IEEE Transactions on Services Computing,Volume:15,Issue:4,Issue Date:01.July-Aug.2022

Identifying Health Insurance Claim Frauds Using Mixture of Clinical Concepts

Md Enamul Haque and Mehmet Engin Tozal

***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.**

***Index Terms*—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, 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.

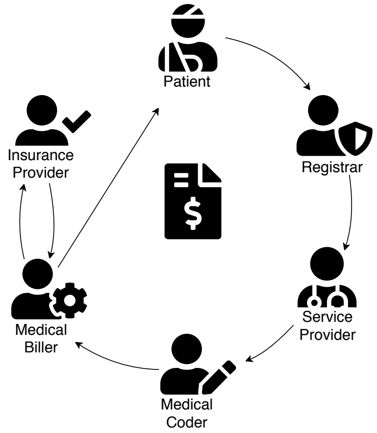


Fig. 1. An overview of the entities interacting in a typical claim reconciliation process [2].

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.

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; revised November 2020

- 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.

II. Related Work

Fraud and abuse are among the most prominent issues in the massive healthcare system. In addition to frauds, accidental errors in documentation causes 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 [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.

Bauder and Khoshgoftaar [19] proposed a general outlier detection model using Bayesian inference to screen healthcare claims. They used Stan model which is similar to [20] 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 based encodings.

III. Mixture of Clinical Concepts

In this section, we first briefly present the medical coding in healthcare. Next, we formally introduce the fraud detection with only diagnosis and procedure codes problem. Lastly, we present our solution approach using health insurance claims representation based on latent clinical concepts.

A mixture of concepts is assigned to a claim based on different health conditions which are inherent characteristics of a treatment. The World Health Organization (WHO) introduced the International Statistical Classification of Diseases and Related Health Problems (ICD) as concise representations of diagnoses and procedures in the form of alphanumeric codes. ICD codes are revised and improved at times. CMS have replaced ICD-9 by ICD-10 coding format since 2015 [21]. We also observe different levels of coding methods for procedures using Healthcare Common Procedure Coding System (HCPCS). The level-I HCPCS codes are equivalent to Current Procedural Terminology (CPT) codes which are used to describe physician services such as blood transfusion. Level-II HCPCS codes are separate from CPT codes and used to describe non-physician services such as ambulance rides, wheelchairs, durable medical equipment [22]. Level-III HCPCS codes were developed for specific programs, but their use have been dropped since 2003. HCPCS level-I and level-II codes consist of only numeric and alpha-numeric values, respectively.

Table I presents an example claim from outpatient claims that includes ICD diagnosis, HCPCS level-I (CPT) procedures, and HCPCS level-II procedure codes [23]. The claim represents the treatment of a patient with diseases of the circulatory system using ICD and HCPCS codes. The diagnoses contain both numeric and alpha-numeric codes that use ICD coding format. On the other hand, procedure codes include both level-I numeric and level-II alphanumeric HCPCS codes. Note that the context descriptions in Table I are not part of the claim dataset, though one can obtain them from https://www.findacode.com/search/.

TABLE I
AN OUTPATIENT CLAIM DATA CONTAINING ICD FORMAT DIAGNOSIS AND HCPCS FORMAT PROCEDURE CODES REPRESENTING TREATMENT / SERVICE RELATED TO THE "DISEASES OF THE CIRCULATORY SYSTEM".

Diagnosis code	ICD9 v3 context
V719	Observation for unspecified suspected condition
7230	Spinal stenosis in cervical region
4359	Unspecified transient cerebral ischemia

Procedure code	HCPCS context
36415	Under venous procedures
36430	Transfusion, blood or blood components
83880	Assay of natriuretic peptide
82043	Under chemistry procedures
80053	Under organ or disease oriented panels
A0021	Ambulance service

A. Problem Statement

Let us assume we are given a dataset of verified and reimbursed (or positive) insurance claims, $C^+ = \{c_1, c_2, \ldots, c_{|C^+|}\}$, where $|C^+|$ is the number of the claims. Each claim c_i 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 $D = \{d_1, d_2, \ldots, d_{|D|}\}$ and procedure codes $P = \{p_1, p_2, \ldots, p_{|P|}\}$, 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, introduced in Section IV.

The overall problem statement is that given ground truth, positive claims and a new incoming test claim c_t, can we determine if c_t has any inconsistent diagnosis and procedure codes implying a fraudulent or erroneous claim? Let us consider that the test claim c_t consists of codes $\{d_{2761}, d_{4271}, p_{395}, p_{428}, p_{272}\}$ 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, d_{2761} and d_{4271} 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. p_{428} denotes *Other repair of esophagus* which is related to disease of respiratory system. On the other hand, p_{395} and p_{272} 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 c_t should be identified as fraudulent (or erroneous) and spared for further investigation due to the existence of the irrelevant procedures, p_{395} and p_{272}.

B. Problem Solution

In this part, 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 structred 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) codes in Figure 2. 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 Figure 2. The parent node of the tree represents a broader diagnosis or procedure. However, node 238 is not an abosolute 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.

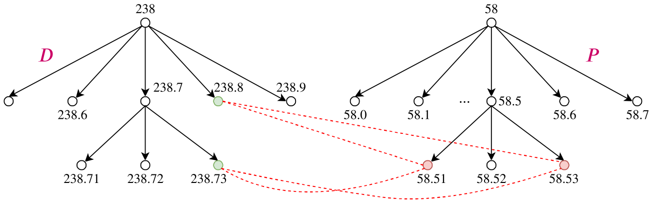


Fig. 2. Hierarchy of clinical codes within a claim represented using ICD-9/10 CM coding format.

The objective of the medical codes representation learning is to find vector-based claim representations such that each claim c_i is represented as a k dimensional vector v_i. 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 encodes a claim using a small set of concepts and the concepts use only a small set of frequently used clinical codes. In practice, this 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 $z_{i,j}$ from a claim, the following generative process is considered.

1) $\theta_i \sim \text{Dirichlet}_K(\alpha)$, where $i \in \{1, \ldots, |C|\}$ and Dirichlet(α) is a Dirichlet distribution with sparse symmetric parameter α.
2) $\phi_k \sim \text{Dirichlet}_V(\beta)$, where $k \in \{1, \ldots, K\}$ and β is sparse.
3) For each of the code positions i, j, where $i \in \{1, \ldots, |C|\}$ and $j \in \{1, \ldots, N_i\}$
   a) Choose a concept $z_{ij} \sim \text{Multinomial}(\theta_i)$
   b) Choose a code $w_{ij} \sim \text{Multinomial}(\phi_{z_{ij}})$

where α and β are hyper-parameters of Dirichlet priors. z_{ij} is the identity of concept of code j in claim i and an integer between 1 and K. w_{ij} is the identity of code w in claim i and and integer between 1 and V. N_i denotes number of codes in the i^{th} claim. Note that, different prior distributions can be assumed based on problem domains.

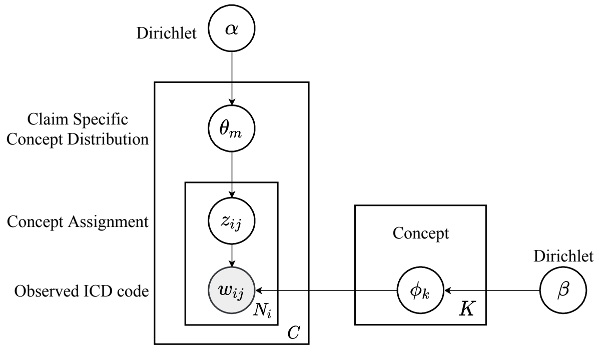


Fig. 3. Plate diagram for generating concepts using Latent Dirichlet Allocation.

We demonstrate the generative process of MCC using the following plate diagram in Figure 3. The process begins by initializing the concept as a Multinomial distribution over all diagnosis and procedure codes in the training claims which is parameterized by ϕ_k. The α and β hyperparameters of Dirichlet priors denote the document-concept and concept-clinical code density, respectively. A smaller α contributes to imposing less number of concept for a claim. Similarly, a high β contributes to making a concept using most of the codes.

We present a simplified architecture of the Recurrent Neural Network with LSTM blocks used in our experiments in Figure 4. In the enhanced learning step, the clinical concept distribution over a defined number of concepts are treated as the features of a claim. The features are then fed into the recurrent neural network with two LSTM layers for enhanced claim representation. Each claim is initially passed through an embedding layer. The encoder is a two sequential LSTM layers with *sigmoid* activation functions. The output layer is passed through another sigmoid function for the binary classification. The main purpose of using LSTM for encoding the mixture data is to differentiate the unique features from every claim.

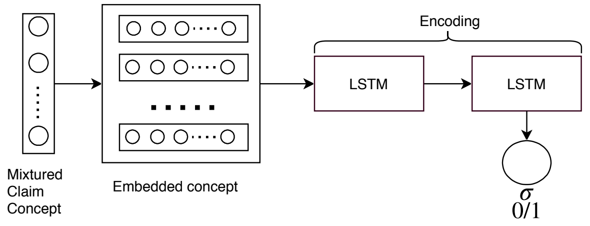


Fig. 4. Two-layer sequential LSTM architecture for enhanced representations of claims based on mixtures of clinical concepts.

In addition to LSTM, we used Robust Principal Component Analysis [27] (RPCA) to remove noise from the features. RPCA is a variant of Principal Component Analysis. The main idea behind RPCA is that the errors in the dataset can occur in large quantity but sparsely (that is with only a few entries). So a balance between the sparseness of the error and the quantity of the error is considered to calculate the principal components optimally. Robust PCA decomposes a matrix $\mathbf{D}$ into $\mathbf{L}$ and $\mathbf{S}$ where $\mathbf{D} = \mathbf{L} + \mathbf{S}$ by solving the following optimization problem:

$$\begin{aligned} \min_{L,S} \quad & ||\mathbf{L}||_* + \lambda||\mathbf{S}||_\mathbf{1} \\ \text{s.t.} \quad & ||\mathbf{D} - \mathbf{L} - \mathbf{S}||_\mathbf{F}^\mathbf{2} = \mathbf{0} \end{aligned} \tag{1}$$

where $\mathbf{L}$ is the low rank matrix, $\mathbf{S}$ denotes the sparse matrix, $\mathbf{D}$ is the original data matrix, $||\mathbf{L}||_*$ is the nuclear norm of $\mathbf{L}$, $||\mathbf{S}||_\mathbf{1}$ is the one norm of $\mathbf{S}$, $||\mathbf{D} - \mathbf{L} - \mathbf{S}||_\mathbf{F}^\mathbf{2}$ is the squared Frobenius norm of $\mathbf{D} - \mathbf{L} - \mathbf{S}$ and λ is the sparsity parameter. RPCA recovers the underlying low-rank data matrix, $\mathbf{L}$, even in the presence of outliers, large errors and noise, captured in $\mathbf{S}$.

IV. Empirical Evaluations

In this section, we first introduce the dataset used in our experiments. Next, we demonstrate the negative claim data preparation from positive claims. Finally, we present and discuss the experimental results.

A. Datasets

We collect medicare and medicaid data from Centers for Medicare and Medicaid Services (CMS) website [28]. The dataset contains inpatient and outpatient claims from years

TABLE II
A SAMPLE INPATIENT CLAIM DATA CONTAINING ICD-9 V3 DIAGNOSIS AND PROCEDURE CODES.

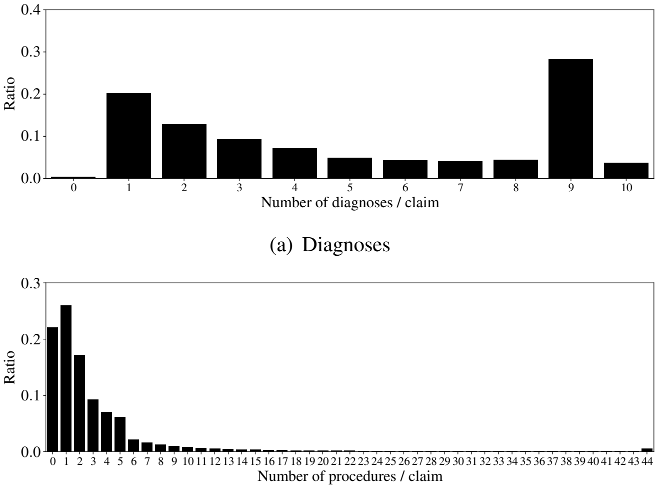
Diagnosis code	**ICD9 v3 context**
41041	Acute myocardial infarction of other inferior wall, initial episode of care
30000	Anxiety state, unspecified
4139	Other diagnostic procedures on spleen
41401	Coronary atherosclerosis of native coronary artery
V5869	Long-term (current) use of other medications
25000	Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled
2721	Biopsy of bony palate
5601	Paralytic ileus
2948	Other persistent mental disorders due to conditions classified elsewhere

Procedure code	**ICD9 v3 context**
4019	Other diagnostic procedures on lymphatic structures
2724	Biopsy of mouth, unspecified structure
V5861	Long-term use anticoagulants
66	Operations on fallopian tubes

between 2008-2010 containing 20 files each. The claims include medical diagnosis and procedure codes along with other de-identified zip code, location, beneficiary payments, provider data and patient specific information. In our model we omit this information, because not only it is irrelevant to our problem definition but also it limits our approach to the information captured by a certain software or healthcare provider. On the other hand, diagnosis and procedure codes, whether country-specific or international, can always be handled in the same way.

We conducted two sets of experiments to demonstrate the effectiveness of Mixture of Clinical Concepts (MCC) using 66,773 *inpatient* and 79,079 *outpatient* claims from a randomly selected single data file. A patient is categorized as inpatient if the hospital stay is longer and prescribed by an authorized doctor for relevant procedures. On the contrary, a patient is categorized as an outpatient if he/she gets lab test, X-rays, or any other hospital services without the written order from a doctor to be admitted to a hospital as an inpatient. Inpatient medicaid and medicare claims consist of ICD format coding for both diagnoses and procedures. On the other hand, outpatient claims consist of ICD, CPT, and HCPCS coding formats.

The primary goal of our experiments is to predict whether a claim contains inconsistent procedures with respect to its diagnosis set. The inpatient data contains maximum of 10 and 5 diagnosis and procedure codes, respectively. On the other hand, the outpatient data contains maximum of 10 and 44 diagnosis and procedure codes, respectively. Similar to [29], we assume that our dataset contains valid, positive ground truth claims. By positive ground truth claim, we mean that the claims were properly analyzed and verified by the insurance providers before the reimbursements of the claimed amounts to the service providers. Table II shows a sample inpatient claim consisting of diagnosis and procedure codes along with their descriptions. The claim in the table consists of ICD codes for both diagnosis and procedures. The diagnosis codes of the claim indicates that the patient suffered from anxiety, diabetes, paralysis of the intestinal muscles, and coronary related diseases. In addition, the procedure codes reflect the treatment of the diagnoses using biopsy and operations on fallopian tubes. Note that the context descriptions in Table II are not part of the claim dataset, though one can obtain them from https://www.findacode.com/search/. The actual inpatient sample dataset can be found from https://tiny.cc/inpatient_claim/.



(a) Diagnoses

(b) Procedures

Fig. 5. Diagnosis and procedure frequency distributions per claim.

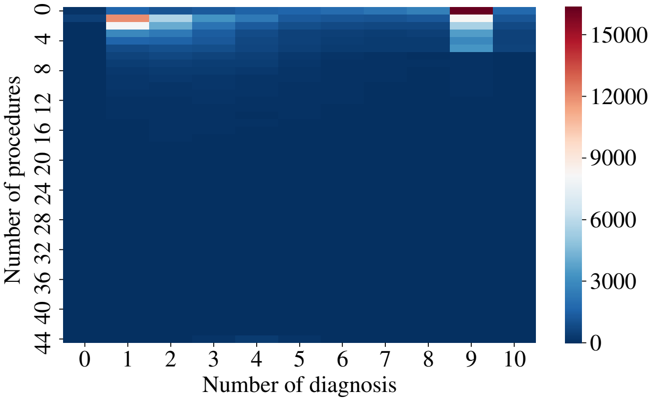


Fig. 6. Co-occurrence map of diagnosis and procedure codes' frequencies

Figures 5(a) and 5(b) present the frequency distributions of the diagnosis and procedure codes per claim in our dataset. As shown in the figures both distributions have long tails as expected, because doctors often make zero to a few diagnoses and apply zero to several treatment procedures. Insurance claims with so many diagnosis and so many procedures are rare. Figure 6 presents a heatmap denoting the co-occurrence of diagnosis and procedure frequencies in our claims dataset. Consistent with Figure 5, Figure 6 demonstrates that around 8% of the claims have one diagnosis and one procedure, followed by 6% having one diagnosis and two procedures. In addition, 7% of the claims have two diagnosis and one or two procedures, followed by 9% having nine diagnosis and one or two procedures. In the dataset we only have two claims that have ten diagnosis and 44 procedures.

One significant challenge to identify fraudulent claims under *supervised learning* setting is the lack of negative ground truth claims. We generate synthetic negative claims by replacing the procedure codes of the positive claims by a certain probability, τ, while preserving their diagnosis codes. Our goal is to generate inconsistent procedures or irrelevant diagnoses and procedures by randomly replacing the procedure codes in a claim while preserving the diagnosis codes.

Algorithm 1 presents the pseudocode for the synthetic negative claim generation process. The algorithm expects the set of positive claims, C^+, and the procedure replacement probability, $\tau \in (0, 1]$, as the input and returns the set of negative claims, C^- as the output. The outer loop at line 1 goes through every claim in C^+ and creates a candidate negative claim, c^-, by cloning a positive claim at line 2. The inner loop at line 3 goes through all procedures in c^- and attempts to replace each procedure with another procedure in set $P \setminus P_i$ by probability τ at lines 4 to 7. If at least one procedure is replaced in c^-, the algorithm adds c^- to the set of negative claims C^- at lines 9 to 11. Finally, line 13 returns the set of negative claims.

Algorithm 1: Negative Claim Generation Process

```
Input: Set of positive claims, $C^+ = \{c_i^+ | c_i^+ = (D_i, P_i)\}$
Input: Replacement probability, $\tau \in (0, 1)$
Output: Set of negative claims, $C^- = \{c_j^- | c_j^- = (D_j, P_j)\}$
 1: for $\forall c_i^+ \in C^+$ do
 2:     $c_i^- \leftarrow c_i^+$
 3:     for $\forall p_{ij} \in P_i$ of $c_i^-$ do
 4:         $p \leftarrow$ generate a random value between 0 and 1
 5:         if $p \leq \tau$ then
 6:             replace $p_{ij}$ by a randomly
 7:              selected procedure from$P \setminus P_i$
 8:         end if
 9:     end for
10:     if at least one procedure is randomly replaced in $c_i^-$
        then
11:         $C^- \leftarrow C^- \cup c_i^-$
12:     end if
13: end for
14: return  $C^-$
```

B. Experimental Results

In our dataset, the inpatient claim data allocates ten fields for diagnosis codes and five fields for the corresponding procedure codes for every claim. On the other hand, the outpatient claim data allocates ten fields for diagnosis codes and 44 fields for the corresponding procedure codes for every claim.

In the first step, we employ LDA on the claim dataset to compute the concept level weights (probabilities) with respect to the individual claims. Next, we use the concept weights of each claim as an input feature to the embedding layer of the LSTM network. Then, the resulting dense representation of the claim is fed into two sequential LSTM layers to extract the significant concepts. In the output layer we use the sigmoid function for binary classification. In addition to using LSTM on top of the clinical concept mixture for each claim, we use Robust Principal Component Analysis (RPCA) to extract significant concepts without noise. We present two heat-maps of fifty randomly selected RPCA processed claims in Figure 7. The figures demonstrate that each claim has fewer concepts which light-up more than the others. Note that, the claims show distinguishable patterns on the lower concept mixture.

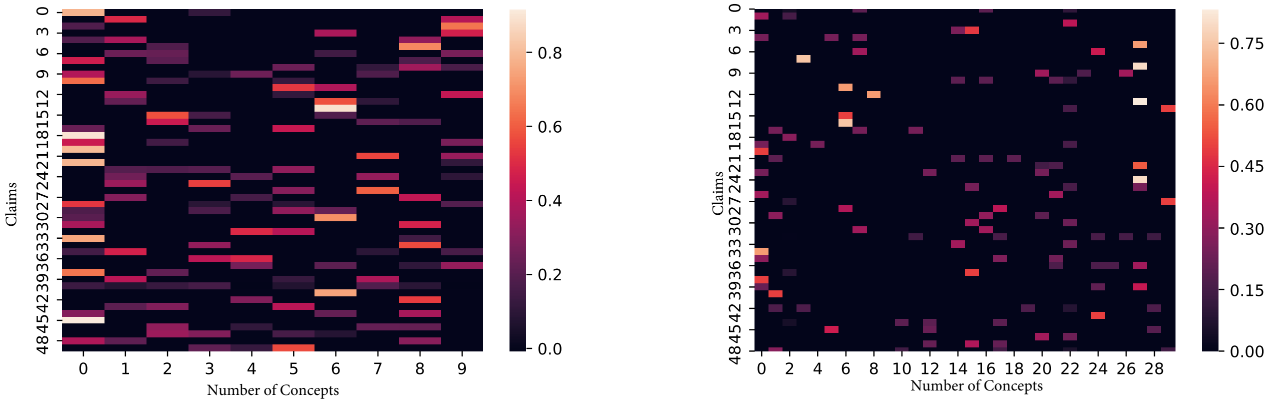
Methods: We use MCC and two other variants denoted as MCC + LSTM and MCC + RPCA as hybrid methods. We compare our methods with Multivariate Outlier Detection (MOD) [11] and Baseline classifier that always predicts most frequent label in the training claims. MCC uses the basic LDA model for the initial representations of the claims. MCC + LSTM uses additional LSTM networks for enhanced representation of claim codes. MCC + RPCA uses a variant of principal component analysis to compute a low rank and a sparse matrix. We employ the low rank matrix for later evaluation. We use Support Vector Machine (SVM) for classification with MCC and MCC + RPCA.

Model parameters and specifications: We use varying model parameters for both LDA and LSTM methods. In our proposed method, as LDA requires predefined number of clinical concepts, we select a range of concepts from 10 to 100 with intervals of 10. We use the same concept frequency as embedding vector length in LSTM networks. In addition, we use two sequential LSTM layers with dropout as 0.5 to avoid model over-fitting. The sequential layers use sigmoid as activation function. We used *scikit learn* package in Python to implement our models.

In the following, we present the experimental results of all the methods with respect to varying concept sizes and procedure replacement probability thresholds. Next, we present a brief discussion after selecting roughly the best parameters using concept size and procedure probability threshold.

1) Effect of Clinical Concept Size: We use different clinical concept sizes to represent a claim as a feature vector. MCC, MCC + LSTM, MCC + RPCA, Multivariate Outlier Detection, and Baseline with respect to varying concept sizes are presented in Figures 8 and 9 for inpatient and outpatient datasets. Note that, we present the average results in both figures for each concept size where results are averaged over ten different procedure replacement threholds.

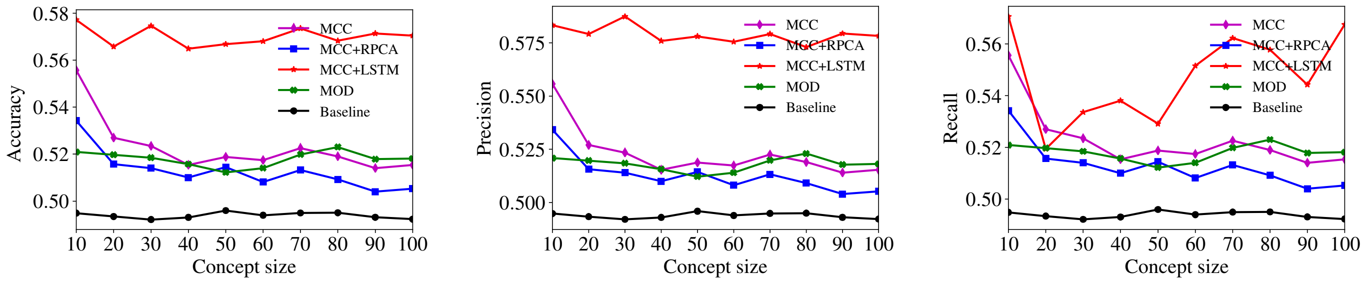
Figure 8 presents accuracy, precision, and recall scores for all the methods with respect to varying concept sizes on



(a) 10 concepts

(b) 30 concepts

Fig. 7. Heatmap on the mixture claim representations of 50 randomly selected claims for (a) 10 and (b) 30 concepts.

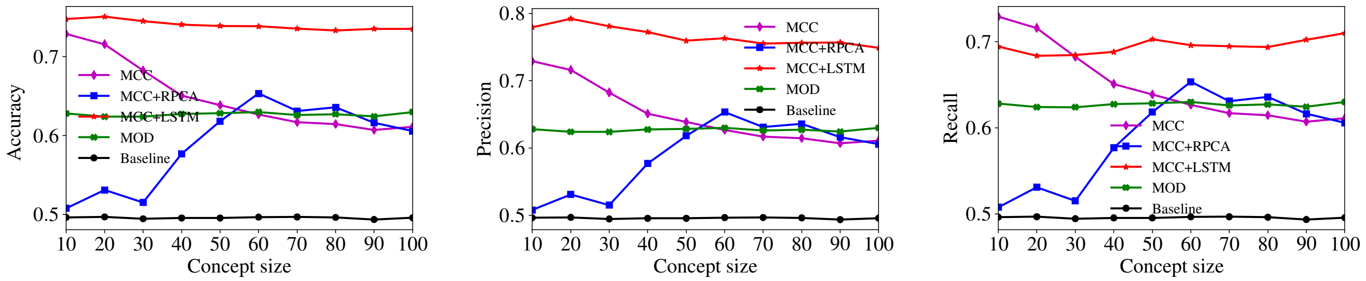


(a) Accuracy

(b) Precision

(c) Recall

Fig. 8. Evaluation metrics for four different methods with respect to concept size on inpatient claims.



(a) Accuracy

(b) Precision

(c) Recall

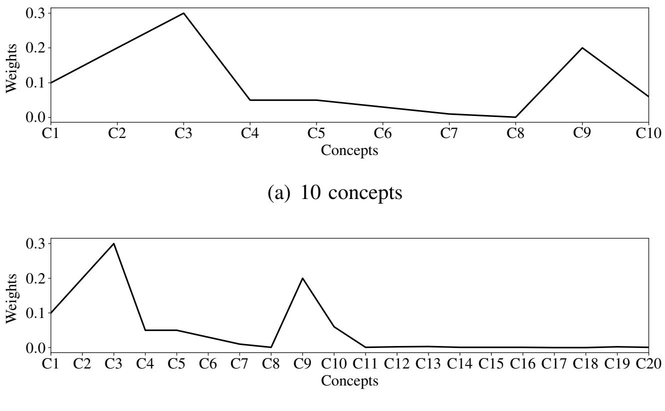
Fig. 9. Evaluation metrics for four different methods with respect to concept size on outpatient claims.

inpatient dataset. Figure 16(a) demonstrates that the accuracy scores of MCC and MCC + RPCA are roughly constant with respect to the concept sizes. We observe similar behavior for MOD and Baseline approaches. Generally, the number of concepts increases as the claim frequency grows with the inclusion of various diagnosis and procedure codes. Due to the effect of latent concept size of a claim, our proposed approaches using MCC and the variants perform similarly with respect to different concept size initializations. However, MCC + LSTM performs better for concept sizes of 10, 70, and 100 for inpatient dataset shown in Figure 16(c). We also observe that both MCC and MCC + RPCA works better for smaller concept sizes of 20 or less with respect to accuracy, precision, and recall. With the increase in concept size, the scores do not exhibit significant improvement. The reason for this phenomena is that both MCC and MCC + RPCA does not capture distinguishable concepts with higher number of initialized concepts in the model. On the other hand, MCC + LSTM demonstrates better results for higher concept sizes such as 70. The reason for LSTM based MCC to perform better for higher concept sizes is that the LSTM network is able to extract significant concepts from longer input sequences. LSTM + MCC also shows poor recall scores for concept sizes between 20 and 50.

Figure 9 presents the accuracy, precision, and recall scores of all the methods with respect to concept sizes on outpatient data. Note that, outpatient claims have typically more clinical codes per claim than the inpatient claims. An inpatient claim has maximum of ten and five diagnosis and procedure codes, respectively. On the other hand, outpatient claims have maximum of ten and fourty-four diagnosis and procedure codes, respectively. This is very important due to the fact that the number of codes within a claim determines the number of concepts a claim can be associated with. Unlike the results obtained from the methods on inpatient data, we observe inconsistencies in results with respect to outpatient dataset when MCC and MCC + RPCA are applied. Both methods perform similarly with larger concept sizes of 50 or more. MCC performs better for smaller concept sizes such as 10 and 20, followed by gradual decrease for the higher concept sizes. We can conclude that default MCC with lower concept size have relatively more distinguishing features. However, MCC + RPCA demonstrates different pattern where smaller concept sizes show poor results, followed by a steep increase for higher concept sizes. More specifically, MCC + RPCA works better for concept sizes of 50 or higher. The reason for this phenomena is that the low rank features extracted by the RPCA procedure from the MCC generated data are very similar for both positive and negative claims. MCC + LSTM, MOD, and Baseline approaches exhibit constant results with respect to the concept sizes. Note that, MCC + LSTM performs the best among all the variants of MCC and the other approaches.

We present the evaluations for the multivariate regression splines residuals with Bayesian univariate outlier detection model [11] (MOD) to detect fraudulent claims based on the MCC generated features in Figures 8 and 9 with respect to inpatient and outpatient claims, respectively. As all the features are probability distributions over concepts, most of the feature values are very close to each other, except a few. Therefore, the MOD model performs poorly compared to MCC + LSTM on all concept sizes. It also performs poorly compared to MCC and MCC + RPCA for lower concept sizes on inpatient data. In addition, it shows poor performance compared to MCC and MCC + RPCA for lower and higher concept sizes, respectively on outpatient claims. Note that, in the original study [11], the authors employ MOD to detect "anomalous provider payments" in Medicare claims data using 13 variables such as zip code, year, number of services provided by providers, and average payment amount. Since these features are irrelevant to our problem setup, we used the proposed approach, MOD, with MCC generated features. In addition, an "outlier" is not necessarily a "fraudulent" claim. A claim with inconsistent diagnosis procedure codes may not deviate from the population in terms of zip code, year, number of services or payment amount. In fact, our negative sample generation process (Algorithm 1) generates such negative claims. Hence, the difference in performance should be attributed to different problem formulations rather then the strengths of the methods.

We observe that the evaluation of MCC and MCC + RPCA on inpatient dataset demonstrate constant performance in Figure 8. On the other hand, MCC + LSTM demonstrates consistent results for both types of claims data in Figures 8 and 9. The primary reason of the phenomena can be described using the latent concept size. As we extract varying number of concepts from the same claim, we should expect an optimal number that is applicable to a claim. For instance, a claim might contain 10 significant concepts in actual analysis. As a result, we should expect to have 10 significant concepts in our concept weight distribution for larger concept sizes as well. We explain the phenomena in Figure 10 where we assume that a claim consists of five significant concepts denoted as $C1, C2, C3, C4,$ and $C9$. We use concept sizes 10 and 20 to extract those five significant concepts using MCC and its variants such as MCC + LSTM and MCC + RPCA. Although the initialization of concept sizes are 10 and 20 for MCC, we notice that both Figures 10(a) and 10(b) show similar weights for the corresponding concepts. Therefore, we can conclude that the concept weight distribution of a claim does not change significantly with respect to different concept size initializations in MCC.



(a) 10 concepts

(b) 20 concepts

Fig. 10. An example of a claims' concept weight distribution for (a) 10 and (b) 20 concepts.

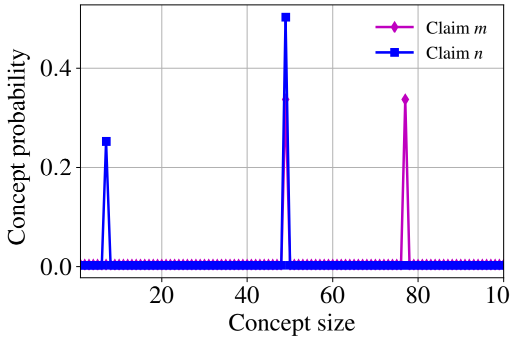
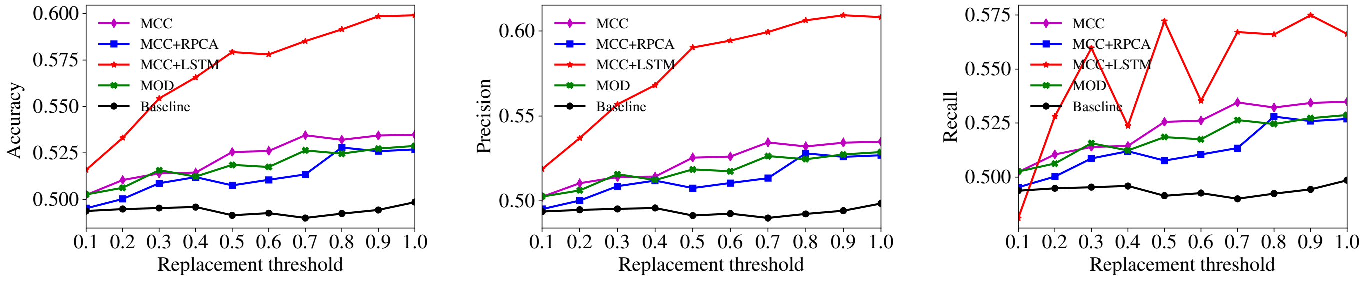


Fig. 11. Similarity between two claims with respect to the probability distribution of concepts.

Overall, we observe that MCC and MCC + RPCA performs better for concept size of 10 on inpatient data. Similar pattern is observed for MCC on outpatient data as well. However, MCC + RPCA performs better when concept size is 50 or more on outpatient data. MCC + LSTM performs constantly better on both inpatient and outpatient claims with respect to most of the concept sizes.



(a) Accuracy

(b) Precision

(c) Recall

Fig. 12. Evaluation metrics for four different methods with respect to the replacement probability in negative procedure sampling on inpatient claims.

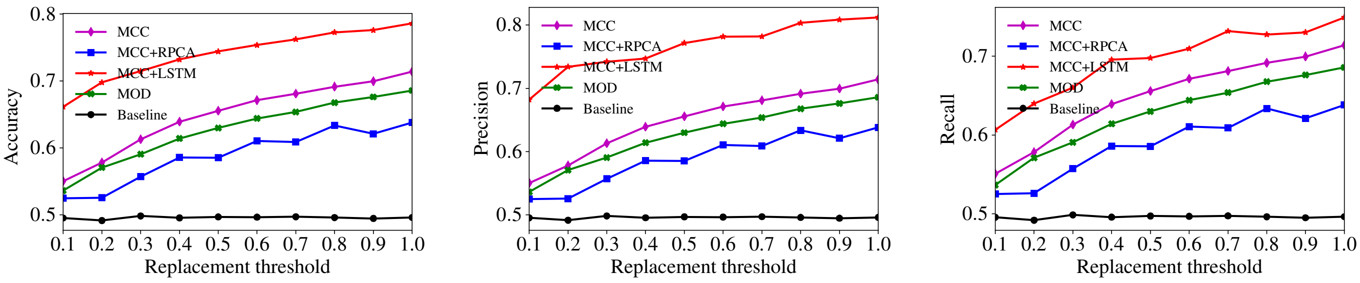
Our empirical analysis finds that positive claims exhibit lower number of major concepts compared to the negative claims which is also intuitive, because a claim with higher number of medical concepts are not common. We present two example positive claims m and n from our dataset based on the cosine similarity of their MCC based feature representation when initial concept size is chosen as 100. The cosine similarity between the representations of these two claims is 0.64. Claim m consists of procedure codes 96.71 and 34.59. On the other hand, claim n consists of 96.71, 42.731, and 53.085. We present the similarity in Figure 11 where claim m and n share two concepts. Procedure code 96.71 denotes *continuous invasive mechanical ventilation for less than 96 consecutive hours*. Procedure code 34.59 indicates *decortication of lung* which is a medical procedure to help the lung re-expand to normal state. Additionally, procedure 42.731 indicates *atrial fibrillation* which is used to express rapid heart rate condition. However, 53.085 denotes health issues related to *under incision procedures on the urethra*. Therefore, the figure suggests that the claims share majority of the concepts based on heart related problems. Similarly, negative claims differ from positive claims by significant concept probability.

2) *Effect of Procedure Replacement Probability:* In this part, we present the average results for all the methods with respect to the procedure replacement probability threshold for ten different concept sizes. The random sampling process enables us to create a dataset with procedure codes that are inconsistent with each other or with their respective diagnosis codes. Naturally, the negative dataset has claims with varying concepts that are difficult to distinguish from the positive ones, because the procedure codes are replaced randomly by probability τ.

The evaluation results presented in Figure 12 shows that the replacement probability, τ, in Algorithm 1 does not have a significant effect on the overall performance, except for MCC + LSTM method on inpatient dataset. We notice similar results for MCC + RPCA compared to MCC as the aim for both methods is to emphasize the concepts that are significant within a claim. MCC + LSTM demonstrated improved accuracy and precision scores in Figures 12(a) and 12(b) with respect to increased probability threshold. However, we notice interesting results of MCC + LSTM for different procedure replacement probabilities where recall scores fluctuate in Figure 12(c). As the LSTM network uses dependency among the concepts to embed a claim, the final representation of two different claims with negligible change in concept weights become significantly different. As a result the final layer of LSTM with sigmoid function can assign them to the classes. However, we believe that the phenomena occurs due to the significant presence of claims with no procedure codes. As a consequence of such presence, the positive claim data will contain concept features solely based on the diagnosis codes of a claim. For example a claim might contain “pain” and “fever” concepts based on diagnosis codes only. On the other hand, another claim may contain the same concepts based on both diagnosis and procedure codes.

Figure 13 presents accuracy, precision, and recall scores for all the methods with respect to procedure probability threshold on outpatient claims. Unlike the results of MCC + LSTM on inpatient data with respect to the probability threshold, we observe impressive performance on outpatient data. The accuracy, precision, and recall scores increase linearly for all the methods, except the Baseline approach as it predicts a claim based on the majority of the class labels in the training dataset. MCC and MCC + LSTM methods perform better for all the replacement thresholds, except MCC + RPCA which, unlike inpatient dataset, performs poorly compared to MOD approach. The reason for MCC + RPCA to work poorly relates to the higher number of clinical codes within a claim in outpatient dataset. The low rank data extracted from the MCC generated features have more overlapping component within a claim which increases both false positive and false negatives. We observe better results when only MCC generated features are used for classification for the similar reason. In addition, we notice rising accuracy, precision, and recall for all the methods with respect to increasing probability threshold. This happens due to the higher number of procedure codes changed in the negative claims. The best performing method, MCC + LSTM, achieves nearly 75% and 80% accuracy for 50% and 100% procedure changes in the negative claims, respectively. MCC and MCC + RPCA achieves nearly 71% and 65% accuracy for the similar procedure percentage change on the outpatient dataset.

3) *Discussions:* In this part we present a rough approximation of the evaluation metrics with respect to both concept size and procedure replacement probability threshold. In addition, we present a comparative evaluation of all the methods using limited concept size with respect to inpatient and outpatient



(a) Accuracy

(b) Precision

(c) Recall

Fig. 13. Evaluation metrics for four different methods with respect to the replacement probability in negative procedure sampling on outpatient claims.

claim datasets.

TABLE III
COMPARATIVE RESULTS WITH RESPECT TO ACCURACY, PRECISION, AND RECALL FOR FIVE DIFFERENT CLAIM IDENTIFICATION APPROACHES ON INPATIENT DATA.

Methods	Evaluation Metrics		
	Accuracy	Precision	Recall
MCC	0.527	0.527	**0.527**
MCC+LSTM	**0.587**	**0.612**	0.503
MCC+RPCA	0.519	0.519	0.519
MOD	0.518	0.518	0.518
Baseline	0.493	0.493	0.493

TABLE IV
COMPARATIVE RESULTS WITH RESPECT TO ACCURACY, PRECISION, AND RECALL FOR FIVE DIFFERENT CLAIM IDENTIFICATION APPROACHES ON OUTPATIENT DATA.

Methods	Evaluation Metrics		
	Accuracy	Precision	Recall
MCC	0.740	0.740	**0.740**
MCC+LSTM	**0.784**	**0.832**	0.721
MCC+RPCA	0.523	0.523	0.523
MOD	0.679	0.679	0.679
Baseline	0.491	0.491	0.491

First, we select lower concept size 20 which roughly works better for all the approaches analyzing the results demonstrated in Figures 8, 9, 12, and 13. Table III presents accuracy, precision, and recall scores of all five methods using inpatient data for concept size 20 and procedure replacement probability threshold 0.9. We select 0.9 as the replacement threshold for the discussion as it shows improved performance for all the methods. In all cases, MOD and Baseline methods do not perform well as MOD considers all the weights as features in the regression model and Baseline favors the most frequent labels in the training data as predicted labels for test claims. On the other hand, our methods use some form of filtering on the concept weights to represent as a claim feature. The experimental results demonstrate that our proposed methods have improvement scope with respect to accuracy, precision, and recall scores for inpatient dataset. We believe the methods can be further improved by identifying different group of claims with respect to the frequency of procedures and actual concept hierarchies present in a claim. The concept hierarchies can be extracted from multiple sources such as clinical notes and patient logs which is not publicly available. In addition, we consider the inclusion of hierarchy as out of scope to our study as our goal is to use minimal data to find out fraudulent claims.

Next, we present accuracy, precision, and recall scores for all the methods with respect to the same concept size 20 and procedure replacement threshold of 0.9 on outpatient data using Table IV. In this case, MCC and MCC + LSTM performs better compared to MOD approach. However, MOD outperforms MCC + RPCA as the later method generates significant false positive and negative outcomes. In all cases, MCC + LSTM outperforms the remaining approaches because of the transformation of concept weights. Based on the previous analysis and results from Table IV, we suggest to use smaller concept size based on the number of claims and the average number of clinical codes present in the outpatient claim. It reduces the computational complexity for larger numbers of claim processing. We use Robust PCA (RPCA) instead of PCA which is able to extract the low-rank structured principal components even with extreme sparsity in the data. RPCA exploits both nuclear and ℓ_1 norm minimization to allow sparsity and noise by applying a convex optimization. Hence, by using the low rank component of the claim data we ensure that the final feature representation of a claim can effectively identify the underlying concept mixtures. As the number of codes within inpatient claims are lower compared to outpatient claims, we notice a moderate difference in the classification outcome, which is an indirect consequence of using low rank component from a claim.

Finally, we present experimental results using only accuracy scores in Table V for five concept sizes with respect to two types of datasets and five methods. We select both small and large concept sizes due to lower claim counts in the dataset and also because higher concept size does not provide improved results for most methods. We select the procedure replacement probability as 0.5 for unbiased evaluation. In general, both small and large concept sizes work better for both datasets when MCC + LSTM is used. However, smaller concept sizes are preferred on inpatient dataset when other MCC approaches are used. In addition, MCC + RPCA works best on outpatient dataset for larger concept sizes. Note that, the medical concepts are very widely spread in the outpatient claims as it contains additional HCPCS level-I (CPT) and level-II codes in the procedure code set. As a result, MOD

TABLE V

ACCURACY SCORES FOR DIFFERENT METHODS WITH RESPECT TO CONCEPT SIZES ON BOTH DATASETS.

Concept Size	Inpatient					Outpatient				
	MCC	MCC+LSTM	MCC+RPCA	MOD	Baseline	MCC	MCC+LSTM	MCC+RPCA	MOD	Baseline
10	**0.570**	**0.586**⋆	**0.543**	0.510	0.494	**0.734**	**0.743**⋆	0.510	0.629	0.498
20	0.535	**0.580**⋆	0.516	**0.532**	0.487	0.728	**0.750**⋆	0.536	0.624	**0.499**
50	0.513	**0.574**⋆	0.507	0.507	0.488	0.639	**0.743**⋆	**0.638**	**0.633**	0.491
70	0.525	**0.573**⋆	0.495	0.521	0.488	0.617	**0.746**⋆	0.629	0.629	0.495
100	0.524	**0.572**⋆	0.498	0.525	**0.494**	0.598	**0.734**⋆	0.582	0.632	0.497

and the Baseline methods perform poorly on inpatient data compared to outpatient dataset. In addition, we observe that the increase in concept size does not have significant impact on the results. However, we observe performance improvements on outpatient dataset that have increased unique diagnosis and procedure codes and fewer claims with null procedures.

Overall, MCC and MCC + LSTM generate more consistent and better results for varying concept sizes and replacement probabilities over MCC + RPCA, MOD, and Baseline. MCC + LSTM reaches an accuracy, precision, and recall scores of 59%, 61%, and 50% (Table III), respectively on the inpatient dataset. Besides, it presents 78%, 83%, and 72% (Table IV) 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 also share our implementation at `http://tiny.cc/mcc-concpet/` to support reproduction of our results in future studies.

V. CONCLUSIONS

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.

APPENDIX A

A. Supplemental Results

In this supplemental part, we demonstrate the significance of the accuracy scores of different methods. First, we present additional results from Logistic regression, Random forest, and Decision tree algorithm to augment the analysis from our previous discussions. We chose MCC + LSTM from our previous discussion because the method provides superior results among other methods. Next, we present a pairwise comparison among all the methods when applied on inpatient dataset. We chose to use inpatient dataset because outpatient claims demonstrate superior performance in both concept and replacement probability settings.

We present the results of Logistic regression (LR), Random forest (RF), Decision tree (DT), and LST based MCC in Figures 14, 16, 15, 17. Figure 14 presents accuracy, precision, and recall scores of MCC based LR, DT, RF, and LSTM applied on inpatient dataset with respect to different concept size. The results demonstrate that both DT and LSTM performs better compared to other methods. Figure 15 presents similar results applied on inpatient dataset with respect to different procedure replacement threholds.

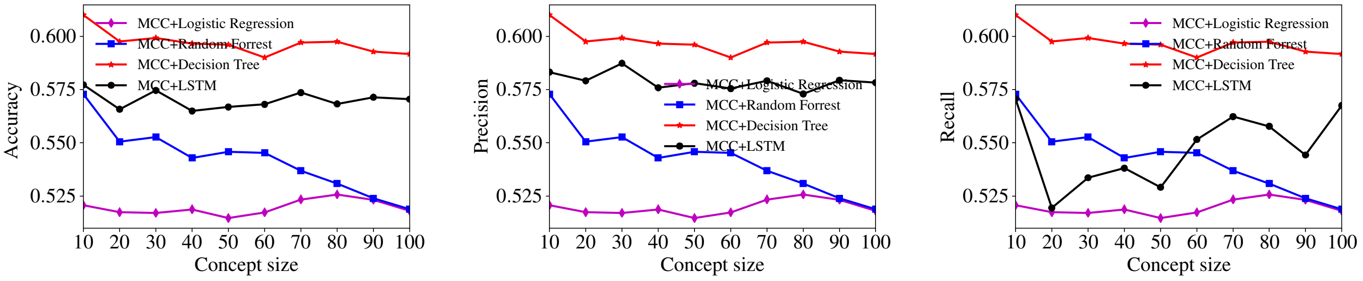
Figure 16 and 17 presents accuracy, precision, and recall scores of LR, DT, RF, and LSTM based methods applied on outpatient dataset with respect to concept size and procedure replacement probability threhosld, respectively. We observe similar results in both figures, however, LSTM remains the top performer in both concept and probability threshold cases.

Finally, we present p-values from non-parametric Nemenyi pairwise comparison among all the methods applied on the average accuracy scores of every procedure replacement threshold on inpatient dataset in Table VI. The scores clearly demonstrates that there are significant differences between the results of MOD and MCC based LSTM and DT. Similarly, MCC based LSTM, DT, and RF are significantly different from the baseline. We consider 95% significance level with $p < 0.05$ to measure the differences. We can interpret the remaining p-values from the symmetric Table VI and consult with the accuracy, precision, and recall plots to provide an in-depth differences of results.

REFERENCES

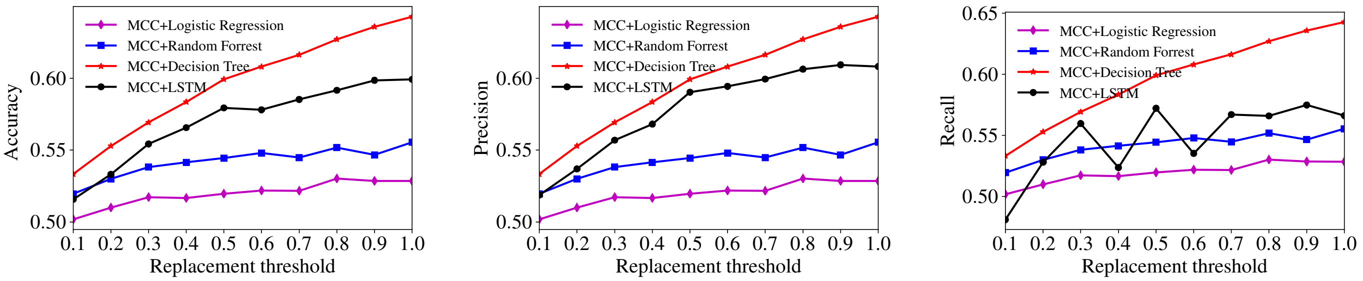
[1] National Health Care Anti-Fraud Association, “The challenge of health care fraud,” https://www.nhcaa.org/resources/health-care-anti-fraud-resources/the-challenge-of-health-care-fraud.aspx, 2020, accessed January, 2020.

[2] Font Awesome, “Image generated by free icons,” https://fontawesome.com/license/free, 2020, online.



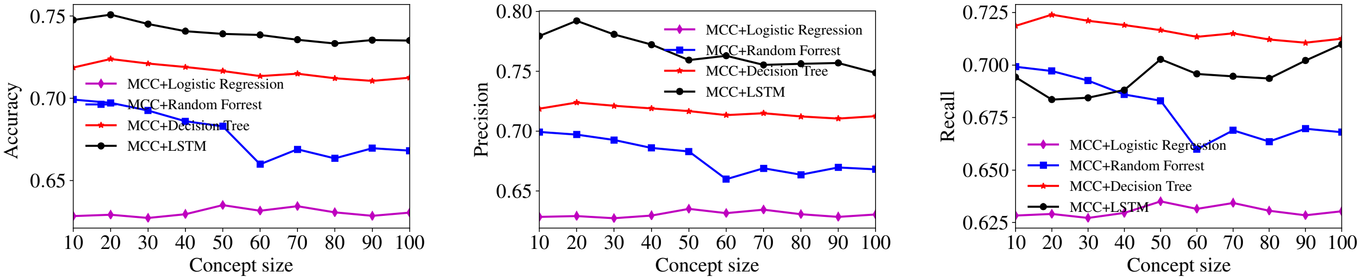
(a) Accuracy (b) Precision (c) Recall

Fig. 14. Evaluation metrics for four different methods with respect to concept size on inpatient claims.



(a) Accuracy (b) Precision (c) Recall

Fig. 15. Evaluation metrics for four different methods with respect to the replacement probability in negative procedure sampling on inpatient claims.



(a) Accuracy (b) Precision (c) Recall

Fig. 16. Evaluation metrics for four different methods with respect to concept size on outpatient claims.

[3] National Health Care Anti-Fraud Association, "Consumer info and action," https://www.nhcaa.org/resources/health-care-anti-fraud-resources/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.

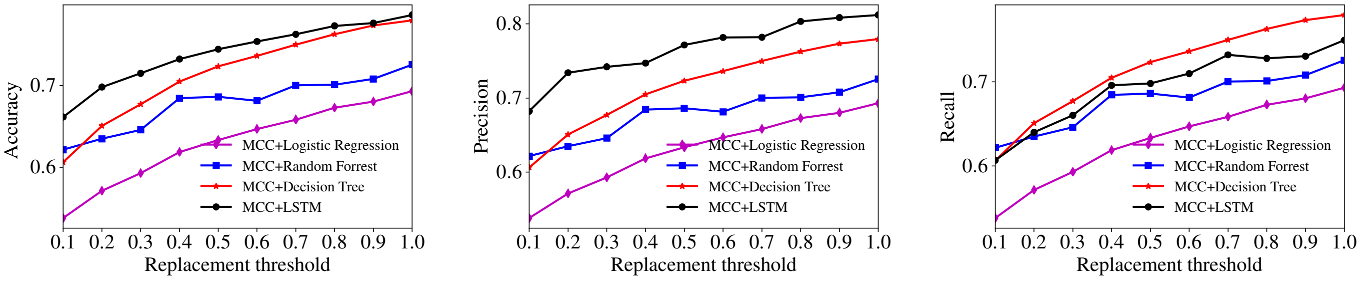
[12] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A probabilistic programming language," *Journal of statistical software*, vol. 76, no. 1, 2017.

[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



(a) Accuracy

(b) Precision

(c) Recall

Fig. 17. Evaluation metrics for four different methods with respect to the replacement probability in negative procedure sampling on outpatient claims.

TABLE VI

p-VALUES FROM PAIRWISE COMPARISONS USING NEMENYI MULTIPLE COMPARISON TEST FOR UN-REPLICATED BLOCKED DATA AMONG DIFFERENT METHODS. $p < 0.05$ DETERMINES SIGNIFICANCE. (LR=LOGISTIC REGRESSION, DT=DECISION TREE, RF=RANDOM FOREST)

	MCC	MCC+RPCA	MCC+LSTM	MOD	Baseline	MCC+LR	MCC+RF	MCC+DT
MCC	1	0.697	0.138	0.9	0.087	0.9	0.642	0.008
MCC+RPCA	0.697	1	0.001	0.9	0.9	0.751	0.016	0.001
MCC+LSTM	0.138	0.001	1	0.040	0.001	0.111	0.9	0.9
MOD	0.9	0.9	0.040	1	0.254	0.9	0.358	0.001
Baseline	0.087	0.9	0.001	0.254	1	0.111	0.001	0.001
MCC+LR	0.9	0.751	0.111	0.9	0.111	1	0.587	0.006
MCC+RF	0.642	0.016	0.9	0.358	0.001	0.587	1	0.587
MCC+DT	0.008	0.001	0.9	0.001	0.001	0.006	0.587	1

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-learning-based 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," https://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. Candès, 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-and-Systems/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*, 2018.

Md Enamul Haque is currently working as a Research Data Scientist at Stanford University School of Medicine, Palo Alto, CA, USA. He earned his Ph.D. degree in Computer Science at the School of Computing and Informatics, University of Louisiana at Lafayette, LA, USA. He completed M.Sc. in Computer Engineering from King Fahd University of Petroleum and Minerals, Saudi Arabia in 2015. He received B.Sc. in computer science from the Islamic University of Technology, Bangladesh. He worked as a Data Scientist at American Family Insurance, Chicago, IL, USA. He also worked as a Software Engineer at Grameenphone Ltd., Bangladesh. His research interest includes complex networks, graph analytics, and machine learning.

Mehmet Engin Tozal is an assistant professor in the School of Computing and Informatics at The University of Louisiana at Lafayette and a member of the Informatics Program. He received Ph.D. degree in Computer Science from The University of Texas at Dallas in 2012. His ongoing research involves analyzing, modeling and sampling real world complex systems including network topologies, social networks and information networks. He also works on designing secure and statistically reliable network protocols for extreme environments.