This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2694446, IEEE Access

Disease Prediction by Machine Learning over Big Data from Healthcare Communities

Min Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang*

***Abstract*—With big data growth in biomedical and healthcare communities, accurate analysis of medical data benefits early disease detection, patient care and community services. However, the analysis accuracy is reduced when the quality of medical data is incomplete. Moreover, different regions exhibit unique characteristics of certain regional diseases, which may weaken the prediction of disease outbreaks. In this paper, we streamline machine learning algorithms for effective prediction of chronic disease outbreak in disease-frequent communities. We experiment the modified prediction models over real-life hospital data collected from central China in 2013-2015. To overcome the difficulty of incomplete data, we use a latent factor model to reconstruct the missing data. We experiment on a regional chronic disease of cerebral infarction. We propose a new convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm using structured and unstructured data from hospital. To the best of our knowledge, none of the existing work focused on both data types in the area of medical big data analytics. Compared to several typical prediction algorithms, the prediction accuracy of our proposed algorithm reaches 94.8% with a convergence speed which is faster than that of the CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm.**

***Index Terms*—Big data analytics; Machine Learning; Healthcare**

I. Introduction

According to a report by McKinsey [1], 50% of Americans have one or more chronic diseases, and 80% of American medical care fee is spent on chronic disease treatment. With the improvement of living standards, the incidence of chronic disease is increasing. The United States has spent an average of 2.7 trillion USD annually on chronic disease treatment. This amount comprises 18% of the entire annual GDP of the United States. The healthcare problem of chronic diseases is also very important in many other countries. In China, chronic diseases are the main cause of death, according to a Chinese report on nutrition and chronic diseases in 2015, 86.6% of deaths are caused by chronic diseases. Therefore, it is essential to perform risk assessments for chronic diseases. With the growth in medical data [2], collecting electronic health records (EHR) is increasingly convenient [3]. Besides, [4] first presented a bio-inspired high-performance heterogeneous vehicular telematics paradigm, such that the collection of mobile users' health-related real-time big data can be achieved with the deployment of advanced heterogeneous vehicular networks. Chen et.al [5]–[7] proposed a healthcare system using smart clothing for sustainable health monitoring. Qiu et al. [8] had thoroughly studied the heterogeneous systems and achieved the best results for cost minimization on tree and simple path cases for heterogeneous systems. Patients' statistical information, test results and disease history are recorded in the EHR, enabling us to identify potential data-centric solutions to reduce the costs of medical case studies. Qiu et al. [9] proposed an efficient flow estimating algorithm for the telehealth cloud system and designed a data coherence protocol for the PHR(Personal Health Record)-based distributed system. Bates et al. [10] proposed six applications of big data in the field of healthcare. Qiu et al. [11] proposed an optimal big data sharing algorithm to handle the complicate data set in telehealth with cloud techniques. One of the applications is to identify high-risk patients which can be utilized to reduce medical cost since high-risk patients often require expensive healthcare. Moreover, in the first paper proposing healthcare cyber-physical system [12], it innovatively brought forward the concept of prediction-based healthcare applications, including health risk assessment. Prediction using traditional disease risk models usually involves a machine learning algorithm (e.g., logistic regression and regression analysis, etc.), and especially a supervised learning algorithm by the use of training data with labels to train the model [13], [14]. In the test set, patients can be classified into groups of either high-risk or low-risk. These models are valuable in clinical situations and are widely studied [15], [16]. However, these schemes have the following characteristics and defects. The data set is typically small, for patients and diseases with specific conditions [17], the characteristics are selected through experience. However, these pre-selected characteristics maybe not satisfy the changes in the disease and its influencing factors.

With the development of big data analytics technology, more attention has been paid to disease prediction from the perspective of big data analysis, various researches have been conducted by selecting the characteristics automatically from a large number of data to improve the accuracy of risk classification [18], [19], rather than the previously selected characteristics. However, those existing work mostly considered structured data. For unstructured data, for example, using convolutional neural network (CNN) to extract text characteristics automatically has already attracted wide attention and also achieved

M. Chen, Y. Hao, and L. Wang are with the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China (minchen@ieee.org).

K. Huang is with the university of Southern California, Los Angeles, California, 90089, USA (kaihwang@usc.edu).

L. Wang is with the Research Center for Tissue Engineering and Regenerative Medicine, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology , Jiefang Avenue 1277, Wuhan, 430022, China, and also with department of Clinical Laboratory, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, 430022, China.

*Corresponding author: Lin Wang (lin_wang@hust.edu.cn).

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2694446, IEEE Access

very good results [20], [21] . However, to the best of our knowledge, none of previous work handle Chinese medical text data by CNN. Furthermore, there is a large difference between diseases in different regions, primarily because of the diverse climate and living habits in the region. Thus, risk classification based on big data analysis, the following challenges remain: How should the missing data be addressed? How should the main chronic diseases in a certain region and the main characteristics of the disease in the region be determined? How can big data analysis technology be used to analyze the disease and create a better model?

To solve these problems, we combine the structured and unstructured data in healthcare field to assess the risk of disease. First, we used latent factor model to reconstruct the missing data from the medical records collected from a hospital in central China. Second, by using statistical knowledge, we could determine the major chronic diseases in the region. Third, to handle structured data, we consult with hospital experts to extract useful features. For unstructured text data, we select the features automatically using CNN algorithm. Finally, we propose a novel CNN-based multimodal disease risk prediction (CNN-MDRP) algorithm for structured and unstructured data. The disease risk model is obtained by the combination of structured and unstructured features. Through the experiment, we draw a conclusion that the performance of CNN-MDPR is better than other existing methods.

The remainder of this article is organized as follows. We describe the dataset and model in Section II. The methods used in this paper are described in Section III. The performance of CNN-UDRP and CNN-MDRP algorithms is discussed in Section IV. We provide the overall results in Section V. Finally, Section VI concludes this paper.

II. Dataset and Model Description

In this section, we describe the hospital datasets we use in this study. Furthermore, we provide disease risk prediction model and evaluation methods.

A. Hospital Data

The hospital dataset used in this study contains real-life hospital data, and the data are stored in the data center. To protect the patient's privacy and security, we created a security access mechanism. The data provided by the hospital include EHR, medical image data and gene data. We use a three year data set from 2013 to 2015. Our data focus on inpatient department data which included 31919 hospitalized patients with 20320848 records in total. The inpatient department data is mainly composed of structured and unstructured text data. The structured data includes laboratory data and the patient's basic information such as the patient's age, gender and life habits, etc. While the unstructured text data includes the patient's narration of his/her illness, the doctor's interrogation records and diagnosis, etc. As shown in Table I, the real-life hospital data collected from central China are classified into two categories, i.e., structured data and unstructured text data.

In order to give out the main disease which affect this region, we have made a statistics on the number of patients, the sex ratio of patients and the major disease in this region every year from the structured and unstructured text data, the statistical results are as shown in Table II. From Table II, we can obtain that the proportion of male and female patients hospitalized each year have little difference and more patients admitted to the hospital in 2014. Moreover, the hospitalization resulted by chronic diseases has always been occupying a large proportion in this area through the statistics of the data. For example, the number of patients hospitalized with the chronic diseases of cerebral infarction, hypertension, and diabetes accounted for 5.63% of the total number of patients admitted to the hospital in 2015, while the other diseases occupied a small proportion. In this paper, we mainly focus on the risk prediction of cerebral infarction since cerebral infarction is a fatal disease.

B. Disease Risk Prediction

From Table II, we obtain the main chronic disease in this region. The goal of this study is to predict whether a patient is amongst the cerebral infarction high-risk population according to their medical history. More formally, we regard the risk prediction model for cerebral infarction as the supervised learning methods of machine learning, i.e., the input value is the attribute value of the patient, $X = (x_1, x_2, \cdots, x_n)$ which includes the patient's personal information such as age, gender, the prevalence of symptoms, and living habits (smoking or not) and other structured data and unstructured data.

The output value is C, which indicates whether the patient is amongst the cerebral infarction high-risk population. $C = \{C_0, C_1\}$, where, C_0 indicates the patient is at high-risk of cerebral infarction, C_1 indicates the patient is at low-risk of cerebral infarction. The following will introduce the dataset, experiment setting, dataset characteristics and learning algorithms briefly.

For dataset, according to the different characteristics of the patient and the discussion with doctors, we will focus on the following three datasets to reach a conclusion.

- Structured data (S-data): use the patient's structured data to predict whether the patient is at high-risk of cerebral infarction.
- Text data (T-data): use the patient's unstructured text data to predict whether the patient is at high-risk of cerebral infarction.
- Structured and text data (S&T-data): use the S-data and T-data above to multi-dimensionally fuse the structured data and unstructured text data to predict whether the patient is at high-risk of cerebral infarction.

In the experiment setting and dataset characteristics, we select 706 patients in total as the experiment data and randomly divided the data into training data and test data. The ratio of the training set and the test set is 6:1 [22], [23] , i.e., 606 patients as the training data set while 100 patients as the test data set. We use the C++ language to realize the machine learning and deep learning algorithms and run it in a parallel fashion by the use of data center. In this paper, for S-data, according to the discussion with doctors and Pearson's correlation analysis, we extract the patient's demographics characteristics and some

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2694446, IEEE Access

TABLE I
ITEM TAXONOMY IN CHINA HOSPITAL DATA

Data category	Item	Description
Structured data	Demographics of the patient	Patient's gender, age, height, weight, etc.
	Living habits	Whether the patient smokes, has a genetic history, etc.
	Examination items and results	Includes 682 items, such as blood, etc.
	Diseases	Patient's disease, such as cerebral infarction, etc.
Unstructured text data	Patient's readme illness	Patient's readme illness and medical history
	Doctor's records	Doctor's interrogation records

TABLE II
INITIAL STATISTICS FROM HOSPITAL DATA IN WUHAN,CHINA

Statistics	2013	2014	2015
Number of inpatients	7265	24756	10552
Males	42.88%	50.36%	57.60%
Females	57.12%	49.64%	42.40%
Proportion of patients with cerebral infarction	1.47%	1.01%	1.66%
Proportion of hypertensive patients	1.06%	1.04%	1.98%
Proportion of diabetics	1.17%	0.99%	1.99%

of the characteristics associated with cerebral infarction and living habits (such as smoking). Then, we obtain a total of patient's 79 features. For T-data, we first extract 815073 words in the text to learn Word Embedding. Then we utilize the independent feature extraction by CNN.

We will introduce machine learning and deep learning algorithms used in this work briefly. For S-data, we use three conventional machine learning algorithms, i.e., Naive Bayesian (NB), K-nearest Neighbour (KNN), and Decision Tree (DT) algorithm [24], [25] to predict the risk of cerebral infarction disease. This is because these three machine learning methods are widely used [26]. For T-data, we propose CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm to predict the risk of cerebral infarction disease. In the remaining of the paper, let CNN-UDRP(T-data) denote the CNN-UDRP algorithm used for T-data. For S&T data, we predict the risk of cerebral infarction disease by the use of CNN-MDRP algorithm, which is denoted by CNN-MDRP(S&T-data) for the sake of simplicity. In the following section, the details about CNN-UDRP(T-data) and CNN-MDRP(S&T data) will be given.

C. Evaluation Methods

For the performance evaluation in the experiment. First, we denote TP, FP, TN and FN as true positive (the number of instances correctly predicted as required), false positive (the number of instances incorrectly predicted as required), true negative (the number of instances correctly predicted as not required) and false negative (the number of instances incorrectly predicted as not required), respectively. Then, we can obtain four measurements: accuracy, precision, recall and F1-measure as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$
$$\text{Precision} = \frac{TP}{TP+FP}, \quad \text{Recall} = \frac{TP}{TP+FN}$$
$$\text{F1-Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}},$$

where the F1-Measure is the weighted harmonic mean of the precision and recall and represents the overall performance.

In addition to the aforementioned evaluation criteria, we use receiver operating characteristic (ROC) curve and the area under curve (AUC) to evaluate the pros and cons of the classifier. The ROC curve shows the trade-off between the true positive rate (TPR) and the false positive rate (FPR), where the TPR and FPR are defined as follows:

$$TPR = \frac{TP}{TP+FN}, \quad TFR = \frac{FP}{FP+TN}$$

If the ROC curve is closer to the upper left corner of the graph, the model is better. The AUC is the area under the curve. When the area is closer to 1, the model is better. In medical data, we pay more attention to the recall rather than accuracy. The higher the recall rate, the lower the probability that a patient who will have the risk of disease is predicted to have no disease risk.

III. METHODS

In this section, we introduce the data imputation, CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm and CNN-based unimodal disease risk prediction (CNN-MDRP) algorithm.

A. Data Imputation

For patient's examination data, there is a large number of missing data due to human error. Thus, we need to fill

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2694446, IEEE Access

Algorithm 1 Stochastic Gradient Descent Algorithm

Input:

γ learning rate;
$\lambda_i, i = 1, 2$ regularization constant;
N the maximum number of iterations;
p_u^0 the initialization of P_u;
q_v^0 the initialization of q_v;

Output:

$\widehat{r}_{uv}$ real data;

1: $t := 0$, $n := 0$, $\widehat{r}_{uv}^0 = p_u^{0'} q_v^0$, $e_u v^0 = r_u v - \widehat{r}_{uv}^0$.
2: $t := t + 1$, $n := n + 1$.
3: Given the error $e_u v^{t-1} = r_{uv} - \widehat{r}_{uv}^{t-1}$ in the previous iteration.
4: Replace $p_u^t = p_u^{t-1} + \gamma(e_{uv}^{t-1} q_v^{t-1} - \lambda_1 p_u^{t-1})$, $q_v^t = q_v^{t-1} + \gamma(e_{uv}^{t-1} p_u^{t-1} - \gamma_2 q_v^{t-1})$, $\widehat{r}_{uv}^t = p_u^{t'} q_v^t$ and $e_{uv}^t = r_{uv} - \widehat{r}_{uv}^t$.
5: If e_{uv}^t approximately equals 0 or $n > N$, return $\widehat{r}_{uv} = \widehat{r}_{uv}^t$ for all possible (u, v); else, go to Step 2.

the structured data. Before data imputation, we first identify uncertain or incomplete medical data and then modify or delete them to improve the data quality. Then, we use data integration for data pre-processing. We can integrate the medical data to guarantee data atomicity: i.e., we integrated the height and weight to obtain body mass index (BMI). For data imputation, we use the latent factor model [27] which is presented to explain the observable variables in terms of the latent variables. Accordingly, assume that $R_{m\times n}$ is the data matrix in our healthcare model. The row designation, m represents the total number of the patients, and the column designation, n represents each patient's number of feature attributes. Assuming that there are k latent factors, the original matrix R can be approximated as

$$R_{(m\times n)} \approx P_{m\times k} Q_{n\times k}^T \tag{1}$$

Thus, each element value can be written as $\widehat{r}_{uv} = p_u^T q_v$, where p_u is the vector of the user factor, which indicates the patient's preference to these potential factors, and q_v is the vector of the feature attribute factor. The p_u and q_v values in the above formula are unknown.

To solve the problem, we can transform this problem into an optimization problem:

$$\min_{\{p,q\}} \left(\sum_{(u,v)} \left(r_{uv} - p_u^T q_v\right)^2 + \lambda_1 ||p_{\mathrm{u}}||^2 + \lambda_2 ||q_{\mathrm{v}}||^2 \right) \tag{2}$$

where r_{uv} is real data, p_u, q_v are the parameters to be solved, and $\lambda_i, i = 1, 2$ is a regularization constant, which can prevent overfitting in the operation process. We can solve it by the use of the stochastic gradient descent method. Define $e_{uv} = \widehat{r}_{uv} - r_{uv}$. Through the derivation above the optimization problem, we can get the specific solution as shown in Algorithm 1, which can fill missing data.

B. CNN-based Unimodal Disease Risk Prediction (CNN-UDRP) Algorithm

For the processing of medical text data, we utilize CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm which can be divided into the following five steps.

1) Representation of text data: As for each word in the medical text, we use the distributed representation of Word Embedding in natural language processing, i.e. the text is represented in the form of vector. In this experiment, each word will be represented as a R^d-dimensional vector, where $d = 50$. Thus, a text including n words can be represented as $T = (t_1, t_2, \cdots, t_n), T \in R^{d\times n}$.

2) Convolution layer of text CNN: Every time we choose s words, where $s = 5$ in Fig. 1(b). In other words, we choose two words from the front and back of each word vector $t_i^{'}$ in the text, i.e. use the row vector as the representation, to consist a $50 \times 5 = 250$ row vector, i.e. $s_i = (t_{i-2}^{'}, t_{i-1}^{'}, t_i^{'}, t_{i+1}^{'}, t_{i+2}^{'})$. As shown in Fig. 1(b), for s_1, s_2, s_{n-1} and s_n, we adopt an zero vector to fill. The selected weight matrix $W^1 \in R^{100\times 250}$ is as shown in Fig. 1(a), i.e., weight matrix W^1 includes 100 convolution filters and the size of each filter regions is 250. Perform convolution operation on W^1 and $s_i (i = 1, 2, \cdots, n)$, as shown in Fig.1(c). Specific calculation progress is that:

$$h_{i,j}^1 = f(W^1[\mathrm{i}] \cdot s_j + b^1) \tag{3}$$

where $i = 1, 2, \cdots, 100$, $j = 1, 2, \cdots, n$. $W^1[i]$ is the i-th row of weight matrix. $\cdot$ is the dot product (a sum over element-wise multiplications), $b^1 \in R^{100}$ is a bias term, and $f(\cdot)$ is an activation function (in this experiment, we use tanh-function as activation function). Thus we can get a $100 \times n$ feature graph

$$h^1 = (h_{i,j}^1)_{100\times n} \tag{4}$$

3) Pool layer of text CNN: Taking the output of convolution layer as the input of pooling layer, we use the max pooling (1-max pooling) operation as shown in Fig. 1(d), i.e., select the max value of the n elements of each row in feature graph matrix

$$h^1 : h_j^2 = \max_{1\le i\le n} h_{i,j}^1, j = 1, 2, \cdots, 100 \tag{5}$$

After max pooling, we obtain 100×1 features h^2. The reason of choosing max pooling operation is that the role of every word in the text is not completely equal, by maximum pooling we can choose the elements which play key role in the text. In spite of different length of the input training set samples, the text is converted into a fixed length vector after convolution

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2694446, IEEE Access

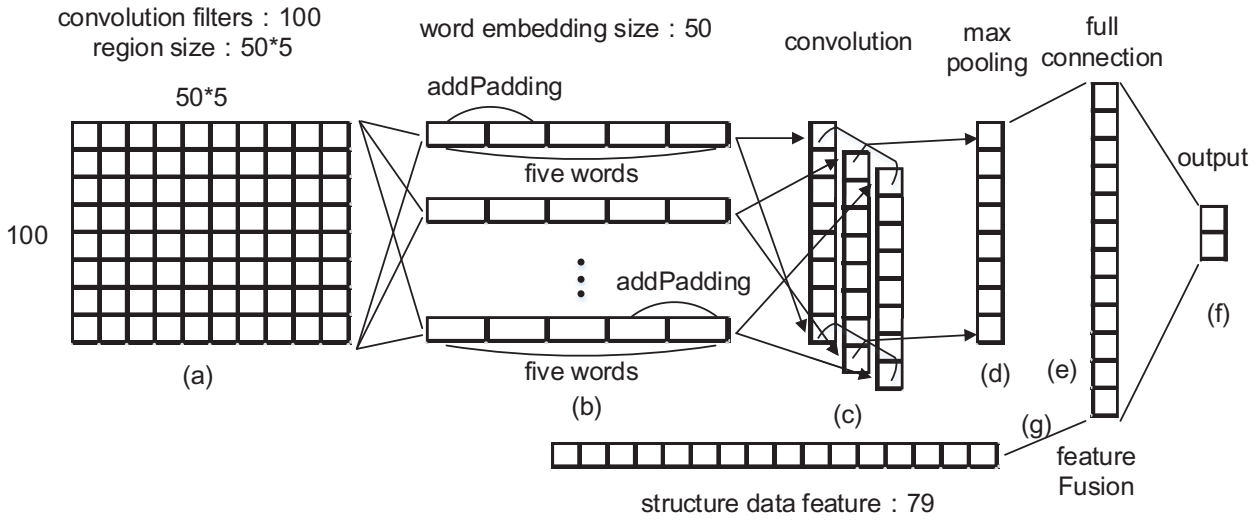


Fig. 1. CNN-based multimodal disease risk prediction (CNN-MDRP) algorithm.

layer and pooling layer, for example, in this experiment, after convolution and pooling, we get 100 features of the text.

4) Full connection layer of text CNN: Pooling layer is connected with a fully connected neural network as shown in Fig. 1(E), the specific calculation process is that:

$$h^3 = W^3 h^2 + b^3 \tag{6}$$

where h^3 is the value of the full connection layer, W^3 and b^3 is the corresponding weights and deviation.

5) CNN classifier: The full connection layer links to a classifier, for the classifier, we choose a softmax classifier, as shown in Fig. 1(f).

C. CNN-based Multimodal Disease Risk Prediction (CNN-MDRP) Algorithm

From what has been discussed above, we can get the information that CNN-UDRP only uses the text data to predict whether the patient is at high risk of cerebral infarction. As for structured and unstructured text data, we design a CNN-MDRP algorithm based on CNN-UDRP as shown in Fig. 1. The processing of text data is similar with CNN-UDRP, as shown in Fig. 1(a-d), which can extract 100 features about text data set. For structure data, we extract 79 features. Then, we conduct the feature level fusion by using 79 features in the S-data and 100 features in T-data, as shown in Fig. 1(g). For full connection layer, computation methods are similar with CNN-UDRP algorithm Since the variation of features number, the corresponding weight matrix and bias change to W^3_{new}, b^3_{new}, respectively. We also utilize softmax classifier. In the following we will introduce how to train the CNN-MDRP algorithm, the specific training process is divided into two parts.

1) Training word Embedding: Word vector training requires pure corpus, the purer the better, that is, it is better to use a professional corpus. In this paper, we extracted the text data of all patients in the hospital from the medical large data center. After cleaning these data, we set them as corpus set. Using ICTACLAS [28] word segmentation tool, word2vec [29] tool n-skip gram algorithm trains the word vector, word vector dimension is set to 50, after training we get about 52100 words in the word vector.

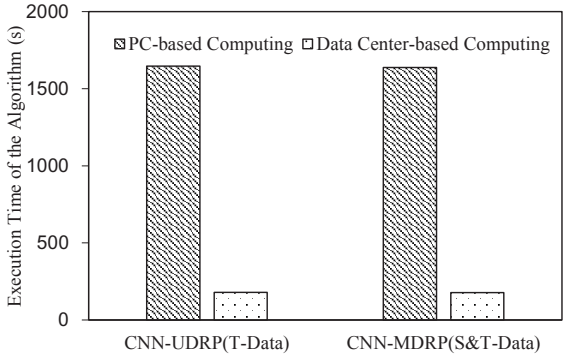


Fig. 2. Running time comparison of CNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithms in personal computer (PC) and data center.

2) Training parameters of CNN-MDRP: In CNN-MDRP algorithm, the specific training parameters are $W^1, W^3_{new}, b^1, b^3_{new}$. we use stochastic gradient method to train parameters, and finally reach the risk assessment of whether the patient suffers from cerebral infarction. Some advanced features shall be tested in future study, such as fractal dimension [30], biorthogonal wavelet transform [31], [32] etc.

IV. Experimental Results

In this section, we discuss the performance of CNN-UDRP and CNN-MDRP algorithms from several aspects, i.e.,the run time, sliding window, iterations and text feature.

A. Run Time Comparison

We compare the running time of CNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithms in personal computer (2core CPU, 8.00G RAM) and data center (6core*2*7=84core CPU, 48*7=336G RAM). Here, we set the same CNN iterations which are 100 and extract the same 100 text features.

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2694446, IEEE Access

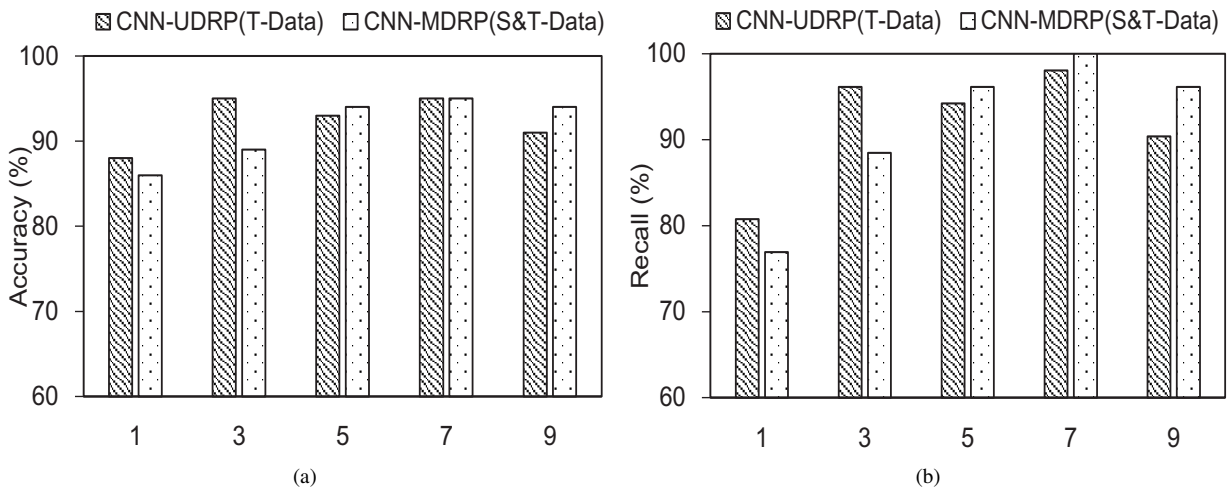


Fig. 3. Effect of sliding window (word number) in the algorithm. (a) The corresponding accuracy of the CNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithms when the number of words for sliding window are 1, 3, 5, 7 and 9. (b) The corresponding recall of the CNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithms when the number of words for sliding window are 1, 3, 5, 7 and 9.

As shown in Fig. 2, for CNN-UDRP (T-data) algorithm, the running time in data center is 178.5s while the time in personal computer is 1646.4s. For CNN-MDRP (S&T-data) algorithm, its running time in data center is 178.2s while the time in personal computer is 1637.2s. That is, the running speed of the data center is 9.18 times on the personal computer. Moreover, we can see the running time of CNN-UDRP (T-data) and CNN-MDRP (S&T-data) are basically the same from the figure, i.e. although the number of CNN-MDRP (S&T-data) features increase after adding structured data, it does not make a significant change in time. The later experiments are based on the running results of the data center.

B. Effect of Sliding Window (Word Number)

When taking convolution of CNN, we need to confirm the number of words for sliding window first. In this experiment, the selected number of words for the sliding window are 1, 3, 5, 7 and 9. The iterations of CNN are 200 and the size of convolution kernel is 100. As shown in Fig. 3, when the number of words for the sliding window are 7, the accuracy and recall of CNN-UDRP (T-data) algorithm are 0.95 and 0.98, respectively. And the accuracy and recall of CNN-MDRP (S&T-data) algorithm are 0.95 and 1.00. These results are all higher than we choose other number of words for sliding window. Thus, in this paper, we choose the number of words for sliding window are 7.

C. Effect of Iterations

We give out the change of the training error rate and test accuracy along with the number of iterations. As shown in Fig. 4, with the increase of the number of iterations, the training error rate of the CNN-UDRP (T-data) algorithm decreases gradually, while test accuracy of this method increases. The CNN-MDRP (S&T-data) algorithm have the similar trend in terms of the training error rate and test accuracy. In Fig. 4, we can also obtain when the number of iterations are 70, the training process of CNN-MDRP (S&T-data) algorithm is already stable while the CNN-UDRP (T-data) algorithm is still not stable. In other words, the training time of MDRP(S&T data) algorithm is shorter, i.e. the convergence speed of CNN-MDRP (S&T-data) algorithm is faster.

D. Effect of Text Features

The number of features extracted from structured data is certain, i.e. 79 features. However, the feature number of unstructured text data extracted by CNN is uncertain. Thus, we research the effect of text feature number on accuracy and recall of CNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithms. We extract $10, 20, \cdots, 120$ features from text by using CNN. Fig. 5 shows the accuracy and recall of each feature after it go through 200 times of iteration. From the Fig. 5(a) and Fig. 5(b), when the feature number of text is smaller than 30, the accuracy and recall of CNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithms are smaller than the feature number of text is bigger than 30 obviously. This is because it is not able to describe a large number of useful information contained in the text when the text feature number is relatively small. Moreover, in the Fig. 5(a), the accuracy of CNN-MDRP (S&T-data) algorithm is more stable than CNN-UDRP (T-data) algorithm, i.e. the CNN-MDRP (S&T-data) algorithm is reduced fluctuation after adding structured data. As shown in Fig. 5(b), after adding structured data, the recall of CNN-MDRP (S&T-data) algorithm is higher than CNN-UDRP (T-data) algorithm obviously. This shows that the recall of algorithm is improved after adding structured data.

V. Analysis of Overall Results

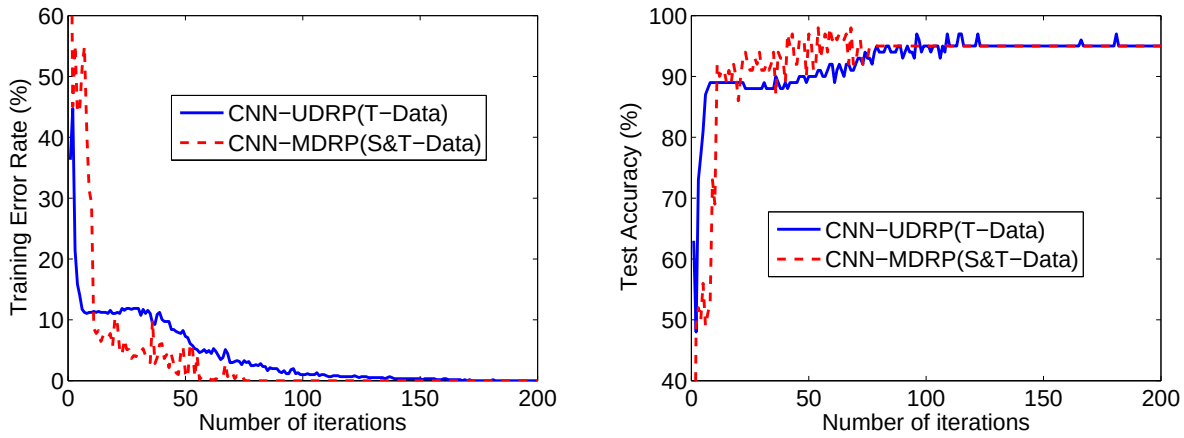
In this section, we describe the overall results about S-data and S&T-data.

A. Structured Data (S-data)

For S-data, we use traditional machine learning algorithms, i.e., NB, KNN and DT algorithm to predict the risk of cerebral

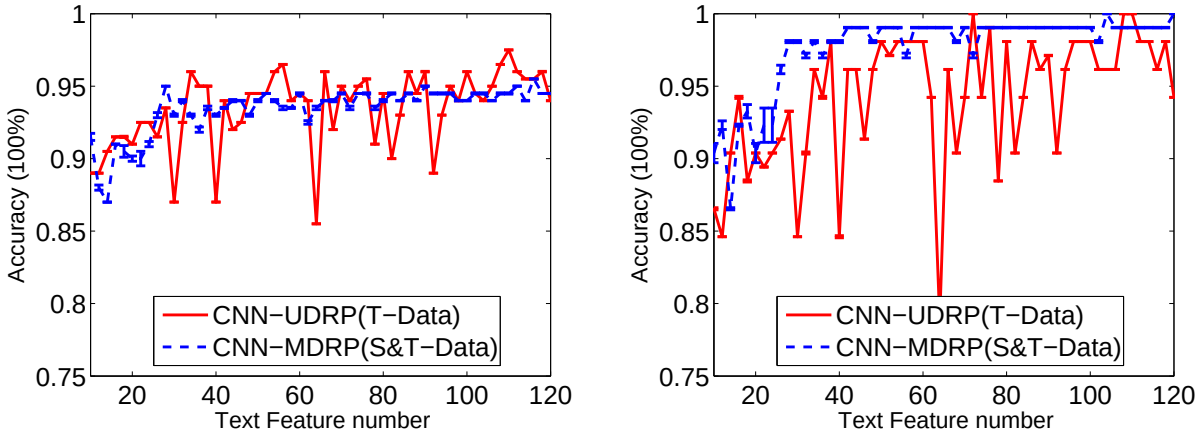
2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2694446, IEEE Access



(a) (b)

Fig. 4. Effect of iterations on the algorithm. (a) The trend of training error rate with the iterations for CNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithms. (b) The trend of test accuracy with the iterations for CNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithms.



(a) (b)

Fig. 5. Effect of text features on the algorithm. (a) The accuracy trend of the CNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithms along with the increased number of text features. (b) The recall trend of the CNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithms along with the increased number of features.

infarction disease. NB classification is a simple probabilistic classifier. It requires to calculate the probability of feature attributes. In this experiment, we use conditional probability formula to estimate discrete feature attributes and Gaussian distribution to estimate continuous feature attributes. The KNN classification is given a training data set, and the closest k instance in the training data set is found. For KNN, it is required to determine the measurement of distance and the selection of k value. In the experiment, the data is normalized at first. Then we use the Euclidean distance to measure the distance. As for the selection of parameters k, we find that the model is the best when k = 10. Thus, we choose k = 10. We choose classification and regression tree (CART) algorithm among several decision tree (DT) algorithms.

To determine the best classifier and improve the accuracy of the model, the 10-fold cross-validation method is used for the training set, and data from the test set are not used in the training phase. The model's basic framework is shown in Fig. 6. The results are shown in Fig. 7(a) and Fig. 7(b). From Fig. 7(a), we can see that the accuracy of the three machine learning algorithms are roughly around 50%. Among them, the accuracy of DT which is 63% is highest, followed by NB and KNN. The recall of NB is 0.80 which is the highest, followed by DT and KNN. We can also draw from Fig. 7(b) that the corresponding AUC of NB, KNN and DB are 0.4950, 0.4536 and 0.6463, respectively. In summary, for S-data, the NB classification is the best in experiment. However, it is also observed that we cannot accurately predict whether the patient is in a high risk of cerebral infarction according to the patient's age, gender, clinical laboratory and other structured data. In other word, because cerebral infarction is a disease with complex symptom, we cannot predict whether the patient is in a high risk group of cerebral infarction only in the light of these simple features.

B. Structured and Text Data (S&T-data)

According to the discussion in Section IV, we give out the accuracy, precision, recall, F1-measure and ROC curve under

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2694446, IEEE Access

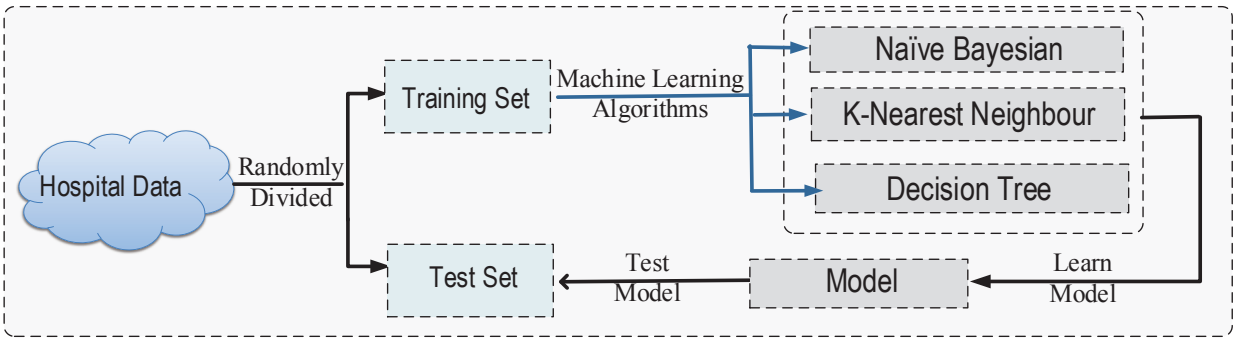
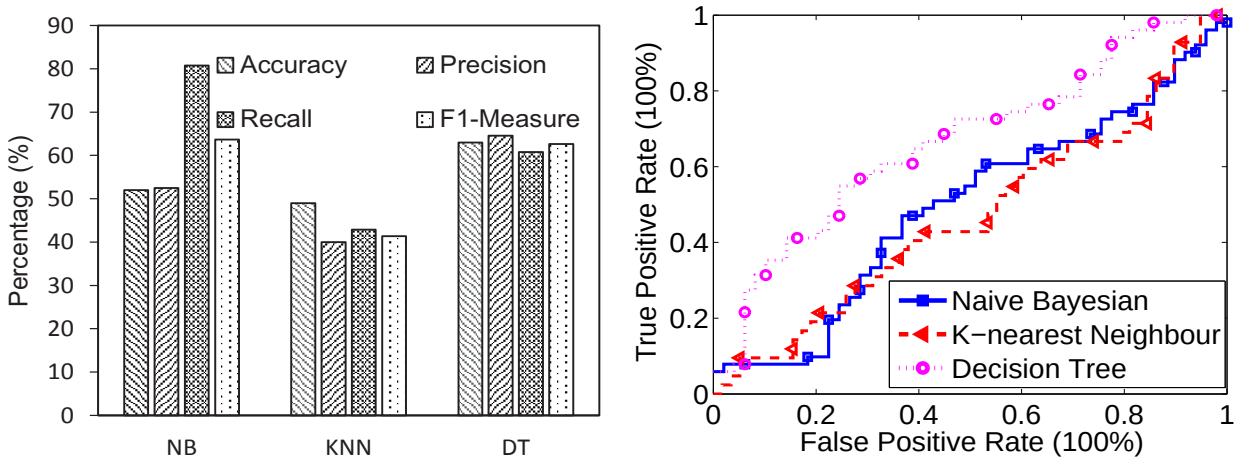
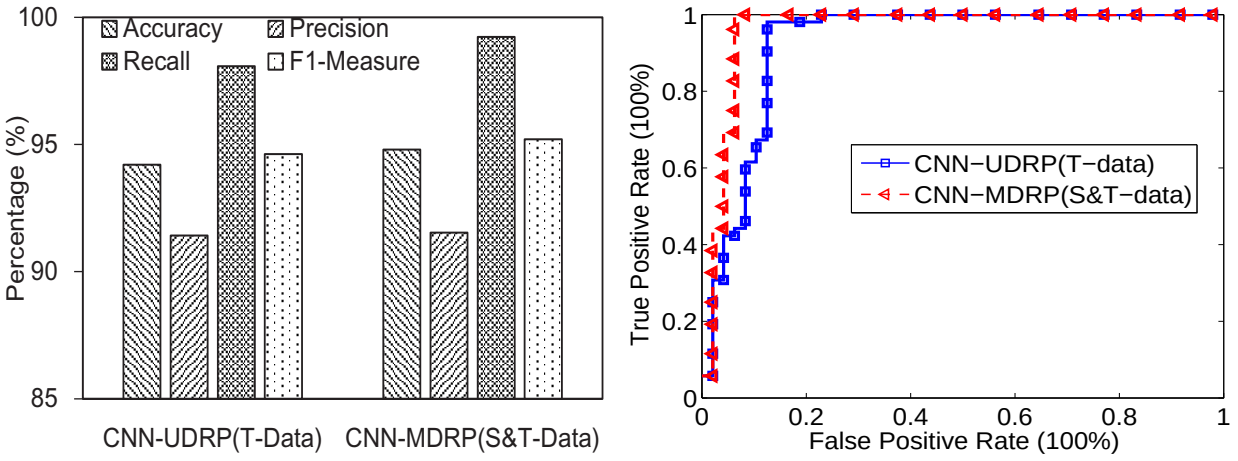


Fig. 6. The three machine learning algorithms userd in our disease prediction experiments.



(a) (b)

Fig. 7. Overall results of S-data. (a) Comparison of accuracy, precision, recall and F1-Measure under S-data for NB, KNN and DT, in which NB = naive Bayesian, KNN = k-nearest neighbour, and DT = decision tree. (b) ROC curves under S-data for NB, KNN and DT.



(a) (b)

Fig. 8. Overall results of S&T-data. (a) Comparison of accuracy, precision, recall and F1-measure under CNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithms. (b) ROC curves under CNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithms.

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2694446, IEEE Access

CNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithms. In this experiment, the selected number of words is 7 and the text feature is 100. As for CNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithms, we both run 5 times and seek the average of their evaluation indexes. From the Fig. 8, the accuracy is 0.9420 and the recall is 0.9808 under CNN-UDRP (T-data) algorithm while the accuracy is 0.9480 and the recall is 0.99923 under CNN-MDRP (S&T-data) algorithm. Thus, we can draw the conclusion that the accuracy of CNN-UDRP (T-data) and CNN-MDRP (S&T-data) algorithms have little difference but the recall of CNN-MDRP (S&T-data) algorithm is higher and its convergence speed is faster. In summary, the performance of CNN-MDRP (S&T-data) is better than CNN-UDRP (T-data).

In conclusion, for disease risk modelling, the accuracy of risk prediction depends on the diversity feature of the hospital data, i.e., the better is the feature description of the disease, the higher the accuracy will be. For some simple disease, e.g., hyperlipidemia, only a few features of structured data can get a good description of the disease, resulting in fairly good effect of disease risk prediction [33]. But for a complex disease, such as cerebral infarction mentioned in the paper, only using features of structured data is not a good way to describe the disease. As seen from Fig. 7(a) and Fig. 7(b), the corresponding accuracy is low, which is roughly around 50%. Therefore, in this paper, we leverage not only the structured data but also the text data of patients based on the proposed CNN-MDPR algorithm. We find that by combining these two data, the accuracy rate can reach 94.80%, so as to better evaluate the risk of cerebral infarction disease.

VI. Conclusion

In this paper, we propose a new convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm using structured and unstructured data from hospital. To the best of our knowledge, none of the existing work focused on both data types in the area of medical big data analytics. Compared to several typical prediction algorithms, the prediction accuracy of our proposed algorithm reaches 94.8% with a convergence speed which is faster than that of the CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm.

Acknowledgement

This research was supported by the National Natural Science Foundation of China (under Grant No. 61572220, and Grant No. 81671904), the International Science and Technology Corporation Program of Chinese Ministry of Science and Technology S2014ZR0340.

References

[1] P. Groves, B. Kayyali, D. Knott, and S. V. Kuiken, "The'big data'revolution in healthcare: Accelerating value and innovation," 2016.

[2] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, 2014.

[3] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," *Nature Reviews Genetics*, vol. 13, no. 6, pp. 395–405, 2012.

[4] D. Tian, J. Zhou, Y. Wang, Y. Lu, H. Xia, and Z. Yi, "A dynamic and self-adaptive network selection method for multimode communications in heterogeneous vehicular telematics," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 6, pp. 3033–3049, 2015.

[5] M. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, C. Youn, "Wearable 2.0: Enable Human-Cloud Integration in Next Generation Healthcare System," *IEEE Communications*, Vol. 55, No. 1, pp. 54–61, Jan. 2017.

[6] M. Chen, Y. Ma, J. Song, C. Lai, B. Hu, "Smart Clothing: Connecting Human with Clouds and Big Data for Sustainable Health Monitoring," ACM/Springer Mobile Networks and Applications, Vol. 21, No. 5, pp. 825C845, 2016.

[7] M. Chen, P. Zhou, G. Fortino, "Emotion Communication System," *IEEE Access*, DOI: 10.1109/ACCESS.2016.2641480, 2016.

[8] M. Qiu and E. H.-M. Sha, "Cost minimization while satisfying hard/soft timing constraints for heterogeneous embedded systems," *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, vol. 14, no. 2, p. 25, 2009.

[9] J. Wang, M. Qiu, and B. Guo, "Enabling real-time information service on telehealth system over cloud-based big data platform," *Journal of Systems Architecture*, vol. 72, pp. 69–79, 2017.

[10] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, "Big data in health care: using analytics to identify and manage high-risk and high-cost patients," *Health Affairs*, vol. 33, no. 7, pp. 1123–1131, 2014.

[11] L. Qiu, K. Gai, and M. Qiu, "Optimal big data sharing approach for tele-health in cloud computing," in *Smart Cloud (SmartCloud), IEEE International Conference on*. IEEE, 2016, pp. 184–189.

[12] Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, "Health-cps: Healthcare cyber-physical system assisted by cloud and big data," *IEEE Systems Journal*, 2015.

[13] K. Lin, J. Luo, L. Hu, M. S. Hossain, and A. Ghoneim, "Localization based on social big data analysis in the vehicular networks," *IEEE Transactions on Industrial Informatics*, 2016.

[14] K. Lin, M. Chen, J. Deng, M. M. Hassan, and G. Fortino, "Enhanced fingerprinting and trajectory prediction for iot localization in smart buildings," *IEEE Transactions on Automation Science and Engineering*, vol. 13, no. 3, pp. 1294–1307, 2016.

[15] D. Oliver, F. Daly, F. C. Martin, and M. E. McMurdo, "Risk factors and risk assessment tools for falls in hospital in-patients: a systematic review," *Age and ageing*, vol. 33, no. 2, pp. 122–130, 2004.

[16] S. Marcoon, A. M. Chang, B. Lee, R. Salhi, and J. E. Hollander, "Heart score to further risk stratify patients with low timi scores," *Critical pathways in cardiology*, vol. 12, no. 1, pp. 1–5, 2013.

[17] S. Bandyopadhyay, J. Wolfson, D. M. Vock, G. Vazquez-Benitez, G. Adomavicius, M. Elidrisi, P. E. Johnson, and P. J. O'Connor, "Data mining for censored time-to-event data: a bayesian network model for predicting cardiovascular risk from electronic health record data," *Data Mining and Knowledge Discovery*, vol. 29, no. 4, pp. 1033–1069, 2015.

[18] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, "A relative similarity based method for interactive patient risk prediction," *Data Mining and Knowledge Discovery*, vol. 29, no. 4, pp. 1070–1093, 2015.

[19] A. Singh, G. Nadkarni, O. Gottesman, S. B. Ellis, E. P. Bottinger, and J. V. Guttag, "Incorporating temporal ehr data in predictive models for risk stratification of renal function deterioration," *Journal of biomedical informatics*, vol. 53, pp. 220–228, 2015.

[20] J. Wan, S. Tang, D. Li, S. Wang, C. Liu, H. Abbas and A. Vasilakos, "A Manufacturing Big Data Solution for Active Preventive Maintenance", *IEEE Transactions on Industrial Informatics*, DOI: 10.1109/TI-I.2017.2670505, 2017.

[21] W. Yin and H. Schütze, "Convolutional neural network for paraphrase identification." in *HLT-NAACL*, 2015, pp. 901–911.

[22] N. Nori, H. Kashima, K. Yamashita, H. Ikai, and Y. Imanaka, "Simultaneous modeling of multiple diseases for mortality prediction in acute hospital care," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 855–864.

[23] S. Zhai, K.-h. Chang, R. Zhang, and Z. M. Zhang, "Deepintent: Learning attentions for online advertising with recurrent neural networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1295–1304.

[24] K. Hwang, M. Chen, "Big Data Analytics for Cloud/IoT and Cognitive Computing," Wiley, U.K., ISBN: 9781119247029, 2017.

[25] H. Chen, R. H. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact." *MIS quarterly*, vol. 36, no. 4, pp. 1165–1188, 2012.

[26] S. Basu Roy, A. Teredesai, K. Zolfaghar, R. Liu, D. Hazel, S. Newman, and A. Marinez, "Dynamic hierarchical classification for patient risk-of-readmission," in *Proceedings of the 21th ACM SIGKDD international*

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2694446, IEEE Access

conference on knowledge discovery and data mining. ACM, 2015, pp. 1691–1700.

[27] J. C. Ho, C. H. Lee, and J. Ghosh, "Septic shock prediction for patients with missing data," *ACM Transactions on Management Information Systems (TMIS)*, vol. 5, no. 1, p. 1, 2014.

[28] "Ictclas," http://ictclas.nlpir.org/.

[29] "word2vec," https://code.google.com/p/word2vec/.

[30] Y.-D. Zhang, X.-Q. Chen, T.-M. Zhan, Z.-Q. Jiao, Y. Sun, Z.-M. Chen, Y. Yao, L.-T. Fang, Y.-D. Lv, and S.-H. Wang, "Fractal dimension estimation for developing pathological brain detection system based on minkowski-bouligand method," *IEEE Access*, vol. 4, pp. 5937–5947, 2016.

[31] Y.-D. Zhang, Z.-J. Yang, H.-M. Lu, X.-X. Zhou, P. Phillips, Q.-M. Liu, and S.-H. Wang, "Facial emotion recognition based on biorthogonal wavelet entropy, fuzzy support vector machine, and stratified cross validation," *IEEE Access*, vol. 4, pp. 8375–8385, 2016.

[32] S.-H. Wang, T.-M. Zhan, Y. Chen, Y. Zhang, M. Yang, H.-M. Lu, H.-N. Wang, B. Liu, and P. Phillips, "Multiple sclerosis detection based on biorthogonal wavelet transform, rbf kernel principal component analysis, and logistic regression," *IEEE Access*, vol. 4, pp. 7567–7576, 2016.

[33] S.-M. Chu, W.-T. Shih, Y.-H. Yang, P.-C. Chen, and Y.-H. Chu, "Use of traditional chinese medicine in patients with hyperlipidemia: A population-based study in taiwan," *Journal of ethnopharmacology*, vol. 168, pp. 129–135, 2015.

2169-3536 (c) 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.