Predicting Chronic Diseases with Health IT: a Survey on Popular Techniques

Tongbin Zhang[#1], Li Cai[#2], Chuandi Pan[*3]

[#1]*Student, School of the 1st Clinical Medical Sciences, School of Information and Engineering, Wenzhou Medical University, China*

[#2]*Manager, Wenzhou Yuekang Information Technology Limited Liability Company, China*

[*3]*Professer, Department of Computer Technology and Information, the First Affiliated Hospital of Wenzhou Medical University, China*

Abstract

In this paper, we study and compare popular predictive techniques to predict the chronic diseases with the support of health information technology. Interestingly, we show that there is no technique guaranteeing the good predictive outcomes for all diseases. In many cases, the well-known state-of-the-art techniques, such as support vector machine, was significantly outperformed by simpler classical techniques. We also show that using feature selection would improve the predictive performance. However, choosing the right predictive technique is still the crucial factor. Therefore, in health information technology industrial practice, the predictive healthcare system should change from only relying on only one technique to integrating multiple techniques in case-study basis.

Keywords: *Chronic disease prediction, Health IT, Random Forest*

I. INTRODUCTION

Health information technology (Health IT) has been actively participating in improving the quality of healthcare [1, 2]. After many years on success in assisting the management and health data organization, it is claimed that Health IT should aim for more ambitious goal in disease-predictive tasks. However, Health IT has not been very widely and successfully applied in this new area [3-6]. Challenges in mining EHR data include noise, heterogeneity, sparseness, incompleteness, random errors, and systematic biases [7, 8]. Well-known predictive techniques have not been analysed and built to meet these challenges. Therefore, in addition to feature selection and data representation [9, 10], the success of predictive algorithms largely depends on selecting the right technique to predict the right disease.

A typical example for the challenge is one of the latest experiments in predicting future disease done by Google [3]. In this work, the Google team use Deep Learning, the technique winning human in the difficult Go game [11], and Random Forest [12] in predicting new occurrence of diseases from the electronic health record (EHR). Here, the Deep Learning work shows some successes in applying this approach, in which the classification accuracy achieves more than 90% on average. However, in most chronic diseases, the predictive performance is not high, such as in Breast Cancer and Hypertension (accuracy less than 75%) [3]. This performance is already better than the ones using dimensional reduction methods [13-16] and other state-of-the-art algorithms such as Tree-Lasso [17] and Elastic Net [18]. In the other hand, the work in [19] shows that in Diabetes prediction, support vector machine [20], another modern and popular technique, does not show better performance than the classical decision table technique [21]. From these examples, we can see an important point: there is no 'universal' technique good for all types of disease prediction.

In this work, we examine and compare the predictive performance of many techniques in chronic disease predictions. The techniques included in this work are: Decision Table [21], Support Vector Machine [20], Random Forest [12], Artificial Neural Network [22], Random Tree [23] and Hoeffding Tree [24] implemented in Weka [25]. We show that with just the combination of classical statistical t-test [26] for feature selection and these techniques, we could achieve high classification performance in many chronic diseases. We conduct the experiment using the outpatient EHR data at the First Affiliated Hospital, Wenzhou Medical University, Zhejiang, China. Since the data set only spans for nearly 4 years, which is short and small, we could not find enough cases for future disease prediction. Therefore, the work is limited to the disease versus control classification problem.

II. METHODS

A. Acquire and preprocess data

To examine chronic disease and lab test association, we acquire the outpatient dataset from the 1st Affiliated Hospital (1AH), Wenzhou Medical University, Zhejiang, China. Among the data sectors at the 1AH, the outpatient contains the highest number of chronic-disease patients with multiple follow-up visits for further validation. In this work, the chronic disease outpatient EHR is collected between October 2010 and August 2014, specified by the research sponsor. The dataset contains

information on 16,310 patients with chronic diseases (identified by ICD code version 10 [27]). There are 73 unique ICD codes for chronic diseases; however, since one disease may have multiple ICD codes. By manual checking, we find that the dataset covers 29 chronic diseases. The patients' demographic information is completely removed, except the gender, according to the patient privacy regulation in China and the requirement of the research sponsor. These patients made 265,903 visits (identified by visit number) between 2010 and 2014 (averagely 16 visits per patient). We show the number of visits per patient distribution in figure 1a. Figure 1b shows the distribution of comorbidity size per patient. Among these, 1,919 patients only had one visit; therefore, we do not use these patients' information in the analysis. 9,746 patients only had one chronic disease; meanwhile, 6,564 patients showed comorbidity among at least two diseases.

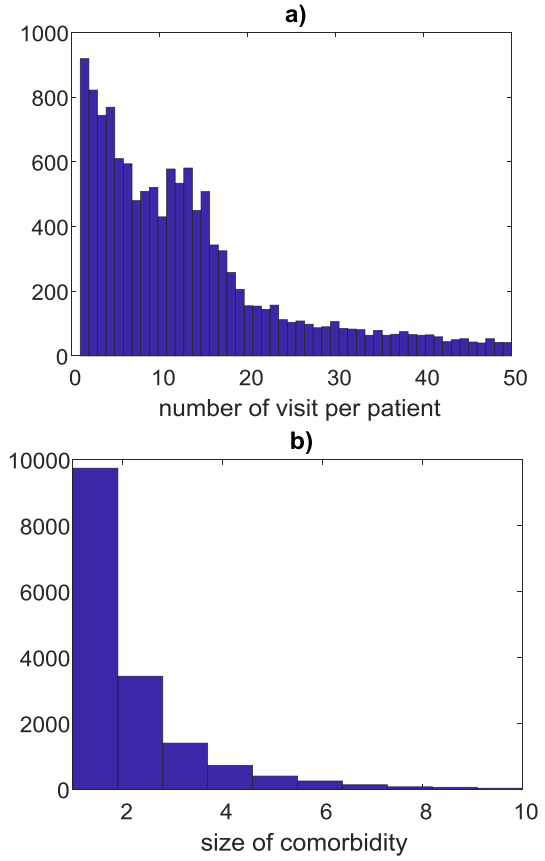

Fig.1 Distribution of number of visit (a) and comorbidity size (b) for each patient

In addition, to form the control class for the statistical analysis, we acquire the EHR from 1000 subjects (500 males and 500 females) who show no abnormality between 2010 and 2014. By no abnormality, we mean that for each subject, all lab test results are marked as 'normal' according to the medical lab test standard at 1AH, and the subject shows no disease between 2010 and 2014. These subjects made 1125 visits. These subjects had neither inpatient nor outpatient visits at 1AH. Therefore, by the scope of the project, we may assume that they are healthy subjects. We choose the control class subject such that their checkup visits are uniformly distributed between 2010 and 2014.

There are 1521 lab tests appearing in the dataset, identified by the lab test ID from the 1AH. Among these lab tests, we only use 47 tests taken by at least 30 outpatients and 30 healthy subjects for analysis to ensure the quality of statistical analysis. We manually translate the test names from Chinese into English and re-identify these tests because some tests have multiple test ID at 1AH. The re-identification ensures that each test has a unique ID, resulting in 46 unique tests.

B. Identify and validate disease-test associations

For the validation purpose, for each disease (positive class), we divide the dataset into two the training set and test set, as shown in figure 2. The training set only contains patients having discovered date, or the earliest date when the patient was diagnosed with the disease, prior to January 1, 2014; while the test set only contains patients having discovered date after January 1, 2014. Then, for each disease analysis, we setup the feature table as follow. In the feature table, each patient represents a row in the table; while each lab test represents a column. For each entry in the table, we only choose the latest available test results after 2 months prior to the discovered date, as shown in figure 3. We adopt this selection since a patient may take different lab tests at different visit after having the disease. In other words, the data may contain missing values. Thus, we mark entries having available test results as 'known', and 'unknown' otherwise. In addition, for the control class, the training set contains subjects whose earliest visit date is prior to January 1, 2014; while the test set contains subjects whose earliest visit date is after January 1, 2014. We also setup the feature table for this class similar to the positive class.

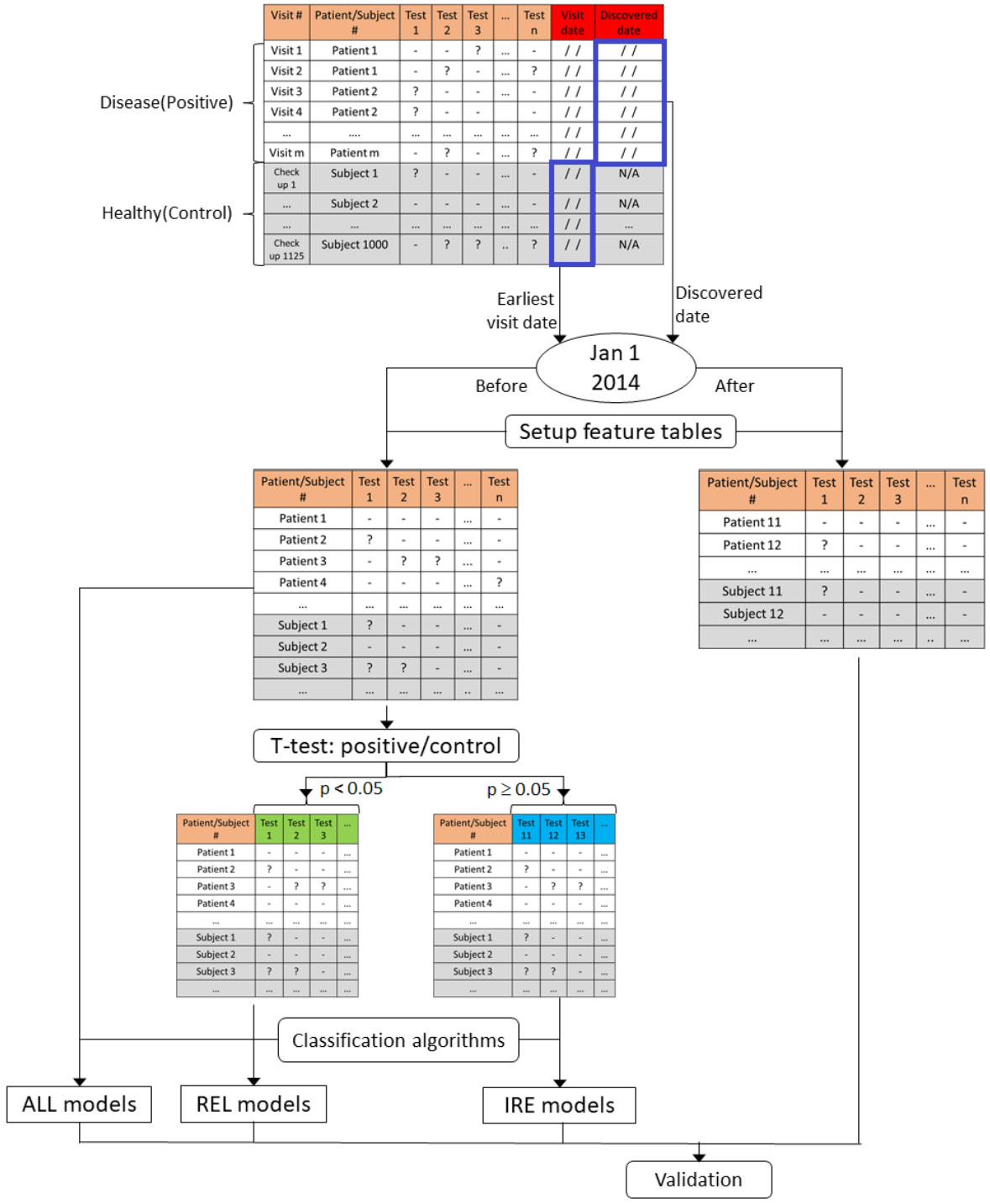

Fig 2 The overall framework in this paper: dividing the dataset into training and test set, setting up features table, finding association between disease and lab tests and validating the result by different classification models. Here, table entry '-' implies that the entry value is known; table entry '?' implies that the entry value is unknown.

We apply statistical and machine learning techniques to detect and validate the disease-lab test associations. To mine the disease-lab test association, we apply the student t-test [27]. As showed in figure 2, for each disease, we define that tests resulting in t-test p-value < 0.05 between the disease and the control classes are associated with the disease. We train the classification models using the training set and measure the performance on the test set, as shown in the above section. For prediction, we use the techniques Decision Table [21], Support Vector Machine [20], Random Forest [12], Artificial Neural Network [22] and Hoeffding Tree [24] (showed in the introduction) implemented in Weka version 3.8 [25].

Disease X

Visit #	Patient #	Test 1	Test 2	Test 3	Visit date	Discovered date
Visit 1	Patient 1	0.1	1	10	Feb 5 2012	Jan 1 2013
Visit 2	Patient 1	0.2	2	?	Dec 24 2012	Jan 1 2013
Visit 3	Patient 1	0.3	3	?	Feb 25 2013	Jan 1 2013
Visit 4	Patient 1	0.4	?	?	Mar 1 2014	Jan 1 2013

Patient #	Test 1	Test 2	Test 3
Patient 1	0.4	3	?

Fig.3 a toy example of setting up the feature table with Disease X and Patient 1 having 4 visits. Here, Visit 1 is not used because the visit date is more than 2 months before X discovered date. Entries for column Test 1, Test 2 and Test 3 are the latest available test results after 2 months of X discovered date. '?' implies that the test result is unknown.

Before the classification, for each disease, we normalize the feature table as follow. For each test, from the 'known' entries, we use the z-score normalization [28-30] to transform them to

$$z_{i,j} = \frac{x_{i,j} - m_j}{s_j} \qquad (1)$$

In this formula, i stands for patient/subject index, j stands for lab test index, $x_{i,j}$ is the patient i's lab test result j, $z_{i,j}$ is the normalization of $x_{i,j}$, m_j is the mean result of test j , s_j is the standard deviation result of test j. By the z-score normalization, the expected normalized test result is 0.

III.RESULTS

A. Comparison of predictive performance among the techniques

Figure 4 shows that the Random Forest techniques, overall, achieves the best predictive performance in both accuracy and AUC. In addition, the performance of Random Forest does not vary much among different diseases. Furthermore, the lowest accuracy of Random Forest is 0.7. These facts suggest that the future development of chronic disease prediction should take Random Forest as the 'benchmark' technique. The Decision Table and Random Tree achieve the AUC similar AUC to Random Forest, but much less accuracy. Surprisingly, the more modern techniques with long-time and strong theoretical support: Support Vector Machine and Artificial Neural Network, show the lowest performance.

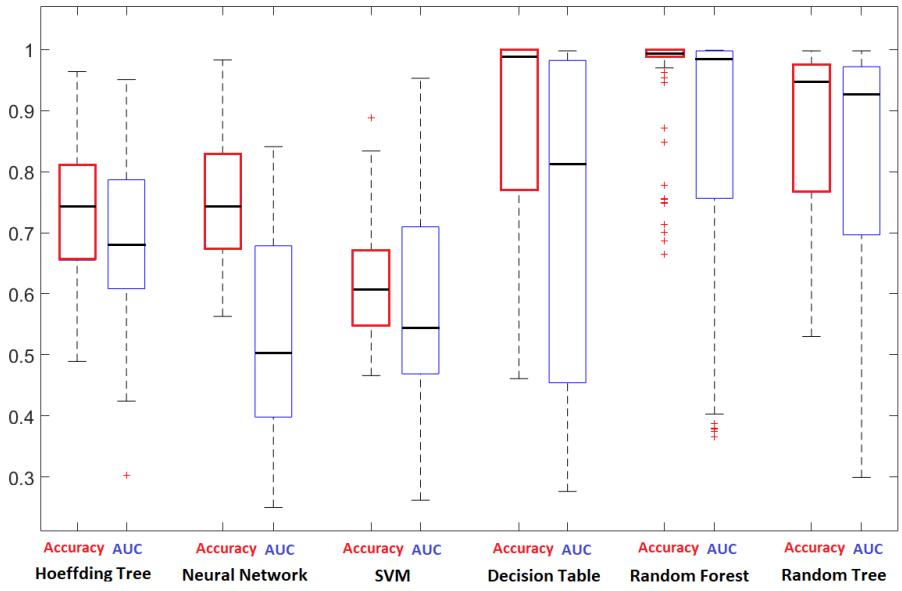

Fig. 4 Overall predictive performances of the techniques.

B. There is no 'universal' the best technique to predict the chronic disease

In figure 5, we show that by the AUC metric, there is no technique that always performs the best for all diseases. The Random Forest is the best one for most of the diseases except Adenomyosis, Chronic Prostatitis, Chronic Glomerulonephritis, Osteoporosis, Arthropathy, Rheumatoid Arthritis, etc. The Support Vector Machine, which is commonly used as the benchmark technique, does not show that it is the best technique in any disease. Therefore, we suggest that this technique should not be used as the 'benchmark' technique in disease prediction.

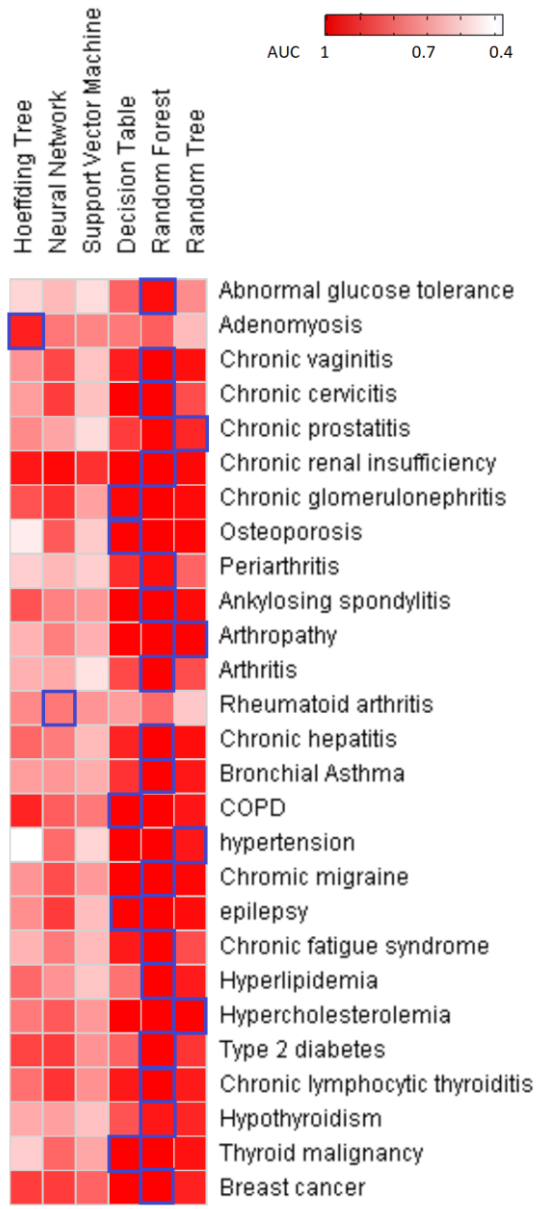

Figure 5 Heatmap of AUC in predicting different diseases: comparison among different techniques.

C. Feature selection improve the predictive performance in most of the techniques

Figure 6 shows that not every predictive technique benefits from feature selection. While Support Vector Machine, Random Tree and Decision Table get significantly better performance when using feature selection, Hoeffding Tree and Neural Network do not. Interestingly, with Random Forest, the one having the best performance in this work, feature selection decreases the predictive performance slightly.

IV. CONCLUSIONS

In this work, we have shown that there is no 'universal' technique achieving the good performance in predicting many chronic diseases in Health information technology. Therefore, the disease predictive system should be built by the integration of many techniques instead of relying on only one technique. However, among these techniques, the Random Forest stands out as the best one in most of the prediction. Therefore, we suggest that for future development of disease prediction model, the Random Forest should be the first option when deciding which predictive technique to integrate. In addition, Random Forest may not need feature selection to get the good performance. In the other hands, the classification performance using associated tests is high and better than the results showed in some state-of-the-art work. However, it is not necessary that the method presented in this work is better, because the work in [3] is completed in a more comprehensive data with longer duration, which allows levelling up the problem to predicting future disease occurrence.

There are several limitations in this work. First, due to the data provider's and project requirement, we only have the data spanning within 4 years. The short duration does not allow truly solving the future disease prediction problem. Second, the data set is originally in Chinese; in addition, the data provider does not apply international standards to identify disease and lab tests fully. Therefore, translation and cleaning up the disease and lab test terminology must be done manually, which may be error-prone.

ACKNOWLEDGEMENT

The authors thank Dr. Thanh Minh Nguyen from the Informatics Institute – the University of Alabama at Birmingham for help editing this manuscript.

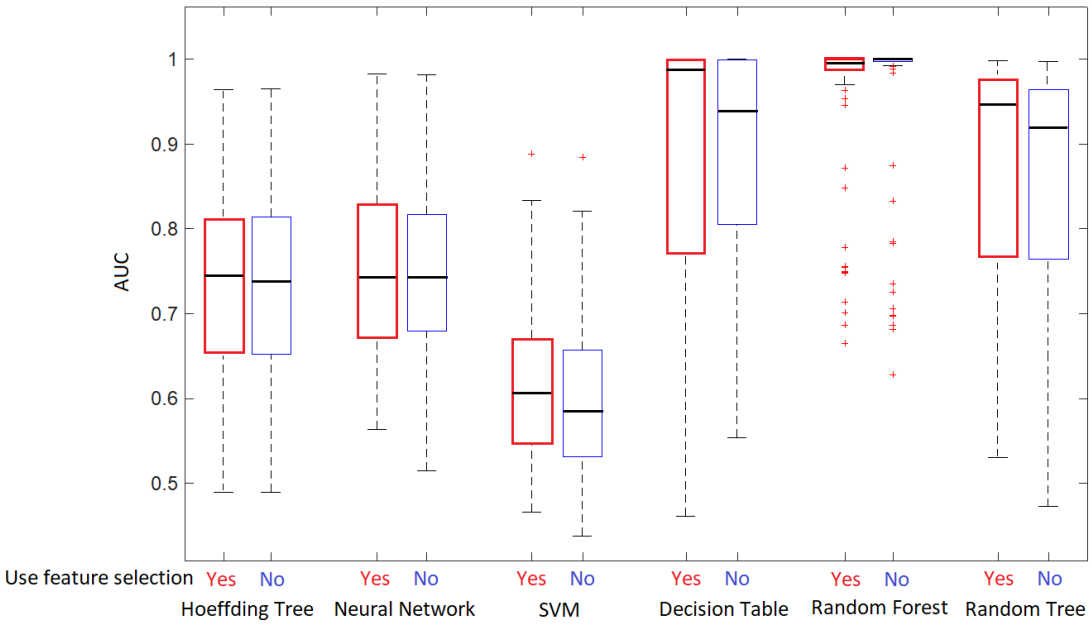

Fig. 6 Predictive performance with feature selection (Yes) and without feature selection (No)

REFERENCES

[1] FK Weigel, TL Switaj, J Hamilton. Leveraging Health information technology to improve quality in federal healthcare. US Army Med Dep J, Oct-Dec, 2015: 68-74

[2] TT Chen, CD Pan. Design and Realization of Digital Intensive Care System. International Journal of Engineering Trends and Technology, vol. 36, no. 7, May-Jun, 2016: 337-342

[3] R Miotto, L Li, BA Kidd, et al. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep, vol. 6, pp. 26094, May 17, 2016.

[4] R Bellazzi, B Zupan. Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform, vol. 77, no. 2, Feb, 2008: 81-97

[5] PB Jensen, LJ Jensen, S Brunak. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet, vol. 13, no. 6, Jun, 2012: 395-405

[6] D Dahlem, D Maniloff, C Ratti. Predictability Bounds of Electronic Health Records. Sci Rep, vol. 5, pp. 11865, Jul 7, 2015.

[7] NG Weiskopf, G Hripcsak, S Swaminathan, et al. Defining and measuring completeness of electronic health records for secondary use. J Biomed Inform, vol. 46, no. 5, Oct, 2013: 830-836

[8] NG Weiskopf, C Weng. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc, vol. 20, no. 1, Jan 1, 2013:144-151

[9] Y Bengio, A Courville, P Vincent. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell, vol. 35, no. 8, Aug, 2013: 1798-1828

[10] MI Jordan, TM Mitchell. Machine learning: Trends, perspectives, and prospects. Science, vol. 349, no. 6245, Jul 17, 2015: 255-260

[11] D Silver, J Schrittwieser, K Simonyan, et al. Mastering the game of Go without human knowledge . Nature, vol. 550, no. 7676, pp. 354, 2017.

[12] A Liaw, M Wiener. Classification and regression by randomForest. R news, vol. 2, no. 3, 2002: 18-22

[13] SH Huang, P LePendu, SV Iyer, et al. Toward personalizing treatment for depression: predicting diagnosis and severity. J Am Med Inform Assoc, vol. 21, no. 6, Nov-Dec, 2014: 1069-75

[14] S Lyalina, B Percha, P LePendu, et al. Identifying phenotypic signatures of neuropsychiatric disorders from electronic medical records . J Am Med Inform Assoc, vol. 20, no. e2, Dec, 2013: 297-305

[15] R Bro, AK Smilde. Principal component analysis. Analytical Methods, vol. 6, no. 9, 2014: 2812-2831

[16] L Van Der Maaten, G Hinton. Visualizing data using t-sne (2008). J Mach Learn Res, vol. 1117, no., 2017: 2579-2605

[17] L Kamkar, SK Gupta, D Phung. Stable feature selection for clinical prediction: exploiting ICD tree structure using Tree-Lasso. J Biomed Inform, vol. 53, Feb, 2015: 277-90

[18] BJ Marafino, WJ Boscardin, RA Dudley. Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to ICU risk stratification from nursing notes. J Biomed Inform, vol. 54, Apr, 2015: 114-120

[19] N Cao, S Zeng, F Shen, et al. Predictive and Preventive Models for Diabetes Prevention using Clinical Information in Electronic Health Record . In IEEE International Conference on Bioinformatics and Biomedicine, Washington DC, 2015.

[20] V Vapnik, SE Golowich, AJ Smola. Support vector method for function approximation, regression estimation and signal processing. Advances in neural information processing system, Feb, 1970: 281-287

[21] R Kohavi. The power of decision tables. Machine Learning: ECML-95, Springer, 1995: 174-189:

[22] R Hecht-Nielsen. Theory of the backpropagation neural network. Neural networks for perception, Elsevier, 1992: 65-93

[23] T Duquesne, JFL Gall. Random trees, levy processes and spatial branching processes. Mathematics, October 2005:113-116

[24] G Holmes, R Kirkby, B Pfahringer. Stress-testing hoeffding trees. Knowledge Discovery in Databases: PKDD 2005, 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, Porto, Portugal, October 3-7, 2005, Proceedings: 495-502.

[25] M Hall, E Frank, G Holmes, et al. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, vol. 11, no. 1, 2009: 10-18

[26] ttest2: two-sample t-test. 08/08/2017. https://www.mathworks.com/help/stats/ttest2.html.

[27] "ICD-10 online versions," World Health Organization, 2014.

[28] R Peck, C Olsen, JL Devore. Introduction to statistics and data analysis: Cengage Learning. Springer International Publishing, 2015.

[29] MJ Zaki, WM Jr. Data Mining and Analysis: Fundamental Concepts and Algorithms, Cambridge University Press; 1 edition, May 12, 2014.
[30] C Cheadle, MP Vawter, WJ Freed, KG Becker. Analysis of microarray data using Z score transformation . J Mol Diagn, vol. 5, no. 2, May, 2003: 73-81

Fund Project:
Wenzhou Major Science and Technology Project "Development of Disease Prevention and Prediction System Based on Cloud Computing and Medical Big Data (Project No.ZG2017020)"

Author Details

A. Tongbin Zhang, first author, medical informatics, Department of Biomedical Engineering.
B. Li Cai, second author, working as a manager from Wenzhou Yuekang Information Technology Limited Liability Company. His main work is the application of information technology.
C. Chuandi Pan, corresponding author, working as a professor from department of biomedical Engineering in Wenzhou medical university, Wenzhou, China. His main research direction is medical informatics.