**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, 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 “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 “upcoding”.

\_ 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.

In this paper, we focus on the problem of healthcare 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 to create models for claim status prediction. Although these methods are successful, they often employ datasets that are not publicly available. Furthermore, the variables featured in those datasets are diverse and generally incompatible, which makes the solutions very difficult to transfer. In this study we 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 concepts with respect to clinical codes of diagnoses and procedures in International Classification of Diseases (ICD) coding format. Moreover, the proposed approach works on other coding formats, e.g., Current Procedural Terminology (CPT) and Healthcare Common Procedure Coding System (HCPCS), or their combinations without any modification.

We represent an insurance claim as a Mixture of latent Clinical 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 latent space. We assume that every claim is a representation of latent 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 procedure codes. 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 of one 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 explicitly label or interpret these concepts, as they are often not obvious, complex or require domain knowledge.

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.

\_ We formulate the fraudulent claim detection problem over a minimal, definitive claim data consisting of procedure and diagnosis codes.

\_ We introduce clinical concepts over procedure and diagnosis codes as a new representation learning approach.

\_ We extend the mixtures of clinical concepts using LSTM and 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 first step, a multivariate regression model is built on 13 hand picked features to generate corresponding residuals. Next, the residuals 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 claim data. The authors use the same CMS (Centers for Medicare and 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 baseline 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%, and 50%, 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. The rest of the paper is organized as follows. Section II presents the related work. We formally introduce the problem and present our solution in Section III. Section IV demonstrates the empirical evaluations. Finally, we conclude the paper in Section V.