqueraded as legitimate transactions so there will always be room for improvement in fraud detection domain (Wright, 2015). Impact of Healthcare Insurance Fraud Healthcare Systems have become an important part of modern life (Abdallah, Maarof, & Zainal, 2016). According to the Canadian Institute of Health Information (CIHI), the total healthcare spending in Canada was estimated to be around $264 billion in the year 2019. As a result, the healthcare industry has become a target for fraudsters (Abdallah, Maarof, & Zainal, 2016). As per the industry standards, 2% to 10 % of the total healthcare spending, an estimation of 5.2 billion to 26 billion CAD (BlueCross, 2016) is lost every year as a result of medical claims fraud. Therefore, healthcare

3. EXISTING SYSTEM

Yang and Hwang developed a fraud detection model usingthe clinical pathways concept and process-mining frameworkthat

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 2, 2024

can detect frauds in the healthcare domain [13]. Themethod uses a module that works by discovering structuralpatterns from input positive and negative clinical instances.The most frequent patterns are extracted from every clinicalinstance using the module. Next, a feature-selection module isused to create a filtered dataset with labeled features. Finally,an inductive model is built on the feature set for evaluating newclaims. Their method uses clustering, association analysis, andprincipal component analysis. The technique was applied on areal-world data set collected from National Health Insurance(NHI) program in Taiwan. Although the authors constructeddifferent features to generate patterns for both normal and abusiveclaims, the significance of those features is not discussed.

Bayerstadler et al. [14] presented a predictive model todetect fraud and abuse using manually labeled claims astraining data. The method is designed to predict the fraud andabuse score using a probability distribution for new claim invoices.Specifically, the authors proposed a Bayesian networkto summarize medical claims' representation patterns usinglatent variables. In the prediction step, a multinomial variablemodeling predicts the probability scores for various fraudevents. Additionally, they estimated the model parametersusing Markov Chain Monte Carlo (MCMC) [15].

Zhang et al. [16] proposed a Medicare fraud detectionframework using the concept of anomaly detection [17]. Firstpart of the proposed method consists of a spatial density basedalgorithm which is claimed to be more suitable compared tolocal outlier factors in medical insurance data. The secondpart of the method uses regression analysis to identify thelinear dependencies among different variables. Additionally,the authors mentioned that the method has limited applicationon new incoming data.

Kose et al. [18] used interactive unsupervised machinelearning where expert knowledge is used as an input to thesystem to identify fraud and abuse related legal cases inhealthcare. The authors used a pairwise comparison methodof analytic hierarchical process (AHP) to incorporate weightsbetween 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 proposedframework is evaluated based on the behavior traitsfound using the storyboard and later used for prescriptions byincluding all related persons and commodities such as drugs.

Bauder and Khoshgoftaar [19] proposed a general outlierdetection model using Bayesian inference to screen healthcareclaims. They used Stan model which is similar to [20] in theirexperiments. Note that, they consider only provider level-frauddetection without considering clinical code based relations.Many of those methods use private datasets or differentdatasets with incompatible feature lists. Therefore, it is verydifficult to directly compare these studies. In addition, HIPAA,GDPR and similar law enforce serious penalties for violationsof the privacy and security of healthcare information, whichmake healthcare providers and insurance companies very reluctantto share rich datasets if not at all. For these reasons, weformulate the problem over a minimal, definitive claim dataconsisting of diagnosis and procedure codes. Under this settingwe tackle the problem of flagging a procedure as legitimate orfraudulent using mixtures of clinical codes along with RNNand RPCA based encodings.

Disadvantages

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

3.1 PROPOSED SYSTEM

We extend the MCC model using Long-Short Term Memorynetworks and Robust Principal Component Analysis. Our goalin extending MCC is to filter the significant concepts fromclaims and classify them as fraudulent or non-fraudulent. Weextend

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 2, 2024

MCC by using the concept weights of a claim as asequence representation within a Long-Short Term Memory (LSTM) network. This network allows us to represent the

claims as sequences of dependent concepts to be classified by the LSTM. Similarly, we apply Robust Principal Component Analysis (RPCA) to filter significant concept weights by decomposing claims into a low-rank and sparse vector representations. The low-rank matrix ideally captures the noise-free weights.

Our unique contributions in this study can be summarizedas follows.

The system formulates the fraudulent claim detection problem overa minimal, definitive claim data consisting of procedureand diagnosis codes.

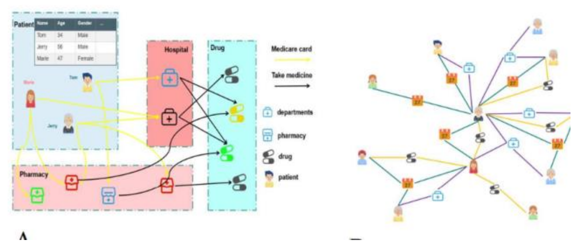
The system introduces clinical concepts over procedure and diagnosiscodes as a new representation learning approach.

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

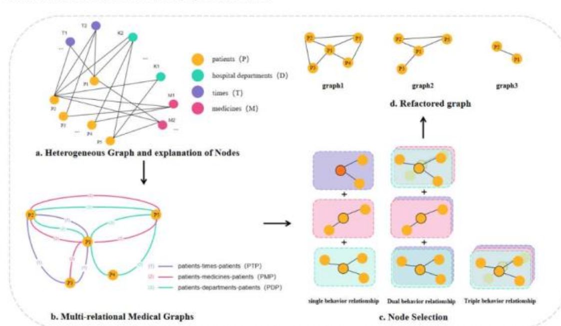
Advantages

- 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 anomalousprovider payments within Medicare claims data.

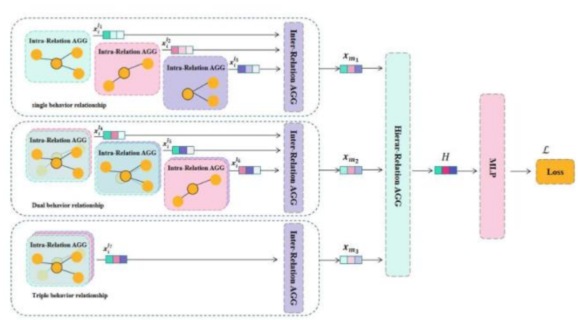
4. OUTPUTSCREENS



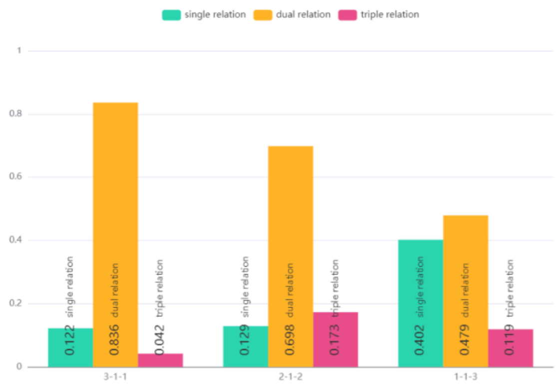
AHIN of the scenario of medical payment service



Process of decomposing and reconstructing heterogeneous graphs



ISSN2454-9940
www.ijasem.org
Vol 18, Issue 2, 2024



5. 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 thenegative claim generation process. MCC + LSTM reachesan accuracy, precision, and recall scores of 59%, 61%, and50%, respectively on the inpatient dataset. Besides, it presents78%, 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.

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 2, 2024

6. REFERENCE

1. Government of Pakistan. Introduction, Sehat Sahulat Program. 2019. https://sehatinsafcard.com/introduction.php. Accessed January 2023.

2. Government of Pakistan. Benefits Package. 2019. https://sehatinsafcard.com/benefits.php. Accessed January 2023.

3. Government of United States. Centers for Medicare and Medicaid Services. 1965. https://www.medicare.gov/. Accessed January 2023.

4. Gee J, Button M, Brooks G. The financial cost of healthcare fraud: what data from around the world shows. 2010.

5. Berwick DM, Hackbarth AD. Eliminating waste in US health care. *JAMA.* 2012;**307**(14):1513–1516. doi: 10.1001/jama.2012.362. [PubMed] [CrossRef] [Google Scholar]

6. M King K. Progress Made, but More Action Needed to Address Medicare Fraud, Waste, and Abuse. 2014. https://www.gao.gov/assets/gao-14-560t.pdf. Accessed January 2023.