International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
ISSN: 2457-0362
www.ijarst.in
IJARST

IDENTIFYING HEALTH INSURANCE CLAIM FRAUDS USING MIXTURE OF CLINICAL CONCEPTS

[1] Avula Lakshmaiah, [2] K.Meghana Sree, [3] J.Karthik Reddy, [4] K.Jaya Prakash, [5] K.Bhanu Teja

[1]Assistant Professor in Department of CSE Sri Indu College Of Engineering And Technology
avulalaxman944@gmail.com
[2,3,4,5] UG Scholars Department of CSE Sri Indu College Of Engineering And Technology

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. Machine learning is an important component of the growing field of data science. Through the use of statistical methods, different type of algorithms is trained to make classifications or predictions, and to uncover key insights in this project. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics. Machine learning algorithms build a model based on this project data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of datasets, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

Keywords - Healthcare, Insurance, Fraud, Mixture Model, Clinical Concepts

I INTRODUCTION

DATA analytics has progressively become crucial to almost any economic development area. Since healthcare is one of the largest financial sectors in the US economy, the massive amount of data, including health records, clinical data, prescriptions, insurance claims, provider information, and patient information “potentially” presents incredible opportunities for data analysts. Health insurance agencies process billions of claims every year and healthcare expenses is over three trillion dollars in the United States [1]. Figure 1 presents a concise flow of a typical healthcare reconciliation process by using different entities involved. First, the service provider’s office ensures that the patient has adequate coverage through his/her insurance plan or other funds before getting any service. Next,the service provider identifies relevant diagnoses based on the initial examinations performed on the patient. The service provider then runs tests on the patient using one or more medical 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 copay 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 creates a "superbill" containing all information about the provider, revised November 2020 patient, visit diagnoses and procedures. The diagnoses and procedures are also translated into medical codes in the superbill. The medical coder electronically sends the superbill to a medical biller who creates a medical claim by ensuring that the claim meets the required coding standards and format. Next, the claim is sent to the corresponding health insurance provider where the validity, correctness, and compliance of the claim is verified. They also prepare a detailed report that describes the coverage of procedures by the patient's insurance plan and send the report to the medical biller. Lastly, the medical biller sends an explanation to the patient describing his/her insurance coverage, benefits and balances.

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 Antifraud Association (NHCAA) defines healthcare fraud as "An intentional deception or misrepresentation made by a person, or an entity, with the knowledge that the deception could result in some unauthorized benefit to him or some other entities" [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 of billions of dollars in the United States [3]. Although there are strict policies regarding fraud and abuse control in healthcare industries, studies show that a very small portion of the losses are 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 not medically necessary.

• Billing for high priced procedures or services instead of the actual procedures, also called "up coding".

• 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.

It is not feasible or practical to apply only domain knowledge to 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 experts to manage the fraudulent activities much better.

II LITERATURE SURVEY

Md Enamul Haque is with the School of Computing and Informatics, University of Louisiana at Lafayette, LA, 70503 USA e-mail: enamul@louisiana.edu.

Mehmet Engin Tozal is an Assistant Professor at the School of Computing and Informatics, University of Louisiana at Lafayette, LA, 70503 USA e-mail: metozal@louisiana.edu.

Manuscript received April, 2020;

Fraud and abuse are among the most prominent issues in the massive healthcare system. In addition to frauds, accidental errors in documentation cause significant losses of money, time and labor.

Several works in the literature propose solutions to the problem of fraud, abuse and error detection in medical, pharmaceutical, and related domains. 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 [102]. 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 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) [42, 30].

Zhang et al. [105] proposed a Medicare fraud detection framework using the concept of anomaly detection [104]. 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. [60] 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) [106] to incorporate weights between actors (patients) and attributes. Expectation maximization (EM) [33] 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 [8] proposed a general outlier detection model using Bayesian inference

to screen healthcare claims. They used Stan model which is similar to [97] in their experiments. Note that, they consider only provider level-fraud detection without considering clinical code based relations. Many of those methods use private datasets or different datasets with incompatible feature lists. Therefore, it is very difficult to directly compare these studies. In addition, HIPAA, GDPR and similar law enforce serious penalties for violations of the privacy and security of healthcare information, which make healthcare providers and insurance companies very reluctant to share rich datasets if not at all. For these reasons, we formulate the problem over a minimal, definitive claim data consisting of diagnosis and procedure codes. Under this setting we tackle the problem of flagging a procedure as legitimate or fraudulent using mixtures of clinical codes along with RNN and RPCA [101] based encoding

III 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 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.

IV PROBLEM STATEMENT

Let us assume we are given a dataset of verified and reimbursed (or positive) insurance claims, C + = {c1, c2, . . . , c|C+|}, where |C +| is the number of the claims. Each claim ci consists of a set of diagnosis and procedure codes summarizing the treatment for a particular patient. Let us denote the set of all diagnosis codes D = {d1, d2, . . . , d|D|} and procedure codes P = {p1, p2, . . . , p|P |}, where |D| and |P| are the number of diagnosis and procedure codes, respectively. The objective is to identify an insurance claim as either fraudulent or legitimate with respect to the mixture of clinical concepts. Note that, a major limitation in healthcare insurance fraud identification is the lack of ground-truth negative claims. We tackle that issue from a statistical sampling perspective, The overall problem statement is that given ground truth, positive claims and a new incoming test claim ct, can we determine if ct has any inconsistent diagnosis and procedure codes implying a fraudulent or erroneous claim? Let us consider that the test claim ct consists of codes {d2761, d4271, p395, p428, p272} where d and p denote diagnoses and procedures, respectively.

We use subscript notation of the code identification numbers with letters d and p to differentiate between diagnosis and procedures. In the claim, d2761 and d4271 diagnoses codes are related to a disease of respiratory systems that denote Hyposmolality/hyponatremia and Paroxysmal ventricular tachycardia, respectively. However, not all the procedure codes in the claim are compatible with the diagnoses. p428 denotes Other repair of esophagus which is related to disease of respiratory system. On the other hand, p395 and p272 denote Other repair of vessels and Diagnostic procedure on oral cavity which are treatments for diseases related to circulatory and dental systems. Therefore, the example claim ct should be identified as fraudulent (or erroneous) and spared for further investigation due to the existence of the irrelevant procedures, p395 and p272.

V PROPOSED SYSTEM

We extend the MCC model using Long-Short Term Memory networks and Robust Principal Component Analysis. Our goal in extending MCC is to filter the significant concepts from claims and classify them as fraudulent or non-fraudulent. We extend MCC by using the concept weights of a claim as a sequence 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 summarized as follows.

The system formulates the fraudulent claim detection problem over a minimal, 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

VI IMPLEMENTATION

We first demonstrate the hierarchical relationships among related diagnosis and procedure codes using an example claim. Next, we present our representation learning process, the Mixture of Clinical Concepts (MCC), which extracts features based on weighted clinical concepts. Then, we present an example claim with both diagnosis and procedure codes to represent the tree structured hierarchy within the actual ICD coding system. Subsequently, the concept weights of a claim are treated as input features to a Long-Short Term Memory (LSTM) [24] based recurrent neural network. The primary objective to use LSTM with the MCC architecture is to model the hierarchical dependencies and relatedness among the concepts. In addition, we separately employ Robust Principal Component Analysis (RPCA) to obtain a low rank data structure which minimizes the impact of noise and outliers in the MCC representation. Usually, health insurance claims consist of multi-level relations among the constituent ICD, HCPCS level-I (CPT), and level-II codes. We demonstrate a simple example of a claim containing four codes including two diagnoses (238.8, 238.73) and two procedures (58.51, 58.53) Both diagnosis and procedure codes follow a hierarchical tree structure in the ICD coding format. Diagnosis and procedure codes are connected using red dashed line in our partial bipartite graph representation of this claim. For example, the root node with diagnosis code 238 denotes Neoplasm of uncertain behavior of other and unspecified sites and tissues refering to the behavior of a tumor which cannot be predicted via pathology. The child nodes of 238 are different versions of the root node which share the same medical concept. Note that,generally a claim involves diagnosis and procedure codes from multiple disjoint trees where each tree Represents a medical concept. We only present single tree structure for simplicity with respect to both diagnoses and procedures in The parent node of the tree represents a broader diagnosis or procedure. However, node 238 is not an absolute root node but an intermediate node of a bigger concept tree. For instance, the node 238 is a sub-concept of Neoplasm which denotes an abnormal growth or death of tissue. The terminal and intermediate nodes provide more specific diagnosis and procedure based on various health issues. The root nodes that represent broader medical concepts are not included in the actual claim for most of the cases. Therefore, we aim to include those latent concepts in the representation of corresponding claims.

The objective of the medical codes representation learning is to find vector-based claim representations such that each claim ci is represented as a k dimensional vector vi . An effective vector representation would place related clinical codes under similar latent concepts. We exploit Latent Dirichlet Allocation (LDA) [25], a popular method from the NLP community that have already been used with success in medical informatics, in our first step of claim representation. Using LDA, each claim is represented as a mixture of different clinical concepts where each claim is considered to have a set of concepts that are assigned to it via LDA. The assignment process is similar to probabilistic latent semantic analysis (pLSA) [26]. The only difference with LDA is that the concept distribution is assumed to have sparse Dirichlet priors which encode a claim using a small set of concepts and the concepts use only a small set of frequently used clinical codes. In process provides a concise and hierarchical representation of clinical codes and a more compact assignment of claims to the concepts. We generate concepts using LDA which assumes that the whole claim data contains predefined K concepts. Generally, each claim is characterized by a distribution over concepts as θ. Additionally, each concept is represented by a distribution over all V clinical codes as φ. Considering LDA to generate concept zi,j from a claim, the following generative process is considered.

VII CONCLUSION

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 no 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 research on fraudulent insurance claim detection using minimal, but definitive data.

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.

REFERENCE

[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.

[11] R. A. Bauder and T. M. Khoshgoftaar, "Multivariate anomaly detection in medicare using model residuals and probabilistic programming," in The Thirtieth International Flairs Conference, 2017.

[12] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker,J. Guo, P. Li, and A. Riddell, "Stan: A probabilistic programming language," Journal of statistical software, vol. 76, no. 1, 2017.

[13] W.-S. Yang and S.-Y. Hwang, "A process-mining framework for the detection of healthcare fraud and abuse," Expert Systems with Applications, vol. 31, no. 1, pp. 56–68, 2006.

[14] A. Bayerstadler, L. van Dijk, and F. Winter, "Bayesian multinomial latent variable modeling for fraud and abuse detection in health insurance," Insurance: Mathematics and Economics, vol. 71, pp.

244–252, 2016.

[15] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, "Introducing markov chain monte carlo,"Markov chain Monte Carlo in practice, vol. 1, p. 19, 1996.

[16] W. Zhang and X. He, "An anomaly detection method for medicare fraud detection," in Big Knowledge (ICBK), 2017 IEEE International Conference on. IEEE, 2017, pp. 309–314.

[17] L. Zhang, J. Lin, and R. Karim, "Adaptive kernel density-based anomaly detection for nonlinear systems," Knowledge-Based Systems, vol. 139, pp. 50–63, 2018.

[18] I. Kose, M. Gokturk, and K. Kilic, "An interactive machine-learningbased electronic fraud and abuse detection system in healthcare insurance," Applied Soft Computing, vol. 36, pp. 283–299,2015.

[19] R. A. Bauder and T. M. Khoshgoftaar, "A probabilistic programming approach for outlier detection in healthcare claims," in Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on. IEEE, 2016, pp. 347–354.

[20] J. Wang and S. Luo, "Augmented beta rectangular regression models: A bayesian perspective," Biometrical Journal, vol. 58, no. 1, pp. 206–221, 2016.

[21] Centers for Medicare and Medicaid Services, "ICD-10,"https://www.cms.gov/Medicare/Coding/ICD10/, 2020, accessed January, 2020.

[22] Medical Billing and Coding, "HCPCS codes," tps://www.medicalbillingandcoding.org/hcpcs-codes/, 2020, accessed January, 2020.

[23] American Academy of Professional Coders, "CPT codes," https://coder.aapc.com/cpt-codes,2020, accessed January, 2020.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of machine Learning research, vol. 3, no. Jan, pp. 993–1022, 2003.

[26] T. Hofmann, "Probabilistic latent semantic analysis," in Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1999, pp. 289–296.

[27] E. J. Candes, X. Li, Y. Ma, and J. Wright, "Robust principal component ` analysis?" Journal of the ACM (JACM), vol. 58, no. 3, p. 11, 2011.

[28] Centers for Medicare and Medicaid Services, "Research, statistics, data and systems,"https://www.cms.gov/Research-Statistics-Data-andSystems/Downloadable-Public-Use-Files/SynPUFs/DE Syn PUF, 2020, accessed January, 2020.

[29] R. A. Bauder and T. M. Khoshgoftaar, "The detection of medicare fraud using machine learning methods with excluded provider labels," in The Thirty-First International Flairs Conference, 2