Predicting Hospital Readmission via Cost-Sensitive Deep Learning

Article *in* IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM · April 2018

DOI: 10.1109/TCBB.2018.2827029

Arbi Ben Abdallah
Washington University in St. Louis
134 PUBLICATIONS **4,576** CITATIONS

Cost-sensitive Deep Learning for Early Readmission Prediction at A Major Hospital

Haishuai Wang†, Zhicheng Cui†, Yixin Chen†, Michael Avidan‡, Arbi Ben Abdallah‡, Alexander Kronzer‡
† Department of Computer Science and Engineering, Washington University in St. Louis
‡ School of Medicine, Washington University in St. Louis
haishuai.wang@wustl.edu,zhicheng.cui@wustl.edu,chen@cse.wustl.edu,
avidanm@wustl.edu,aba@wustl.edu,akronzer@wustl.edu

ABSTRACT

With increased use of electronic medical records (EMRs), data mining on medical data has great potential to improve the treatment quality of hospitals and increase the survival rate of patients. Early readmission prediction enables early intervention, which is essential to preventing serious or life-threatening events, and act as a substantial contributor to reducing healthcare costs. Existing works on predicting readmission often focus on certain vital signs and diseases by extracting statistical features. They also fail to consider skewness of class labels in medical data and different costs of misclassification errors. In this paper, we recur to the merits of convolutional neural networks (CNN) to automatically learn features from vital sign time series, and one-hot encoding to effectively extend feature vectors with heterogeneous clinical features, such as demographics, hospitalization history, vital signs and laboratory tests. Then, both learnt features via CNN and statistical features via one-hot encoding are fed into a multilayer perceptron (MLP) for prediction. We use a cost-sensitive formulation to train MLP during prediction to tackle the imbalance and skewness challenge. We validate the proposed approach on two real medical datasets. Data are from Barnes-Jewish Hospital, and all data in these datasets is taken from historical EMR databases and reflects the kinds of data that would realistically be available at the clinical prediction system in hospitals. We find that early prediction of readmission is possible and when compared with state-of-the-art existing methods used by hospitals, our methods perform significantly better. Based on these results, a system has been deployed in hospital settings with the proposed forecasting algorithms to support treatment.

KEYWORDS

Readmission Prediction, Deep Learning; Electronic Medical Records; Cost-sensitive

KDD'17, Halifax, Canada
© 2017 Copyright held by the owner/author(s). 123-4567-24-567/08/06. . . $15.00
DOI: 10.475/123_4

ACM Reference format:
Haishuai Wang†, Zhicheng Cui†, Yixin Chen†, Michael Avidan‡, Arbi Ben Abdallah‡, Alexander Kronzer‡. 2017. Cost-sensitive Deep Learning for Early Readmission Prediction at A Major Hospital. In *Proceedings of KDD, Halifax, Canada, August 2017 (KDD'17),* 9 pages.
DOI: 10.475/123_4

1 INTRODUCTION

Big-data based predictive algorithms in medical community has been a significant research topic since the Electronic Medical Records (EMRs) captured rich clinical and related temporal information. The applications of machine learning to solve important problems in healthcare, such as predicting readmission, have the potential to revolutionize clinical care and early prevention.

Background and Significance: A hospital readmission is defined as admission to a hospital within a specified time frame after an original admission. Different time frames have been used for research purposes, the most common being 30-day, 90-day, and 1-year readmissions. Readmission may occur for planned or unplanned reasons, and at the same hospital as original or admission or a different one [7]. Readmission prediction is significant for two reasons: quality and cost of health care. Readmission reflects relatively low quality and also has negative social impacts on the patients and on the hospital [12]. Nearly 20 percent of hospital patients are readmitted within 30 days of discharge, a $15 billion problem for both patients and the healthcare system. Avoidable readmissions account for around $17 billion a year [10]. Consequently, readmission is becoming more important as an indicator for evaluating the overall healthcare effectiveness. It is significant to identify and predict readmission early in order to prevent it.

We propose to develop, validate and assess machine learning, forecasting algorithms that predict readmission for individual patients. The forecasting algorithms will be based on data consolidated from heterogeneous sources, including the patient's electronic medical record, the array of physiological monitors in the operating room and the general hospital wards, and evidence-based scientific literature.

There are some existing forecasting algorithms being used to predict readmission [2, 3, 7, 15, 16]. However, these algorithms have some shortcomings, making them inapplicable to our data sets and objectives:

1. They predict patients without considering the misprediction costs of different categories. In a readmission prediction problem where the occurred cases (minority class) are usually quite rare as compared with normal populations (majority class), the recognition goal is to detect patients with readmission. A favorable classification model is one that provides a higher identification rate on the minority class (Positive Prediction Value) under a reasonably good prediction rate on the majority class (Negative Prediction Value).

2. Time-series is commonly used in the medical domain since medical equipments record vital signs data with certain time interval. They involve extracting discriminative features from the original time series, which is ad-hoc and separates the feature extraction part with the classification part, resulting in limited accuracy performance. Consequently, existing approaches ignore temporal information, local structure and overall trend of time series.

3. They use inefficient feature encoding and limited patient characteristics are related to a certain disease. Instead of representing the features as computable sequences, they use the features directly without considering efficiency and order. Thus, an effective feature encoding method is required to improve prediction accuracy. Besides, with the increasing use of EMRs, more existing patient characteristics can result in more effective prediction [15].

4. Though our goal is to make predictions for various medical datasets, such as data from general wards or operating rooms, they have not provided an integrated clinical decision support system for hospitals to predict readmission from heterogeneous, multi-scale, and high-dimensional data.

Nowadays, deep learning has been one of the most prominent machine learning techniques [1, 17]. Deep learning aims to model high-level abstractions in the data using nonlinear transformations. Such abstractions can then be used to interpret the data, or to build better predictive models. Through stacking multiple layers, the model is able to capture much richer structures and learn significantly more complicated functions. Convolutional Neural Networks (C-NN) is reported as a successful technique for time series classification [8, 18, 19] because of its ability to automatically learn complex feature representation using its convolutional layers. Thus, CNN is able to handle time series data without requiring any handcrafted features.

We aim to apply deep learning techniques to develop better models for early predicting readmissions. At the same time, we need to consider the imbalanced or skewed class distribution problem, which yields varying costs information for different types of misclassification errors. In this paper, we present cost-sensitive deep learning models on clinical readmission prediction using data collected by monitoring d-ifferent vital signs, demographics and lab results. Specifically, we first automatically learn the feature representation from the vital signs time series using CNN, and simultaneously construct the feature vectors from discrete and continuous features (such as demographics) by one-hot encoding. Without loss of generality, we also extract statistical features from time series (such as first order and second order features) as discrete features and feed into the feature vector. Then, we combine the learned time series features from CNN and feature vector from one-hot encoding as input to a Multi Layer Perceptron (with multiple hidden layers). At the output layer, a cost-sensitive prediction formulation is used to address the imbalanced challenge. A cost-sensitive prediction can be obtained using Bayesian optimal decision based on a cost matrix. The cost matrix denotes the uneven identification importance between classes, so that the proposed approach put on weights on the learning towards the rare class associated with higher misclassification cost. The method we develop in this paper is focused on a much broader class of patients (ward patients and surgery patients), and deployed in a real system for supporting treatment and decision making. Model performance metrics are compared to state-of-the-art approaches. Our method outperforms the existing methods on real clinical data sets, and is deployed on a real system at a major hospital.

2 DATA DESCRIPTION

The work described in this paper was done in partnership with Washington University School of Medicine and Barnes-Jewish Hospital, one of the largest hospitals in the United States. We used two real data sets from Barnes-Jewish Hospital. A large database is from the general hospital wards (GHWs) between July 2007 and July 2011. GHWs gathered data from various sources, including more than 30 vital signs (pulse, shock index, temperature, heart rate etc.) from routine clinical processes, demographics, real-time bedside monitoring and existing electronic data sources from patients at the general hospital wards (GHWs) at Barnes-Jewish Hospital. Out of 41,503 patient visits, there are only 1983 patients who have readmission, which are less than 0.5%. This extremely skewed class distribution makes the prediction task very difficult.

Another data set is operating room pilot data (ORP), which is derived from heterogeneous sources, including the patient's electronic medical record, the array of physiological monitors in the operating room, laboratory tests, and evidence-based scientific literature. The ORP includes more than 40 vital signs during surgery (such as heart rate which are recorded every minute) and patients' pre-operation information such as demographics, past hospitalization history, surgery information and tests. The demographic features in our data include patients' age, gender, height, weight, race and so on. The surgery information includes surgery type, anesthesia type, and etc.

The purpose is to develop forecasting algorithms that mine and analyze the data to predict the patients' outcomes (specifically, whether or not they would be re-admitted). The forecasting algorithms will be based on data collected from general wards or operating rooms. The algorithm will facilitate patient-specific clinical decision support (such as early readmission prediction) to enable early intervention. The system is implemented and deployed in the Barnes-Jewish Hospital.

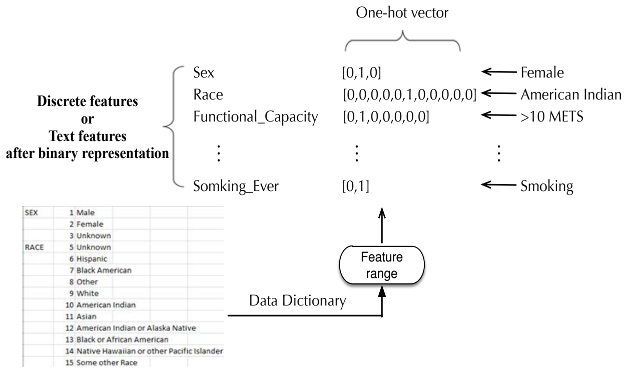

Figure 1: One-hot vector encoding of the medical data. Each discrete feature is represented as an M-dimensional vector, where one dimension is set to 1 and the rest are 0. The value of M is feature range calculated from the data dictionary. The text feature is represented in a binary format, i.e., the value is set to be 1 at the corresponding location in the text feature range, otherwise, it is 0. The text features after binary representation is then concatenated to the one-hot vector.

3 PREPROCESSING AND FEATURES

Data exploration and preprocessing, and feature extraction are critical steps for the success of any application domain. They are especially important for our clinical domain since the data are noisy, complex, and heterogeneous. Thus, prior to feature encoding, several preprocessing steps are applied to eliminate outliers and find an appropriate feature representation of patient's states.

We first preprocess the dataset by removing the outliers. The raw data typically contain many reading and input errors because information are recorded by nurses and there are inevitably errors introduced by manual operations. We list the acceptable ranges of every feature based on the domain knowledge of the medical experts in our team. Then we perform a sanity check of the data and replace the abnormal values that are outside of the feature range by the mean value of the entire population.

Second, not all patients have values for all signs in a real clinical data, and many types of clinical features involved in lab tests are not routinely performed on all patients. We use the mean value of a sign over the entire historical dataset to fill the missing values.

Finally, we normalize the data to scale the values in each bucket of every vital sign so that the data values range in the interval [0,1]. Such normalization is helpful for prediction algorithms such as deep learning and logistic regression.

A key aspect in any application of data mining is building effective features for classification or prediction. Before building our model, we first worked with the physicians from Barnes-Jewish Hospital as well as studied prior work to determine good features, since the input of our model is based on the feature embedding from raw medical data. Based on the characteristics of our data sets, we have discrete features, continuous features, text features and time series features which record the vital values at different time. Thus, the features are inapplicable to a classifier directly. In this paper, we use all the features extracted from all kinds of data in the data sets by one-hot vector encoding, at the same time, we adopt convolutional neural networks to automatically learn discriminative feature from time series data. In this way, the built features not only contain statistical information but also hold temporal and local information as well as the overall trend of time series.

One-hot vector encoding: We use the one-hot vector format to represent features in the EMRs data and make it applicable to general classifiers. Based on the feature ranges in the data dictionary, the discrete features (such as sex and race) are encoded by using one-hot encoding (as shown in Figure 1), and the text features (such as surgery types) are encoded into binary representation (0/1) to add into the one-hot vector. The continuous features (such as height) can be fed into the vector directly since we have the normalization process during preprocessing. Effective features need to be extracted from the time series feature before being added to the vector. To capture the temporal effects in time series, we use a bucketing technique. For the time series data of each patient, we divided it into 2 buckets based on the care time (for ORP) or room start time (for GHWs), and compute the features in each bucket. Then, we extract first order features and second order features from patients' vital sign time series in each bucket. The details of first order and second order feature from time series are as follows:

3.1 First Order Features

We use some traditional statistical features as the first order features. Specifically, the first order features include maximum, minimum, mean (μ), standard deviation(σ), skewness and kurtosis in each bucket. Skewness is a measure of symmetry of the probability distribution of a real-valued random variable. The larger absolute value of skewness means the greater deviation of its distribution. Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. A larger absolute value of the kurtosis represents greater difference between the steepness of its distribution and the normal distribution. The formula of mean, standard deviation, skewness and kurtosis are:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N},\ \sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)}{N}} \tag{1}$$

$$Skewness = \frac{\sum_{i=1}^{N}(\mathbf{x}_i - \mu)^3}{(N-1)\sigma^3} \tag{2}$$

$$Kurtosis = \frac{\sum_{i=1}^{N}(\mathbf{x}_i - \mu)^4}{(N-1)\sigma^4} - 3 \tag{3}$$

3.2 Second Order Features

The most commonly used second order features are co-occurrence features. The co-occurrence features in one-dimensional time series have been shown to perform better than other second-order features [14]. The data is firstly quantized into Q levels, and then a two dimensional matrix $\phi(i,j)(1 \leq i,j \leq Q)$ is constructed. Point (i,j) in the matrix represents the number of times that a point in the sequence with level i is followed, at a distance d_1, by a point with level j. The co-occurrence features we used are Energy (E_1), Entropy (E_2), Correlation ($\rho_{x,y}$), Inertia, and Local Homogeneity (LH). The features are calculated by the following equations:

$$E_1 = \sum_{i=1}^{Q}\sum_{j=1}^{Q}\phi(i,j)^2,\ E_2 = \sum_{i=1}^{Q}\sum_{j=1}^{Q}\phi(i,j) * log(\phi(i,j))$$

$$\rho_{x,y} = \frac{\sum_{i=1}^{Q}\sum_{j=1}^{Q}(i-\mu_x)(j-\mu_y)\phi(i,j)}{\sigma_x\sigma_y}$$

where:

$$\mu_x = \frac{\sum_{i=1}^{Q} i \sum_{j=1}^{Q}\phi(i,j)}{Q},\ \sigma_x^2 = \frac{\sum_{i=1}^{Q}(i-\mu_x)^2\sum_{j=1}^{Q}\phi(i,j)}{Q}$$

$$\mu_y = \frac{\sum_{j=1}^{Q} j \sum_{i=1}^{Q}\phi(i,j)}{Q},\ \sigma_y^2 = \frac{\sum_{j=1}^{Q}(j-\mu_y)^2\sum_{i=1}^{Q}\phi(i,j)}{Q}$$

$$Inertia = \sum_{i=1}^{Q}\sum_{j=1}^{Q}(i-j)^2\phi(i,j),\ LH = \sum_{i=1}^{Q}\sum_{j=1}^{Q}\frac{\phi(i,j)}{1+(i-j)^2}$$

We set $Q = 5$ in our experiments. The extracted first order and second order features are concatenated into the one-hot vector as input to our model.

3.3 Convolutional Neural Network for Time Series Feature Learning

We use Convolutional Neural Network (CNN) to automatically learn features from time series (such as heart rate, temperature and blood pressure which are recorded every minute). In our setting, we regard CNN as feature extractor. The input time series is fed into CNN model, containing several convolutional layers, activation layers and max-pooling layers to learn features.

The convolutional layer contains a set of learnable filters which are updated using the backpropagation algorithm. Convolution operation can capture local temporal information from the time series. We use the same filter size through all convolutional layers.

The activation layer introduces the non-linearity into neural networks and allows it to learn more complex model. We adopt $sigmoid(\cdot)$ as our activation function in all activation layers, where $sigmoid(\alpha) = 1/(1+e^{-\alpha})$.

The max-pooling layer aims to provide an abstracted form of the representation by down-sampling. At the same time, it reduces the computational cost by reducing the number of parameters to learn and provides basic translation invariance to the internal representation.

The statistical features can be combined with features learnt from CNN, and furthure feed them into a multilayer perceptronn for readmission prediction task. In principle, our extracted and learnt features can be used as input to any classification algorithms.

4 PREDICTION METHODOLOGY

A main challenge in our application is that we have severely skewed datasets as there are much more normal patients than those with deterioration. For example, among 2565 records in the GHWs data, only 406 have a 30-day readmission. This extremely imbalanced class distribution makes the prediction task very difficult.

4.1 Classification Algorithms

In the medical domain, the cost of misdiagnosing abnormal patient as healthy is different with misdiagnosing healthy as abnormal patient. In most cases, the proportion of normal patients is larger than abnormal patients (e.g., readmission and ICU patients). Therefore, in our data sets, we have two crucial issues during classification. One is imbalanced outcomes and another one is low sensitivity of abnormal patients. Standard classifiers, however, pay less attention to rare cases in an imbalanced data set. Consequently, test patients belonging to the small class are misclassified more often than those belonging to the prevalent class.

To over this problem, we formalize it as a cost-sensitive classification problem. Cost-sensitive classification considers the varying costs of different misclassification types. A cost matrix encodes the penalty of classifying samples from one class as another. Bayesian optimal decision can help obtain the cost-sensitive prediction . Eq. 4 shows the predicted class label that reaches the lowest expected cost:

$$y_{pred} = \underset{1\leq k\leq K}{\arg\min}\sum_{i=1}^{K}P(y=i|\mathbf{x},\mathbf{W},\mathbf{b})\mathbf{C}(k,i) \tag{4}$$

where $\mathbf{C}(k,i)$ denotes the cost of predicting a sample from class k as class i. The diagonal elements in the cost matrix are set to be 1, others are negative. Smaller value in the cost matrix impose larger penalty. The $P(y=i|\mathbf{x},\mathbf{W},\mathbf{b})$ is to estimate the probability of class i given $\mathbf{x}$. The probability estimator can be any classifiers which the outputs are probability. Nowadays deep neural network models are successful for different applications and attracting increasing attention. In this work, we use a loss function that embeds the cost information. We denote the deep neural network (DNN) with cost sensitive as CSDNN for short.

The DNN consists of one input layer, one output layer and multiple hidden layers. There are m neurons in the input layer, where m is the dimension of input feature vector. The hidden layers are fully-connected with the input layer. Each hidden layer h uses $\mathbf{W}_h$ as a fully-connected weight matrix and $\mathbf{b}_h$ as a bias vector that enters the neurons. Then, for

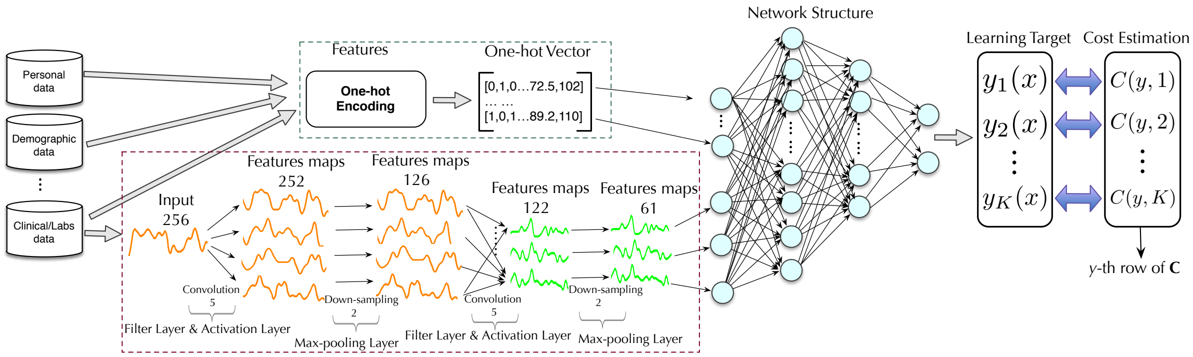

Figure 2: CSDNN framework. It first extracts features from discrete, text, continuous and time series data to an one-hot vector by using one-hot encoding. In the meanwhile, we use convolutional neural networks to automatically learn features from time series data. During prediction, CSDNN considers different costs of misclassification errors with a cost matrix C in the output layer. Once acquiring predicted outcome y, the predicted errors can be calculated with cost matrix C according to the loss function in Eq. (7).

an input feature vector $\mathbf{x}$, the output of the hidden layer is $\mathcal{H}(\mathbf{W}_h\mathbf{x}+\mathbf{b}_h)$, where the activation function $\mathcal{H}$ can be sigmod or tanh. We used tanh in the experiments because it typically yields faster training (and sometimes better local minima), that is, $\mathcal{H}(\alpha) = (e^{\alpha} - e^{-\alpha})/(e^{\alpha}+e^{-\alpha})$. After H hidden layers, the DNN describes a complex feature transform function by computing:

$$\mathcal{F}(\mathbf{x}) = \mathcal{H}(\mathbf{W}_H \cdot \mathcal{H}(\cdots \mathcal{H}(\mathbf{W}_2 \cdot \mathcal{H}(\mathbf{W}_1\mathbf{x}+\mathbf{b}_1)+\mathbf{b}_2)\cdots)+\mathbf{b}_H) \tag{5}$$

Then, an output layer is placed after the H-th hidden layer. From hidden layer to output layer is a logistic regression to output the probability of feature vector $\mathbf{x}$ belonging to each category. Hence, there are K neurons (outputs) in the output layer, where the i-th neuron with weights $\mathbf{W}_o^i$ and bias $\mathbf{b}_o^i$ (the subscript o represents the parameters in the output layer). We used softmax regression in the output layer to make the model capable of handle multi-class classification. Therefore, the estimate of probability of class i given $\mathbf{x}$ can be formulated as follows:

$$\begin{aligned} P(y=i|\mathbf{x},\mathbf{W},\mathbf{b}) &= softmax_i(\mathbf{W}_o\mathcal{F}(\mathbf{x})+\mathbf{b}_o) \\ &= \frac{\exp(\mathbf{W}_o^i\mathcal{F}(\mathbf{x})+\mathbf{b}_o^i)}{\sum_{k=1}^{K}\exp(\mathbf{W}_o^i\mathcal{F}(\mathbf{x})+\mathbf{b}_o^i)} \end{aligned} \tag{6}$$

To learn and optimize the parameters of the model, we set the negative log-likelihood as the loss function and minimize the loss function with respect to $\{\mathbf{W}_h\}_{h=1}^{H}$, $\{\mathbf{b}_h\}_{h=1}^{H}$, $\mathbf{W}_o$ and $\mathbf{b}_o$. The loss function over the training set is as follows:

$$Loss = -\frac{1}{N}\sum_{n=1}^{N} log\big[\sum_{i=1}^{K} P(y=i|\mathbf{x}_n,\mathbf{W},\mathbf{b})\mathbf{C}(y_n,i)\big] + \frac{\lambda}{2}\|\mathbf{W}\|_2^2 \tag{7}$$

where N is the total number of patients, y_n is the readmission indicator for the n-th patient where 1 indicates readmission and 0 control, and $P(y=y_n|\mathbf{x}_n,\mathbf{W},\mathbf{b})$ is the n-th patient calculated by the model. The loss minimization and parameter optimization can be performed through the back-propagation and mini-batch stochastic gradient descent.

The CSDNN framework is shown in Figure 2. Both extracted statistical features by one-hot encoding and learnt features by CNN are input to a multilayer perceptron. The cost matrix is applied to the loss function in Eq. (7) during prediction phase. To estimate parameters of models, we utilize gradient-based optimization method to minimize the loss function. Since backpropagation is an efficient and most widely used gradient-based method in neural networks [20], we use simple backpropagation algorithm to train our CSDNN model. As stochastic gradient descent (SGD) could converge faster than full-batch for large scale data sets, we adopt SGD instead of the full-batch version to update the parameters.

5 EXPERIMENTS AND EVALUATION

5.1 Data sets and Setup

We evaluate performance of proposed CSDNN framework on two real data sets from Barnes-Jewish Hospital. One data is from general hospital wards (GHWs) while another one is pilot data from operating room (ORP). The two data sets are described in Section 2, and more details are as follows:

GHWs data: We aim to predict 30-day and 60-day readmission in the GHWs data. There are 41,503 patient visits in the GHWs data, and 2,565 have the outcomes of readmission or not. In this data set, each patient is measured for 34 indicators, including demographics, vital signs (pulse, shock index, mean arterial blood pressure, temperature, and respiratory rate), and lab tests (albumin, bilirubin, BUN, creatinine, sodium, potassium, glucose, hemoglobin, white cell count, INR, and other routine chemistry and hematology results). A total of 406 patients are readmitted within 30 days and 538 instances are readmitted within 60 days.

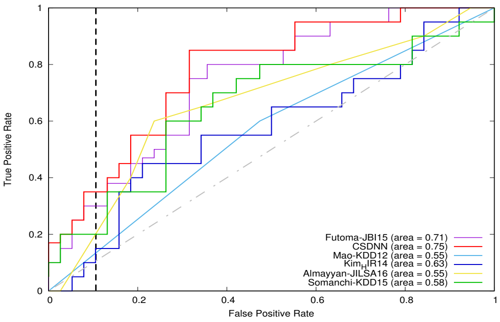

Figure 3: ROC curves of 1-year readmission prediction on the ORP data set.

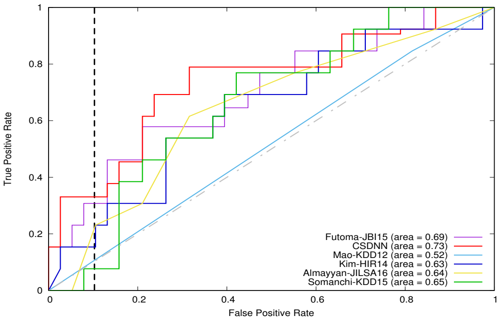

Figure 4: ROC curves of 30-day readmission prediction on the ORP data set.

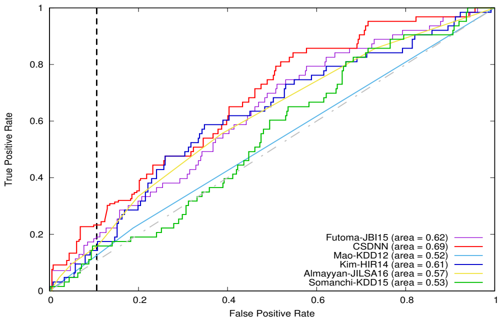

Figure 5: ROC curves of 30-day readmission prediction on the GHWs data set.

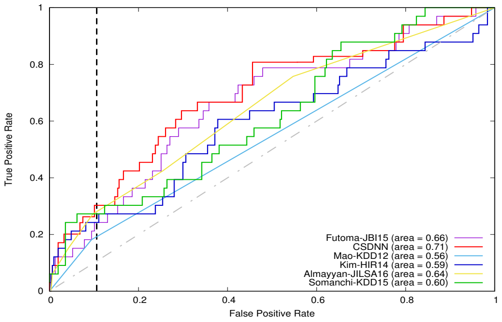

Figure 6: ROC curves of 60-day readmission prediction on the GHWs data set.

ORP data: We aim to predict 30-days and 1-year readmission in the ORP data. There are 700 patients in the pilot data with more than 50 pre-operation features and 26 intra-operation vital signs of each patient. Since there are plenty of null outcomes in the pilot data set, we remove the patients with null outcomes. After removing the null outcomes, we have 403 patients for predicting 1-year readmission, including 157 patients are labelled as readmission, and 359 patients for 30-days readmission, including 124 patients with readmission label.

For both data sets, we randomly select 60%, 15%, and 25% from readmission and non-readmission patients as training data, validation data and test data, respectively. We choose the best parameters through validation data.

5.2 Evaluation Criteria

Following the most common procedure for evaluating models for early predicting readmission, we use: ROC (receives operating Characteristic) Curve, AUC (Area Under (ROC) Curve), Accuracy (Precision), Sensitivity (Recall), Specificity, PPV (Positive Predictive Value), and NPV (Negative Predictive Value) to evaluate the proposed method.

Baselines: We evaluate CSDNN for comparison with existing approaches used in hospitals. From the literature study, the existing predictive methods for readmission are mainly based on feature extraction for specific disease or data set, and then input the extracted features to classifiers. The most widely used classifiers are Support Vector Machine (SVM), Logistic Regression(LR), Decision Tree (DT), Random Forest (RF) and Artificial Neural Networks (ANN). In spite of the settings of our problem are not exactly the same with all the baselines, we implement baselines based on their methodologies used in the state-of-the-art approaches for readmission prediction. Specifically, [Mao et al., KDD'12] [14] proposed an integrated data mining approach with various statistical features, and applied an exploratory undersampling [13] method to deal with the class-imbalance problem. They used RF as classifier and obtain good performance. [Somanchi et al., KDD'15] [15] extracted features from heterogeneous data source (such as demographics and vitals), and employed SVM as classifier for cardiac arrest early prediction. [Kim et al., HIR'11] [11] used extra physiological variables extracted from an APACHE critical care system, and shows DT classifier archives the best performance. [Almayyan, JILSA'16] [2] selected discriminative features using PSO and several feature selection techniques to reduce the features dimension, and applied random forest classifier to diagnose lymphatic diseases. [Futoma, JBI'15] [7] applied ANN for

Table 1: 30-day readmission prediction on the GHWs data set.

Method	Accuracy	Specificity	Sensitivity	F1-Score	AUC	NPV	PPV
Somanchi_KDD15	0.83	0.85	0.08	0.15	0.53	0.97	0.02
Mao_KDD12	0.72	0.86	0.18	0.30	0.52	0.81	0.24
Kim_HIR14	0.85	0.85	0.00	0.00	0.61	1.00	0.00
Almayyan_JILSA16	0.84	0.85	0.11	0.19	0.57	0.98	0.02
Futoma_JBI15	0.84	0.86	0.23	0.36	0.62	0.97	0.06
CSDNN	0.87	0.89	0.27	0.41	0.69	0.83	0.35

Table 2: 60-day readmission prediction on the GHWs data set.

Method	Accuracy	Specificity	Sensitivity	F1-Score	AUC	NPV	PPV
Somanchi_KDD15	0.87	0.91	0.19	0.31	0.60	0.94	0.12
Mao_KDD12	0.84	0.91	0.19	0.31	0.56	0.90	0.21
Kim_HIR14	0.90	0.90	0.00	0.00	0.59	1.00	0.00
Almayyan_JILSA16	0.90	0.91	0.23	0.37	0.61	0.99	0.03
Futoma_JBI15	0.89	0.91	0.23	0.37	0.66	0.98	0.06
CSDNN	0.91	0.93	0.27	0.42	0.71	0.92	0.28

Table 3: 1-year readmission prediction on the ORP data set.

Method	Accuracy	Specificity	Sensitivity	F1-Score	AUC	NPV	PPV
Somanchi_KDD15	0.64	0.70	0.16	0.26	0.58	0.58	0.21
Mao_KDD12	0.67	0.74	0.21	0.32	0.55	0.81	0.15
Kim_HIR14	0.61	0.77	0.31	0.44	0.63	0.63	0.40
Almayyan_JILSA16	0.71	0.76	0.38	0.51	0.55	0.87	0.08
Futoma_JBI15	0.68	0.78	0.35	0.48	0.71	0.80	0.32
CSDNN	0.75	0.79	0.42	0.55	0.75	0.79	0.41

Table 4: 30-day readmission prediction on the ORP data set.

Method	Accuracy	Specificity	Sensitivity	F1-Score	AUC	NPV	PPV
Somanchi_KDD15	0.65	0.72	0.00	0.00	0.65	0.87	0.00
Mao_KDD12	0.65	0.74	0.22	0.34	0.52	0.82	0.15
Kim_HIR14	0.61	0.80	0.24	0.37	0.63	0.65	0.28
Almayyan_JILSA16	0.73	0.78	0.25	0.38	0.64	0.67	0.22
Futoma_JBI15	0.78	0.83	0.41	0.55	0.69	0.72	0.15
CSDNN	0.79	0.85	0.49	0.62	0.73	0.87	0.33

predicting early hospital readmission and get good predictive performance than regression methods.

For simplicity, we use Mao_KDD12, Somanchi_KDD15, Kim_HIR14, Almayyan_JILSA16, and Futoma_JBI15 for short to represent the benchmark approaches.

6 RESULTS AND DISCUSSION

Results: Tables 2-4 and Figures 3-5 present the performance of the different predictive approaches on the GHWs and ORP data sets. In comparison to the state-of-the-art baselines on the test set, we find that our model (CSDNN) performs better than baselines in terms of AUC and PPV. The PPVs of our model are over twice the value of that found in the baselines. Obviously, the PPV is statistically significantly improved by using cost-sensitive deep learning. This is critical since the misclassification costs of readmission patients is more serious And our goal is to make the predictions for abnormal patients as precise as possible under high NPV, which enables the hospital to intervene early, as well as adjust the schedules for physicians and nurses to optimize overall quality of care for all patients. As we can observe from the ROC curves in Figures 3-5, we are able to predict readmission with around 90% sensitivity and 30% false positive rate, which is better than baselines under that same false positive rate.

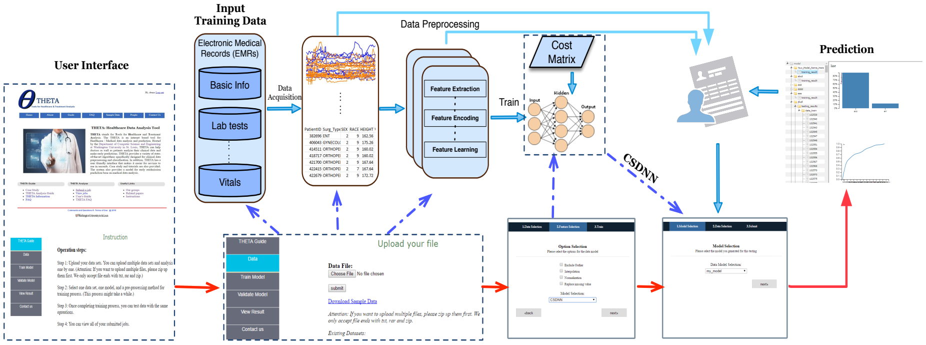

Figure 7: An illustration of the deployed system workflow. The system has user-friendly interfaces and detailed user guide. Physicians can follow the steps and the case study on the guide pages to predict readmission.

Discussion: We achieved high accuracy mainly because we used both sufficient statistical features and automatically learned time series features by convolutional neural networks (CNN), as well as we consider the misclassification costs to improve PPV. Compared with traditional statistical features, CNN can learn a hierarchical feature representation from raw data automatically, which make it possible to improve the accuracy of feature-based methods. Cost-sensitive deep learning approach ensures the prediction of rare but high misclassification cost class, which are developed by introducing cost items into the learning framework. However, this may affect the prediction for normal patient (NPV). Thus, the PPV and NPV need to be tradeoff. As we believe the cost of a false positive is considerably higher than a false negative, relatively low NPV may be a tolerable tradeoff.

Sensitivity analysis: For any test, there is usually a tradeoff between the different measures. This tradeoff can be represented using a ROC curve, which is a plot of sensitivity or true positive rate, versus false positive rate (1-specificity). For practical deployment in hospitals, a high specificity (e.g. >90%) is needed. The ROC figures also show the results of all algorithms with specificity being fixed close to 0.90. Even at this relatively high specificity, the CSDNN approach can achieve a sensitivity of around 35% on the ORP data. The sensitivity of ORP data is relatively higher than GHWs data, because the ORP data is a small pilot data and not very imbalance compared with GHWs data.

7 SYSTEM DEPLOYMENT

The work described here was done in partnership with Barnes-Jewish Hospital, one of the largest hospitals in the United States. Based on our performance, the results is good enough to deploy a decision support system with the proposed predictive algorithms to support treatment. The purpose of the clinical decision support system is to identify prognostic factors and suggests interventions based on novel feature extracting and learning algorithms using heterogeneous data.

We build up a system[1] to deploy our CSDNN algorithm for early readmission prediction. The system architecture is demonstrated in Figure 7. The system is an internet based tool for medical data analysis and outcome prediction, for example, readmission prediction via our CSDD algorithm. There are four key components in the system: 1) Data acquisition. There are user-friendly interfaces to guide user how to submit a job and how to train a model. Physicians can upload historical EMR data to the system following the sample data format. 2) Data preprocessing. After uploading the training data by physicians, the system preprocesses the raw data with several modules, including feature extracting, feature encoding and feature learning. 3) Model selection. Users can select which model will be trained. Since we integrated several models into the system for different tasks, users should select one model for specific purpose. In our case, CSDNN should be selected to predict readmission. Once a model is selected, the system will train the model using the uploaded data. 4) Prediction. Once the training phase is over, test data can be fed into the system. Test data will be analyzed using the trained model (CSDNN). Finally, the system shows the results to indicate whether the patient is readmission or not.

8 RELATED WORK

A number of forecasting algorithms exist that use medical data for outcomes prediction. To predict whether a patient is readmitted to hospital, existing dedicated efforts are mostly focused on extracting effective features and using accurate classifiers. In this section, we give a brief overview of research efforts done along early readmission prediction at hospital.

As readmission act as a substantial contributor to rising healthcare costs, predicting readmission has been identified as one of the key problems for the healthcare domain. However, there are not many solutions are known to be effective. He et al. [9] present a data-driven method to predict hospital readmission solely on administrative claims data. Nevertheless, their method is unable to incorporate clinical laboratory

[1] http://www.cse.wustl.edu/∼z.cui/Theta/

data in the model and as a result is not able to directly compare its performance with other approaches. Applying a comprehensive dataset that make generalization more reasonable. Therefore, [2, 6, 15] leverage a variety of data sources, including patient demographic and social characteristics, medications, procedures, conditions, and lab tests. However, they used the features designed for specific disease. [11, 14, 16] come up with more general statistical features used for predicting readmission with conventional modeling techniques, such as support vector machine (SVM) or logistic regression. All of these methods relies on feature extraction and the ability of classifiers, which limit the performance of their methods.

To date, previous works on early readmission prediction by extracting statistical features from vital signs are inefficient feature representing methods, since they are hard to capture temporal patterns present in longitudinal time series data. Choi et al. [5] show deep learning models outperform the traditional modeling techniques in medical domain, and deep learning can be interpretable for healthcare analysis [4]. However, these works based on deep learning fail to consider the imbalanced data problem.

9 CONCLUSIONS

Readmission is a major source of cost for healthcare systems. Readmission not only degrades the quality of health care but also increases medical expenses. In this paper, we aim to identify those patients who are likely to be readmitted to the hospital. The identified patients can then be considered by health care personnel for application of preventive alternative measures. The goal is to deliver superior prediction quality, with good interpretability and high computational efficiency, that supports early readmission prediction.

Deep learning has been one of the most prominent machine learning techniques nowadays. Deep learning makes possible automatic feature learning from medical data. We propose to use both traditional statistical features via one-hot encoding and learnt features via convolutional neural networks as input to a multilayer perceptron. This way can utilize the advantage of local information, temporal information and overall trends in vital signs time series. However, imbalance or skewed class distribution are challenges in medical data. For most cases, the recognition importance of positive instances is higher than that of negative instances. Therefore, we further propose a cost-sensitive deep learning model to address the imbalanced problem on medical data. The effectiveness of the proposed approach is validated on two real medical data sets from Barnes-Jewish Hospital. Our performance is good enough to warrant an actual clinical trial in hospital setting. Consequently, our model has been deployed in a real system for readmission prediction.

REFERENCES

[1] MM Al Rahhal, Yakoub Bazi, Haikel AlHichri, Naif Alajlan, Farid Melgani, and RR Yager. 2016. Deep learning approach for active classification of electrocardiogram signals. *Information Sciences* 345 (2016), 340–354.

[2] Waheeda Almayyan. 2016. Lymph Diseases Prediction Using Random Forest and Particle Swarm Optimization. *Journal of Intelligent Learning Systems and Applications* 8, 03 (2016), 51.

[3] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1721–1730.

[4] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. In *Advances in Neural Information Processing Systems*. 3504–3512.

[5] Edward Choi, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association* (2016), ocw112.

[6] Shahid A Choudhry, Jing Li, Darcy Davis, Cole Erdmann, Rishi Sikka, and Bharat Sutariya. 2013. A public-private partnership develops and externally validates a 30-day hospital readmission risk prediction model. *Online journal of public health informatics* 5, 2 (2013), 219.

[7] Joseph Futoma, Jonathan Morris, and Joseph Lucas. 2015. A comparison of models for predicting early hospital readmissions. *Journal of biomedical informatics* 56 (2015), 229–238.

[8] John Cristian Borges Gamboa. 2017. Deep Learning for Time-Series Analysis. *arXiv preprint arXiv:1701.01887* (2017).

[9] Danning He, Simon C Mathews, Anthony N Kalloo, and Susan Hutfless. 2014. Mining high-dimensional administrative claims data to predict early hospital readmissions. *Journal of the American Medical Informatics Association* 21, 2 (2014), 272–279.

[10] Stephen F Jencks, Mark V Williams, and Eric A Coleman. 2009. Rehospitalizations among Patients in the Medicare Fee-for-Service Program. *N Engl J Med* 360 (2009), 1418–28.

[11] Sujin Kim, Woojae Kim, and Rae Woong Park. 2011. A comparison of intensive care unit mortality prediction models through the use of data mining techniques. *Healthcare informatics research* 17, 4 (2011), 232–243.

[12] Eun Whan Lee. 2012. Selecting the Best Prediction Model for Readmission. *Journal of Preventive Medicine and Public Health* 45, 4 (2012), 259.

[13] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. 2009. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39, 2 (2009), 539–550.

[14] Yi Mao, Wenlin Chen, Yixin Chen, Chenyang Lu, Marin Kollef, and Thomas Bailey. 2012. An integrated data mining approach to real-time clinical monitoring and deterioration warning. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1140–1148.

[15] Sriram Somanchi, Samrachana Adhikari, Allen Lin, Elena Eneva, and Rayid Ghani. 2015. Early prediction of cardiac arrest (code blue) using electronic medical records. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2119–2126.

[16] Shanu Sushmita, Garima Khulbe, Aftab Hasan, Stacey Newman, Padmashree Ravindra, Senjuti Basu Roy, Martine De Cock, and Ankur Teredesai. 2016. Predicting 30-day risk and cost of" all-cause" hospital readmissions. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.

[17] Christian Szegedy. 2016. An Overview of Deep Learning. *AITP 2016* (2016).

[18] Yujin Tang, Jianfeng Xu, Kazunori Matsumoto, and Chihiro Ono. 2016. Sequence-to-Sequence Model with Attention for Time Series Classification. In *Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on*. IEEE, 503–510.

[19] Zhiguang Wang, Weizhong Yan, and Tim Oates. 2016. Time Series Classification from Scratch with Deep Neural Networks: A Strong Baseline. *arXiv preprint arXiv:1611.06455* (2016).

[20] Yi Zheng, Qi Liu, Enhong Chen, Yong Ge, and J Leon Zhao. 2014. Time series classification using multi-channels deep convolutional neural networks. In *International Conference on Web-Age Information Management*. Springer, 298–310.