Health Risk Prediction Using Big Medical Data - a Collaborative Filtering-Enhanced Deep Learning Approach

Xin Li
Computer Science Department
North Dakota State University
Fargo, ND, USA
xin.li.2@ndsu.edu

Juan Li
Computer Science Department
North Dakota State University
Fargo, ND, USA
j.li@ndsu.edu

***Abstract*—The massive amount of medical data accumulated from patients and healthcare providers has become a vast reservoir of knowledge source that may enable promising applications such as risk predictive modeling, clinical decision support, disease or safety surveillance. However, discovering knowledge from the big medical data can be very complex because of the nature of this type of data: they normally contain large amount of unstructured data; they may have lots of missing values; they can be highly complex and heterogeneous. To address these challenges, in this paper we propose a Collaborative Filtering-Enhanced Deep Learning approach. In particular, we estimate missing values based on patients' similarity, i.e., we predict one patient's missing features based on the values of similar patients. This is implemented with the Collaborative Topic Regression method, which tightly couples topic model and probability matrix factorization and is able to utilize the rich information hidden in the data. Then a deep neural network-based method is applied for the prediction of health risks. This method can help us handle complex and multi-modality data. Extensive experiments on a real-world dataset show improvements of our proposed algorithm over the state-of-the-art methods.**

***Index Terms*—deep neural network, medical data, topic model, Bayesian probability matrix factorization, health risk prediction, diabetes**

I. Introduction

Vast quantities of medical data have been created by the mass adoption of the digitization of all sorts of information, such as clinical data (physician's notes, prescriptions, laboratory, medical images), electronical patient records (EPRs), Internet of Things (IoT) generated real-time big data (such as vital signs), and social media data. To explore the opportunities and challenges introduced by the growing abundance of digital data captured during the delivery of health care, artificial intelligence-based approaches have been used to analyze the big medical/healthcare data to revolutionize the delivery of care, drive new medical discoveries, improve patient outcomes, and optimize practice.

However, to use the big medical data, researchers must address several important challenges of healthcare data: (1) A significant amount of medical data may contain unstructured data, such as text document, audio voice, and email messages. Unstructured data typically requires a human touch to read, capture and interpret properly. Traditional classification/ regression tools have difficulty to use this part of the data in analysis. (2) Medical data are normally highly complex and heterogeneous. Lots of important clinical conditions are also poorly understood or associated with complex, multi-factorial underlying pathologies. A major challenge of analyzing big medical data is the generation of appropriate feature set from various large number of complex features without human intervention. (3) Medical datasets are typically sparse, and may have large volumes of missing variables for each patient, due to various human factors, for instance, human subjects occasionally miss their blood glucose testing.

To address the aforementioned three challenges, in this paper, we propose a novel approach that seamlessly integrates deep neural network with collaborative filtering to realize a personalized risk prediction. Deep neural network is employed in our approach because of its advanced capability on integrated feature learning, and handling complex and multi-modality data [6]. The proposed approach utilizes multiple layers of many hidden neurons of deep neural network to learn data representations or features with multiple levels of abstraction. By using the backpropagation algorithm, deep learning can detect complex structure of big medical data, and then automatically adjust its internal parameters used to compute the representation in each layer from the representation in the previous layer [6]. It can generate features that are more sophisticated and difficult to elaborate in human descriptive means from the complex medical data set.

To enable deep learning to better interface with healthcare/ medical data, our approach needs to address the other two challenges relating to the characteristics of healthcare data (i.e. sparse and unstructured). For this reason, we employed collaborative filtering (CF) [17] as an assistance. CF has been widely used to predict a person's preferences based on other similar persons' preferences. The rationale behind this technique is that users with common preferences are

This work was supported by the National Science Foundation (NSF) with award number: 1722913.

more likely to have additional common preferences. CF-based technologies (such as [16]) can handle very sparse dataset pretty well, and they are also capable of utilizing and processing unstructured data.

Inspired by these advantages of CF and by the analogy between predicting users' preference to predicting patients' health risk, we use CF techniques to assist the health risk prediction. In particular, we match patients and patient characteristics to adverse outcomes like traditional CF finds similarities between users and items. To overcome the data sparsity problem which conventional CF-based methods may suffer, auxiliary information such as the info stored in the unstructured data (patient's demographic data, social economic data) will be utilized. This information will help us uncover unexpected relationships. We choose collaborative topic regression (CTR) [16] approach to tightly couple the patients' disease and patients demographical, social economical information. CTR is a probabilistic model combining topic model and probabilistic matrix factorization (PMF). Since the CTR can utilize the hidden information in the unstructured data and estimate missing data based on patients' similarity, we integrate the CTR into the deep learning to enhance the efficiency and accuracy of deep neural network. By integrating the combined strength of CRT-based CF with deep learning, we expect to overcome the shortcomings of existing approaches and provide a more efficient result.

In this paper, we apply our proposed methodology to predict risk of hospital readmission for diabetic patients. Diabetes is one of the most common and costly chronic diseases. An estimated 23.1 million people in the United States are diagnosed with diabetes at a cost of more than $245 billion per year. [5]. The largest components of medical expenditures of diabetes care are hospital inpatient care (43% of the total medical cost) [13]. With about 25% patients being readmitted within 30 days of discharge [1], unplanned sessions have become a serious issue. As diabetes patient readmission rate is becoming one of the major concerns for many national hospitals in the U.S., it is of great significance to study the possibility and risks of diabetes patients' readmission. Hospital readmission is a high-priority health care quality measure and target for cost reduction, particularly readmission rate within 30 days of discharge (30-day readmission, aka early readmission). This has made healthcare professionals, scientists, and policymakers increasingly focus on the 30-day readmission rates to determine the complexity of patient populations, prepare for procedures and interventions, and eventually improve healthcare quality and reduce cost. Despite the broad interest in early readmission rate, relatively little research effort has been applied specifically on readmission of patients with diabetes.

Although this paper's experiments focus on diabetes case, we believe that the proposed methodology would be general enough to be used more broadly.

The rest of the paper is organized as follows. We review the related work in Section II. We present our proposed collaborative filtering-enhanced deep learning algorithm in Section III. Experimental results and detailed analysis of the results are brought in Section IV. We conclude our work in Section V.

II. RELATED WORK

Nowadays the healthcare sector has generated a huge volume of patient data. Machine learning provides a way to automatically find patterns and making predictions about data. There are quite a few works in recent years on data analytics with different types of digital patient data.

To deal with the complex feature selection problem, deep learning has been used as an analysis approach. As pointed out in [14], deep learning has been used for health informatics to automatically generate optimized high-level features and semantic interpretation from the input data. For example, Cheng et al. [4] proposed a deep learning-based prediction model using Electronic Health Records (EHRs). In this model, EHRs for patients are represented as a temporal matrix with a time dimension and an event dimension. Then a four-layer convolutional neural network (CNN) is used to extract features and assess risks.

In another work, to make healthcare decisions Liang et al. [7] applied a deep learning model to EHRs database to enhance feature representations. They proposed to apply deep belief network (DBN) for unsupervised feature extraction, and then perform supervised learning through a standard support vector machine (SVM). In their work, Nie et al. [12] proposed a deep learning scheme to infer people's possible diseases given the questions of that people asks online. The proposed scheme constructs a sparsely connected deep architecture with three hidden layers. These three layers is constructed via alternative signature mining and pre-training in an incremental way: First, medical signatures are discovered from raw features. Then the raw features and their signatures are acted as input nodes in one layer and hidden nodes in the subsequent layer. Inter-relations between these two layers can be learned. Following that hidden nodes will be served as raw features for the more abstract signature mining. This process repeats until the model is well tuned. More applications of deep learning can be found in medical informatics and public health domains [3], [8], [10], [11], [15].

Various approaches have been studied to address the data sparsity problem of medical data. For example, to process sparse and non-vector input data, Wang et al. [18] proposed a high order extension of sparse logistic regression model, MulSLR, (for Multilinear Sparse Logistic Regression) to predict clinical risk. Their approach solves K classification vectors instead of solving one classification vector as in conventional logistic regression.

To determine patient acuity using incomplete, sparse and heterogeneous clinical data, Ghassemi et al. [19] proposed an approach that transforms this clinical data into a new latent space using the hyperparameters of multi-task GP (MTGP) models. In this way, patients can be compared based on their similarity in the new hyperparameter space. Information in this hyperparameter space could be viewed as timeseries data, and abstracted features can represent the series dynamics. This approach has been approved to increase classification performance on mortality prediction of ICU patients, however, the computational cost of this approach is very high.

Lipton et al. [9] propose an approach to model missing clinical data using recurrent neural networks (RNNs). Unlike classical approaches that treat missing data via heuristic imputation, in their approach, the authors model missingness as a feature. The proposed RNN can use only simple binary indicators for missingness.

GRU-D [2], a deep learning models, was developed to exploit the missing patterns of missing data for effective imputation and improving prediction performance. GRU-D was based on Gated Recurrent Unit (GRU), a recurrent neural network. GRU-D takes two representations of missing patterns, i.e., masking and time interval, and incorporates them into a deep model architecture. This model not only captures the long-term temporal dependencies in time series, but also utilizes the missing patterns to achieve better prediction results.

In spite of the numerous efforts and achievements of existing research in analyzing digital medical data, analyzing medical data is still an important and challenging task. Accurate and efficient risk prediction has always been an important topic attracting many researchers' interests.

III. Methodology

In this section we present the detailed methodology of our proposed Collaborative Filtering-enhanced Deep Learning (CFDL) approach .

A. Problem formulation

Let $\mathcal{X} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), (\mathbf{x}_3, y_3), \ldots, (\mathbf{x}_P, y_P)\}$ be the data set, where P is the number of patients; $\mathbf{x}_i \in \mathbb{R}^d$ is the i-th instance; and $y_i \in \{1, \ldots, Q\}$ represents its label, where $Q \geqslant 2$ is the number of classes. $(\mathbf{x}_i)_j$ is the j-th feature of patient i; for example, it can be one of a patient's demographic features, or disease symptoms, or vital signs. y_i represent the target label, for example, it can be a particular health risk.

To model the missing data, we divide the data set $\mathcal{X}$ into an observed component $\mathcal{X}^o$ and a missing component $\mathcal{X}^m$. Similarly, for some data vector, $\mathbf{x}_i$ is divided into $(\mathbf{x}_i^o, \mathbf{x}_i^m)$ where these data vector may have different missing components. Note that $\mathbf{x}_i^m$ should not be a subset of $\mathbf{x}_i^o$. For some data vector, the whole data instance is lost, hence $\mathbf{x}_i = \mathbf{x}_i^m$.

To model unstructured data, in our case text data (such as doctor's notes, open-question survey), we can model them as a corpus D of M documents $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$ consisting of unstructured data, with a vocabulary of size V.

Given these data, we have two goals. Firstly, we aim to learn the missing data of $\mathcal{X}^m$ from the observed data $\mathcal{X}^o$ and unstructured data of M documents $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$. Secondly, we want to predict the estimated label for the dataset. To realize these two goals, we propose a two-stage methodology presented in the following subsections.

B. Stage 1: identify missing data

The goal of this stage is to learn the missing data $\mathcal{X}^m$ from the existing observed and unstructured data. The main idea is that if an important feature is missing for a particular instance, it can be estimated from similar data that are present. Consider we have the unstructured data, we hypothesize that there is rich information within the unstructured data, which can provide additional source of information for the estimation.

We apply the topic model on the unstructured data of M documents. To do that, we assume that there exist a fixed number of latent topics that appear across multiple documents. Each topic is characterized by a multinomial distribution over the vocabulary of the corpus D, drawn from a Dirichlet distribution denoted as $\phi_k \sim Dir(\beta)$. Each document is characterized by a multinomial distribution over the set of topics in the corpus, which also assumed have a Dirichlet prior denoted as $\theta_j \sim Dir(\alpha)$. The topic distribution has the following probability density:

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}, \tag{1}$$

where the parameter α is a k-vector with components $\alpha_i > 0$, and where $\Gamma(x)$ is the Gamma function.

We follow the CF-based recommendation algorithm, the matrix of patient and health-related features can be decomposed by Matrix Factorization (MF), by denoting a set of values correspond to a set of patients with indices $i \in \{1, 2, \ldots, N\}$ to a set of health factors with indices $j \in \{1, 2, \ldots, M\}$, each entry of the value R_{ij} can be expressed as the inner product of a patient matrix and a health factor matrix:

$$R_{ij} \approx < U_i, V_j > \equiv \sum_{k=1}^{K} U_{ik} V_{kj} \tag{2}$$

where U_i , V_j represent the k-dimensional patient-specific and health factor-specific latent feature vectors of patient i and factor j, respectively.

Collaborative topic regression (CTR) additionally molds the health factor vector as the combining of a latent variables that offsets the topic proportion. According to CTR, for each v_j

$$v_j = \varepsilon_j + \theta_j \tag{3}$$

where θ_j is the topic proportion computed by latent dirichlet allocation (LDA), therefore the values can be expressed as:

$$R_{ij}^t \approx < U_i, \varepsilon_j + \theta_j > \equiv \sum_{k=1}^{K} U_{ik}(\varepsilon_{kj} + \theta_{kj}) \tag{4}$$

CTR places prior distribution on $\mathbf{U}$, $\mathbf{V}$. Specifically zero-mean, independent Gaussian priors are imposed on patient and health factor vector:

$$\begin{aligned} U_i &\sim \mathcal{N}(0, \sigma_u^2 \mathbf{I}), \text{ i = 1 \ldots N}, \\ V_j &\sim \mathcal{N}(0, \sigma_v^2 \mathbf{I}), \text{ j = 1 \ldots M} \end{aligned} \tag{5}$$

where $\mathbf{I}$ is the D−by−D identity matrix.

The conditional distribution over the observed ratings and the prior distributions are given by

$$R_{ij}|\mathbf{U},\mathbf{V} \sim \mathcal{N}(\,U_i,\,\varepsilon_j+\theta_j,\,\alpha^{-1}) \tag{6}$$

In this way, the original complex matrix can be decomposed into simpler computations involving the corresponding patient related matrix and health factor related matrix, as well as the topic proportions.

CTR combines the matrix factorization with the topic modeling to collaboratively predict values and to learn topics. After we have trained our LDA implementation on a separate training corpus and learned the model parameters α and β, the whole model can be learned by maximizing the log-posterior distribution, which takes the following form assuming ratings are made independently conditioned on latent factors:

$$\begin{aligned}
&\ln p(U,V|R,\sigma_R^2,\sigma_U^2,\sigma_V^2)\\
&=\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}^t \ln p(R_{ij}^T|U_i,V_j)+\sum_{i=1}^{N}\ln p(U_i)\\
&+\sum_{j=1}^{m}\ln p(V_j-\theta_j)\\
&=-\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{I_{ij}^t(R_{ij}^T-<U_i,V_j>)^2}{2\sigma_R^2}\\
&-\frac{1}{2}\left(\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}^t\right)\ln\sigma_R^2\\
&-\frac{1}{2\sigma_U^2}\sum_{i=1}^{N}U_i^TU_i-\frac{1}{2\sigma_V^2}\sum_{j=1}^{M}(V_j-\theta_j)^T(V_j-\theta_j)\\
&+\sum_{j=1}^{M}\sum_{s=1}^{W(j)}\ln(\sum_{k=1}^{K}\theta_{jk}\beta_{k,w_{js}})-\ln\sigma_0+C
\end{aligned} \tag{7}$$

where C is a constant that does not depend on the parameters. Maximizing the log-posterior over the latent features with hyperparameters (i.e. the observation noise variance and prior variances) kept fixed is equivalent to minimizing the following sum-of-squared-errors objective functions with quadratic regularization terms:

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{c_{ij}^t(r_{ij}^t-<U_i,V_j>)^2}{2}+\frac{\lambda_u}{2}\sum_{i=1}^{N}U_i^TU_i\\
&+\frac{\lambda_v}{2}\sum_{j=1}^{M}(V_j-\theta_j)^T(V_j-\theta_j)
\end{aligned} \tag{8}$$

where $\lambda_u=\sigma_R^2/\sigma_U^2$, $\lambda_v=\sigma_R^2/\sigma_V^2$, and $\lambda_0=\sigma_R^2/\sigma_0^2$ and Dirichlet prior (α) is set to 1. Note that c_{ij}^t is the confidence parameter for rating r_{ij}^t.

After the parameters have been learned, we can do the prediction for the missing data. For one data instance if only some parts of the feature data are missing, the *point estimate* can be used to approximate their expectations as:

$$\mathrm{E}[r_{ij}^t|X]\approx\mathrm{E}[u_i^t|X](\mathrm{E}[\theta_j^t|X]+\mathrm{E}[\varepsilon_j^t|X]) \tag{9}$$

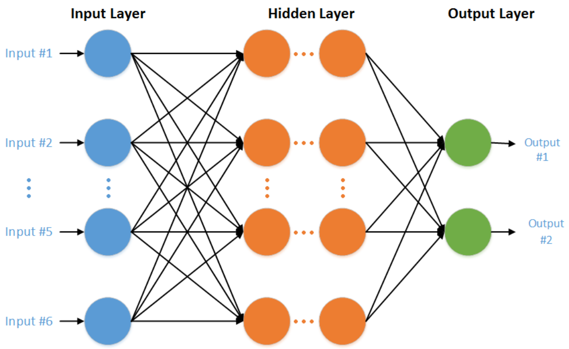

Fig. 1. The deep neural network

$$r_{ij}^{t\ *}\approx(u_i^*)^T v_j^* \tag{10}$$

If the whole data instance is lost, $\mathrm{E}[\varepsilon_j^t|X]=0$ and the missing values can be predicted as:

$$\mathrm{E}[r_{ij}^t|X]\approx\mathrm{E}[u_i^t|X](\mathrm{E}[\theta_j^t|X]) \tag{11}$$

$$r_{ij}^{t\ *}\approx(u_i^*)^T\theta_j^* \tag{12}$$

Using this approach, the missing data can be estimated and added back to the dataset for further processing.

C. Stage 2: deep neural network training and prediction process

After the missing values have been filled by CTR, deep neural network can be used for the model prediction. Fig. 1 describes the architecture of the deep neural network that we used for classification. The system does not need any complicated syntactic or semantic preprocessing. The feature vector is fed into the input nodes of the network. Each node generates an output with an activation function, and the linear combinations of the outputs are linked to the next hidden layers. The activation functions among different layers are different. The training data is defined as $\{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),(\mathbf{x}_3,y_3),\ldots,(\mathbf{x}_P,y_P)\}$ of P samples. $\mathbf{x}$ is the input feature vector, y is the class label information. In succession, the features are respectively extracted and then directly concatenated to form the final feature vector. Finally, to compute the confidence of each relation, the feature vector is fed into a softmax classifier. The output of the classifier is a vector, the dimension of which is equal to the number of predefined classification types. The value of each dimension is the confidence score of the corresponding classification.

In the training process, the input feature goes through the input nodes at the bottom of the deep learning network, where the weights are initialized with random values. Thereafter the weight vectors are fine-tuned in sequence. The training goal of this process is to minimize a comprehensive cost function given as the mean squared error function between the prediction value and the real output value:

$$C(\mathbf{w};\mathbf{x},y)=\|h_w(\mathbf{x})-y\| \tag{13}$$

where $\mathbf{w}$ is the set of the weights in the deep learning network, which needs to be trained in this phase, y is the label, $h_w(\mathbf{x})$

is a hypothesis function which will yield an estimated output, and $\|\bullet\|$ denotes the Frobenius norm.

The overall cost function for a batch training is defined as:

$$J(\mathbf{w}) = \frac{1}{P}\sum_{p} C(\mathbf{w}; \mathbf{x}_p, y_p) \tag{14}$$

We want to obtain the optimal parameter set to achieve the minimization of the objective function as:

$$\mathbf{W}^* = \arg\min_{\mathbf{w}} J(\mathbf{w}) \tag{15}$$

This can be achieved by the back propagation algorithm. In the back propagation algorithm, we use the stochastic gradient method to update the weight vectors from the top layer to the bottom layer as

$$w_{ji}^{n} = w_{ji}^{n-1} + \eta \frac{\partial}{\partial w_{ji}^{n-1}} J(\mathbf{w}) \tag{16}$$

where η is an adaption parameter.

After the neural network structure and the weighted parameter are determined, the deep neural network can make prediction directly.

D. Algorithm design and implementation

Given the number of i for patients, j for health-related factors and K for topics, the proposed two-stages algorithm is summarized as:

Algorithm 1 Algorithm for CFDL

```
Input: The training dataset in
Output: The architecture and weights for DNN out
  Initialization : Initialize the network weights to random number
  for each patient $i$ do
    Draw patient latent vector $u_i \sim \mathcal{N}(0,\ \lambda_u^{-1} I_K)$
  end for
  for each health-related item $j$ do
    Draw topic proportions $\theta_j \sim Dirichlet(\alpha)$
    Draw item latent offset vector $\varepsilon_j \sim \mathcal{N}(0,\ \lambda_v^{-1} I)$ , and set the item latent vector as $v_j = \varepsilon_j + \theta_j$
    for each word $w_{jn}$ do
      Draw topic assignment $z_{jn} \sim Mult(\theta)$
      Draw word assignment $w_{jn} \sim Mult(\beta_{z_{jn}})$
    end for
  end for
  for each patient-health-related item pair $(i, j)$ do
    Draw the rating $x_{ij} \sim \mathcal{N}(\mathrm{u}_i{}^T \mathrm{v}_j,\ \mathrm{c}_{ij}^{-1})$, where $c_{ij}$ is a confidence parameter for rating $x_{ij}$, $a > b$ . $c_{ij} = a$. (higher confidence), if $x_{ij} = 1$ and $c_{ij} = b$, if $x_{ij} = 0$.
  end for

  while the iteration number is greater than 0 do
    for each of the input data
    $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), (\mathbf{x}_3, y_3), \ldots, (\mathbf{x}_P, y_P)\}$ do
      Form the vector of input, run the network forward with the input data to get the network output
    end for
    for each output node do
      Compute $\|h_w(\mathbf{x}) - y\|$
    end for
    for each of the layer do
      Update the weight as $w_{ji}^{n} = w_{ji}^{n-1} + \eta \frac{\partial}{\partial w_{ji}^{n-1}} J(\mathbf{w})$
    end for
  end while
  return the network structure and weights
```

IV. Experiments

In this section, we present our initial efforts on evaluating the proposed methodologies.

A. Datasets

We applied our proposed methodology on the UCI dataset [20]. This dataset contains real medical records of patients diagnosed with diabetes, collected over a period of 10 years (1999-2008) from 130 hospitals in USA. The medical records of each patient included 55 different attributes and a label indicating whether the patient was readmitted to the hospital within 30 days, after 30 days, or never readmitted. The distribution is as follows - 11% of patients were readmitted within 30 days, 35% after 30 days and 54% patients were never readmitted. In total, there were 101,765 encounters available for analysis that satisfy these criteria. Each encounter was labeled with one of three classes (<30, >30, NO) based on whether the patient was readmitted within 30 days (<30), readmitted in more than 30 days (>=30), or did not have a recorded readmission (NO).

B. Pre-processing

For the 10173 patients, they have 24 medicine features, which have four kinds of discrete *values*, *No*, *Up*, *Down*, and *Steady*. we map these values to numerical values from 1 to 4 respectively. To verify the performance of our proposed method on dealing with missing values, we intentionally removed 20% of these medicine values. We molded the medicine feature matrix as a collaborative filtering problem with 10173 users and 24 items. Therefore, in this stage, 1,791,064 points of data were used for training and 447,766 points of data were used for prediction of the CTR model. As there is no unstructured data in our dataset, our model has been degenerated to a PMF (probability matrix factorization) problem.

To avoid over-fitting the model, we removed some of the attributes in the original dataset such as a patient's Encounter ID and patient number. The features we used are summarized in Table I. Some of the features were processed. For example, age is divided into 10 ranges and is calculated as the mean of the interval. The target variable (class label) is readmission or not. It has three levels: '<30' (patient is readmitted within 30 days), '>30' (patient is readmitted after 30 days) and 'no' (patient was not readmitted). In this paper, the target variable has been re-coded to '1' (patient is readmitted) and 0 (patient

TABLE I
LIST OF FEATURE DESCRIPTIONS

Feature name	Type
Gender	Nominal
Race	Nominal
Age	Nominal
Weight	Numeric
Medical specialty	Nominal
Time in hospital	Numeric
Number of lab procedures	Numeric
Number of procedures	Numeric
Number of medications	Numeric
Number of outpatient visits	Numeric
Number of emergency visits	Numeric
Number of inpatient visits	Numeric
Diagnosis 1	Nominal
Diagnosis 2	Nominal
Diagnosis 3	Nominal
Number of diagnoses	Numeric
Glucose serum test result	Nominal
A1c test result	Nominal
Change of medications	Nominal
Diabetes medications	Nominal
24 features for medications	Nominal

is not readmitted) as a binary variable.

C. Training and Testing

We train a five-layer deep neural network with 1 input layer, 1 output layer and 3 hidden layers. The output layer's active function is *softmax*. Each layer consists of some neurons. The number of neurons of the input layer is the same as the input feature dimension, and the number of neurons of the output layer is the same as the output classes. For the neurons in the input layer, they receive a single value on their input and are sent to all of the hidden nodes. The nodes of the hidden and output layers are active, and each layer is fully interconnected. The output layer is responsible for producing and presenting the final network outputs, which are generated from the procedure performed by neurons in the previous layers.

In the deep learning prediction stage, we split the dataset into two parts: a training dataset (70%) and a testing dataset (30%). To conduct the experiment, we take an average of 10 runs by shuffling the dataset. We have used 10 folds cross validation by subdividing the original dataset. For all these folds we got similar results which indicates the stability of the score.

D. Result and discussion

To measure the performance of the proposed algorithm, CFDL, we compare it with the state-of-the-art classification approaches, namely Support Vector Machine (SVM), Decision Tree, and Naive Bayes. The result is represented based on the most commonly used metrics: *accuracy*, *precision*, *recall*, *micro-averaged F1*, *macro-averaged F1*, which are defined as follows:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{17}$$

$$precision = \frac{TP}{TP + FP} \tag{18}$$

$$recall = \frac{TP}{TP + FN} \tag{19}$$

$$micro - avg.F1 = \frac{TP}{num} \tag{20}$$

$$macro - avg.F1 = \frac{2 \times recall \times precision}{recall + precision} \tag{21}$$

TABLE II
CONFUSION MATRIX

		Prediction outcome		
		p	**n**	**total**
actual value	**p′**	True Positive	False Negative	P′
	n′	False Positive	True Negative	N′
	total	P	N	

All these measurements are being calculated using the confusion matrix in Table II based on two possible outcomes: positive (p) and negative (n).

Summaries of these performance measurements for the above-mentioned approaches can be found in Table III and IV. Table III shows the prediction results using the original UCI dataset, while Table IV presents the prediction results using a dataset that were intentionally deleted 20% of the random selected medication data from the UCI dataset. For both scenarios, CFDL is the winner among all of these comparisons. The findings in these two tables clearly indicate that the proposed approach CFDL outperforms other approaches in almost all of the five metrics.

Although the experiments performed on the UCI dataset demonstrate the good performance of our proposed approach, we expect that the proposed method would performs even better compared with other approaches on more complex, heterogeneous, and unstructured data, as our algorithm was designed to overcome the challenges of such dataset.

V. CONCLUSION

In this paper, we have presented a novel approach to analyze big complex medical data to make accurate predictions. In particular, we propose a collaborative filtering-enhanced deep learning algorithm, CFDL. This algorithm can effectively utilize unstructured data and find hidden relationship from

TABLE III
FPERFORMANCE MEASUREMENTS FOR DIFFERENT PREDICTION METHODS - ORIGINAL DATASET

Approach	Micro-avg. F1	Macro-avg. F1	Accuracy	Precision	Recall
CFDL	0.107	0.64	0.8894	0.5018	0.9637
SVM	0.077	0.593	0.8706	0.448	0.875
Decision Tree	0.087	0.35	0.6774	0.2258	0.7788
Naive Bayes	0.077	0.263	0.5693	0.1629	0.6908

TABLE IV
FPERFORMANCE MEASUREMENTS FOR DIFFERENT PREDICTION METHODS- 20% MISSING DATA

Approach	Micro-avg. F1	Macro-avg. F1	Accuracy	Precision	Recall
CFDL	0.087	0.568	0.869	0.448	0.7788
SVM	0.098	0.477	0.78	0.328	0.875
Decision Tree	0.087	0.292	0.5756	0.1796	0.787
Naive Bayes	0.096	0.327	0.634	0.216	0.867

these data; it can accurately estimate missing values of the data set; and finally it can integrated feature learning, and handling complex and multimodality data.

We evaluated the proposed method using a real-world diabetes readmission dataset. We observed that the proposed method outperforms many other approaches. As we explained, due to the limitation of the dataset the advantages of our algorithm cannot be fully demonstrated with the experiments. In the future, we will continue to evaluate and improve the proposed algorithm based on more complex and larger-scale dataset. We also plan to add more context dimensions (such as time) to the algorithm.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation (NSF) under Div. of Information and Intelligent Systems (IIS) with award number: 1722913.

REFERENCES

[1] Kristin E. Bergethon, Christine Ju, Adam D. DeVore, N. Chantelle Hardy, Gregg C. Fonarow, Clyde W. Yancy, Paul A. Heidenreich, Deepak L. Bhatt,Eric D. Peterson, and Adrian F. Hernandez. Trends in 30-Day Readmission Rates for Patients Hospitalized with Heart Failure: Findings from the Get with the Guidelines-Heart Failure Registry. *Circulation: Heart Failure*, 9(6), 2016.

[2] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent Neural Networks for Multivariate Time Series with Missing Values. pages 1–14, 2016.

[3] Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. Distilling Knowledge from Deep Networks with Applications to Healthcare Domain. pages 1–13, 2015.

[4] Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. Risk Prediction with Electronic Health Records: A Deep Learning Approach. *SIAM International Conference on Data Mining*, pages 432–440, 2016.

[5] Centers for Disease Control, Prevention, and Others. National Diabetes Statistics Report: Estimates of Diabetes and Its Burden in the United States. Atlanta, GA: Centers for Disease Control and Prevention; 2014. *US Department of Health and Human Services*, (Cdc):2009–2012, 2017.

[6] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[7] Znaonui Liang, Gang Zhang, Jimmy Xiangji Huang, and Qmming Vivian Hu. Deep learning for healthcare decision making with EMRs. *Proceedings - 2014 IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2014*, (Cm):556–559, 2014.

[8] Zachary C. Lipton, David C. Kale, Charles Elkan, and Randall Wetzel. Learning to Diagnose with LSTM Recurrent Neural Networks. pages 1–18, 2015.

[9] Zachary C. Lipton, David C. Kale, and Randall Wetzel. Modeling Missing Data in Clinical Time Series with RNNs. 56, 2016.

[10] Saaed Mehrabi, Sunghwan Sohn, Dingheng Li, Joshua J. Pankratz, Terry Therneau, Jennifer L. St Sauver, Hongfang Liu, and Mathew Palakal. Temporal Pattern and Association Discovery of Diagnosis Codes Using Deep Learning. *Proceedings - 2015 IEEE International Conference on Healthcare Informatics, ICHI 2015*, pages 408–416, 2015.

[11] Riccardo Miotto, Li Li, Brian A. Kidd, and Joel T. Dudley. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports*, 6(January):1–10, 2016.

[12] Liqiang Nie, Meng Wang, Luming Zhang, Shuicheng Yan, Bo Zhang, and Tat Seng Chua. Disease Inference from Health-Related Questions via Sparse Deep Learning. *IEEE Transactions on Knowledge and Data Engineering*, 27(8):2107–2119, 2015.

[13] Matt Petersen. Economic costs of diabetes in the U.S. in 2012. *Diabetes Care*, 39(7):1033–1046, 2016.

[14] Daniele Ravi, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang Zhong Yang. Deep Learning for Health Informatics. *IEEE Journal of Biomedical and Health Informatics*, 21(1):4–21, 2017.

[15] Shin, Hoo-chang, Le Lu and Ronald M. Summers. Interleaved Text / Image Deep Mining on a Large-Scale Radiology Database. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*, 17:1090–1099, 2015.

[16] Purushotham Sanjay, Yan Liu, and C-C. Jay Kuo. Collaborative topic regression with social matrix factorization for recommendation systems. *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 17:691–698, 2012.

[17] Xiaoyuan Su and Taghi M. Khoshgoftaar. A Survey of Collaborative Filtering Techniques. *Advances in Artificial Intelligence*, 2009(Section 3):1–19, 2009.

[18] Fei Wang, Ping Zhang, Buyue Qian, Xiang Wang, and Ian Davidson. Clinical risk prediction with multilinear sparse logistic regression. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*, pages 145–154, 2014.

[19] Marzyeh Ghassemi, Marco A.F. Pimentel, Tristan Naumann, Thomas Brennan, David A. Clifton, Peter Szolovits, and Mengling Feng. A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data. *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Conference on Artificial.

[20] "UCI Machine Learning Repository: Diabetes 130-US hospitals for years 1999-2008 Data Set", Archive.ics.uci.edu, 2018. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008. [Accessed: 15- Apr- 2018].