IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 12, 2019, accepted May 20, 2019, date of publication May 23, 2019, date of current version July 16, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2918625

Diagnosis and Analysis of Diabetic Retinopathy Based on Electronic Health Records

YUNLEI SUN[1] AND DALIN ZHANG[2]

[1]College of Computer and Communication Engineering, China University of Petroleum (East China), Qingdao 266580, China
[2]National Research Center of Railway Safety Assessment, Beijing Jiaotong University, Beijing 100044, China

Corresponding author: Yunlei Sun (sunyunlei@upc.edu.cn)

This work was supported by the Fundamental Research Funds for the Central Universities under Grant 18CX02019A.

ABSTRACT Diabetic retinopathy (DR) is an important disease leading to blindness in humans, attracting a lot of research interests. Previous breakthrough research findings rely on deep learning techniques to diagnose diabetic retinopathy in patients with medical imaging. Although the medical imaging achieves reasonable recognition accuracy, the application of mass, easy-to-obtain and free electronic health records (EHR) data in life can make an early diagnosis of the DR more convenient and quick. In this paper, we used a set of five machine learning models to diagnose the DR in patients with the EHR data and formed a set of treatment methods. Our experimental data set is formed by processing the data provided by 301 hospitals. The experimental results show that random forest (RF) in the machine learning model can get 92% accuracy with good performance. Subsequently, the input features were analyzed and their importance graded to find that the predisposing factors triggering the human DR disease were associated with renal and liver function. In addition, disease diagnosis methods based on readily available the EHR data will become an integral part of smart healthcare and mobile healthcare.

INDEX TERMS Diabetic retinopathy, disease diagnosis, electronic medical records, machine learning, mobile medical.

I. INTRODUCTION

Early diagnosis of patients with diabetes by EHR, has gradually become an effective measure to prevent DR disease [1], [2]. People are in a fast-growing society and cost is one important factor for the diagnosis and treatment of the disease. Therefore, swift medical practices such as smart medical care and mobile health care, etc. are changing the way people think and live. In our daily life, wearable devices and electronic health devices brought by high-tech have become indispensable to every household, so that a large amount of EHR data can be obtained. Through EHR data, we can diagnose whether potential patients are with diabetic retinopathy. Comparing this approach with conventional diabetic retinopathy diagnosis (e.g., fundus images), it does not only eliminate the time and money costs, but also maintains a high accuracy. Conventional diagnosis of DR requires a professional medical institution to obtain the fundus image of the human body by advanced medical equipment and then to judge by a professional physician based on professional knowledge and experience [3]. Early diagnosis of DR with convenient, easy-to-access, free, low-level EHR data is a simple and convenient way to treat people.

Diabetes mellitus (DM) is a major disease with high penetrance in humans around the globe, a trend that is still on the rise [4], [5]. Because the body's blood glucose level is difficult to control, which may lead to imbalance in the body's sugar levels, a variety of unstable changes to the body, and will trigger a series of complications. According to our statistics on the electronic medical records of 301 hospitals, DR is one of the most common diabetes complications, as shown in Figure 1. When a patient has diabetes, it often causes various complications. It is not diabetes itself, but various complications that affect the patient's physical condition. This paper focuses on diabetic retinopathy, and another paper in our group focused on the complications of Diabetic Nephropathy. DR as a large number of complications of diabetes is due to a long time to maintain high blood sugar in the human body, which will damage the retinal blood vessels, leading to retinopathy [3]. DR can lead to human vision damage, further blindness has become one of the main causes of blindness in Western countries [6].

People pay more and more attention on their own health condition and increase the number of medical

The associate editor coordinating the review of this manuscript and approving it for publication was Tai-Hoon Kim.

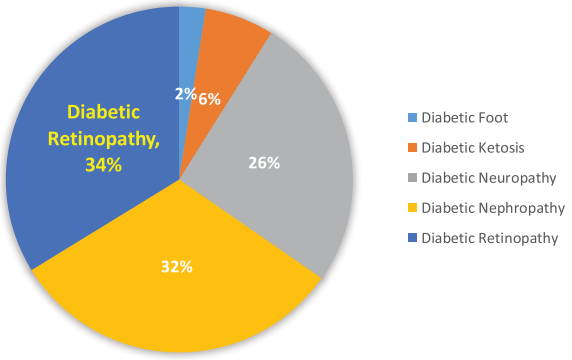

FIGURE 1. Diabetic complications (from 301 hospitals 5057 records of diabetes complications).

examinations going to authoritative medical institutions. As a result, the generated EHR becomes a huge amount of data. EHR data contains a large amount of patient information, which can be extracted through data mining implicit hidden patterns, and EHR data is also a very easy to get, low-cost data, it only needs to be extracted from the daily physical examination report after simple processing. But EHR data is challenging to represent and model due to its high dimensionality, noise heterogeneity, sparseness, incompleteness, random errors, and systematic biases [7]–[10]. Data mining has become a hot research in recent years for cheap and easy accessible EHR.

The traditional DR diagnostic method is through the human fundus image, we focus on getting knowledge from the EHR, mining the information, through these numerical data to diagnose whether a diabetic patient is DR. In this paper, Machine learning models (Logistic Regression, Support Vector Machines, Random Forests, Decision Trees, Naïve Bayes) have been used to diagnose diabetic potential patients suffering from DR. The results of these algorithms are compared to analyze the reasons for the differences in accuracy.

In our experimental results, we found that the Random Forest (RF) model was used to diagnose DR potential patient to achieve the desired accuracy. Diagnostic accuracy of RF model as high as 92%. Subsequently, we also prioritized the input features in the dataset to determine the predisposing factors that trigger the development of human DR in relation to kidney and liver function. Our experiments use EHR data to diagnose potential patients with DR, which is instructive for more researchers in the future to mine disease-related EHR data.

The main purpose of this paper is to find a benchmark model of machine learning for DR diagnostic through EHR data. The remaining part of the paper is organized as follows. In Section 2, it is the description of the data set and data preprocessing process and the methods used in this experiment; Section 3 discusses the results and analysis of the experiment; Section 4 deals with summarizing the related work done under the scope of this paper; Section 5 concludes the paper.

II. METHODOLOGIES

A. DATASETS

Our data is from the Medical Big Data Center of the 301 Hospital, which includes a total of 3 years (2009 to 2011) of patient physical examination data, and contains about 4 million records. The data includes patient information form, detailed information form, diagnosis form, sickness sign record form, biochemical indicator form, glycated indicator form and follow-up data. The Medical Big Data Center has done desensitization before sharing the data, so the data we got is desensitized data, and does not contain the patient's private information (patient name, patient's date of birth, phone number, address, ID card Number, medical record number). For our study of DR, to filter out the appropriate data that EHR data related to diabetic retinopathy, we adopted a strategy to pre-filter to select the relevant patient. The EHR samples we selected should meet the following criteria: i) a diagnostic record with DR, ii) multiple records at different times for one patient's serial number, using only the data detected at the first hospitalization as a record. Through this process, we successfully obtained 1708 cases of DR related records. To ensure the plausibility of the prediction, we also screened out from the data set non-DR patients as a control sample to ensure a 1:1 ratio between DR and non-DR. Finally, we created a dataset suitable for this experimental, consisting of 3416 records of DR patients and non-DR patients. These 3,416 records are all suffering from diabetes. We divided the patient sample data into a training set and a test set according to a ratio of 8:2. The training set was used to train a diagnostic model of DR disease, and the test set was used to predict the accuracy of the disease diagnosis result. The ''Diagnostic Results'' column in the test set is hidden during the test. After the test, manual comparison can be performed to ensure that it is DR instead of DM.

B. DATASETS PRE-PROCESSING

In order to ensure the performance of disease diagnosis, it is necessary to build good features from the EHR data for the machine learning model. This is because raw EHR data is often noisy, sparse, and contain unstructured information (e.g., text) [11].

In our work, we retain some general demographic data (e.g., gender, age), numerical data in medical records (e.g., glucose), and diagnostic labels for each case in the data set. For experimental accuracy, we removed all clinical records and difficult text descriptions. Time columns and duplicate columns, redundant columns in the data set were also deleted.

Perhaps the loss of human causes or the lack of objective reasons, there are a large number of missing values in the data set, if these missing are kept, they will affect the subsequent diagnosis test and largely reduce the accuracy of the diagnosis. At the same time, the format inconsistencies of the original data will also affect subsequent work. Therefore, we choose to use the relevant tools (R language, python) on the data for some related pretreatment.

Transformation steps include:

- Replacing missing values, and
- ID Mapping, and
- Data type classification.

During data preprocessing, we consider that feature engineering may improve DR disease diagnosis accuracy for patients. As is often the case, data and features determine the upper limit of machine learning, and models and algorithms only approximate this upper limit. Therefore, feature engineering occupies a very important position in machine learning. Our dataset, which eliminates hard-to-handle unstructured data (text), the rest is mainly numerical data. The numerical data of feature engineering is to make a specific transformation of data. Our data line is a sample, a column is a feature, according to the different column data types to do the corresponding transformation, include:

- LabelBinarization of values, and
- MinMaxScaleration of values, and
- StandardScaleration of values, and
- Normalization of values.

These could make it easy to use the dataset. The data in different sections of each column are standardized to a common range. Through data pre-processing, the features of the dataset are changed. Finally, the features of the dataset is 99 dimensions in total. A total of 3416 records are divided into training and testing sets according to a ratio of 8:2.

C. EXPERIMENTAL

The widely-used classification model such as Logistic Regression (LR), Support Vector Machines (SVM), Random Forests (RF), Decision Trees (DT), Naïve Bayes (NB). These classification models are frequently utilized in a wide range of fields, and are recognized as popular choices for classification tasks [12]–[14].

In this experiment, we selected these five classification models. We divided the patient sample data into training sets and test sets according to the ratio of 8: 2. Training set is used to train a diagnostic model of DR disease, and the test set is used to predict the accuracy of the disease diagnosis. Afterwards, based on the feature transformation we made on the dataset, DR potential patients were diagnosed and the classification models were used to test the impact of our input features on the diagnosis of DR potential patients. Figure 2 shows the flow chart of the DR classification.

III. RESULTS AND ANALYSIS

A. RESULTS

The main focus of our work is to demonstrate that machine learning models, which are widely-used in daily life, are both appropriate and feasible for a particular task (DR disease diagnosis) and find a benchmark model with good diagnosis performance from multiple machine learning models. In order to maximize the use of valid data, and get the real test results, we have taken cross-validation in the experiment to improve the accuracy of the model. The parameters of

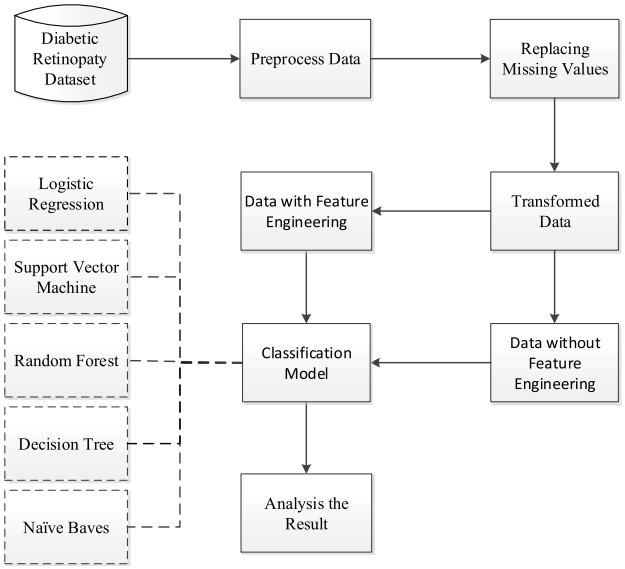

FIGURE 2. Diagnostic flowchart of DR on patients through HER.

the model settings, we based on the default recommended parameters, for some of the parameters that may affect the classification accuracy, we perform the hyperparameter tuning.

We guarantee the same data sets, and divide the input data into two types: feature engineering transformation and non-featureless engineering transformation, then put them into each model to evaluate their performance. The results of the diagnosis are shown in the Figure 3 and Table 1. Input data without feature engineering transformation, RF and DT show better accuracy than the other three classification models, RF classification accuracy is 92%. After the data is processed for feature engineering, the RF and DT still keep a high accuracy.

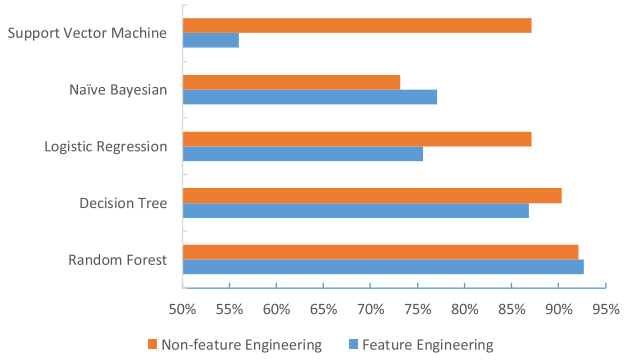

FIGURE 3. DR diagnostic accuracy.

TABLE 1. Five classification models of dr diagnostic accuracy (3416 samples).

Classifier	Feature Engineering	Non-feature Engineering
Random Forest	92.69%	92.10%
Decision Tree	86.82%	90.35%
Logistic Regression	75.58%	87.13%
Naïve Bayesian	77.07%	73.14%
Support Vector Machine	55.99%	87.13%

By comparing the data before and after the implementation of feature engineering, it was found that the diagnostic accuracy of LR and DT has been improved to some extent after feature engineering. The accuracy of DT is improved from 86% to 90%, while the accuracy of LR is improved from 76% to 87%. SVM accuracy has been significantly improved from 56% to 87%. However, the diagnostic accuracy of RF is not affected by the characteristics of the project, but only on the basis of the original fluctuations. After performing feature engineering, NB reduced its accuracy from 77% to 73%.

Our experimental results show that the most suitable machine learning model for DR disease diagnosis of EHR data is RF. Select the model does not require the data feature engineering operation, 92% of the diagnostic accuracy can be reached.

B. ANALYSIS

1) THEORY ANALYSIS

RF is a classifier that uses multiple trees to train and predict samples [15], [16]. It contains classifiers for multiple decision trees and the categories they output are determined by the order of the categories that the individual trees output. In random forest, multiple samples are extracted from the original sample by using the repeated sampling method, the decision tree is modeled for each sample, and the prediction of multiple decision trees is combined to obtain the final result. In order to improve the precision, RF introduces randomness, selects N features for each decision-making split and selects the best features for splitting. However, among the N best split features, RF takes the form of random selection.

For our data in this experiment, there is no correlation between the features in the data, which is of high degree of independence. RF selects the features randomly when making decision splitting, and then scoring integrates the result of splitting later, so as to select the best features of each step. The purpose of our experiment is to diagnose whether patients are with DR, based on the RF theory to solve the dichotomy problem, as long as the construction of a reasonable number of trees within a reasonable range, you will get a higher diagnostic accuracy. RF picks the features each time they make a decision, sorting the input features to an important degree, and the model sorts them accordingly, even without performing feature engineering, Therefore, there is no difference between the diagnostic accuracy of the RF model without the featureless engineering.

SVM is a supervised learning used for classification. The SVM removes the over fit nature of samples which increases the prediction accuracy [15], [16]. SVM posses a linear hyperplane with a margin which divides the dataset into positive and negative samples [15], [16].

In this experiment, SVM was the most improved model among the five classification models after feature engineering. The accuracy of this model used in this experiment was 56%, after the feature transformation the accuracy rate is increased to 87%. The SVM classification is to find a hyperplane by which two types of samples are divided and a support vector close to the classification plane is used to determine the hyperplane. The feature transformation makes the features of the sample larger, so that when the algorithm performs the dichotomy problem on the sample. The distance between samples can be increased according to the obvious difference features. Before the algorithm can be based on less sample features, some sample points in the wrong classification area are selected as the support vector. By increasing the number of sample features, the SVM will have more basis to determine the differences between the samples, so as to distinguish the data samples to a greater degree, then the classification hyperplane obtained will be more accurate, This allows the SVM to dramatically improve the accuracy of feature-oriented data processing.

NB [15], [16] basic idea is to solve the probability of occurrence of each category under the given condition for the given item to be classified and to classify the item to be classified as the item with the highest probability. Bayes formula is as follows:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{1}$$

In this experiment, NB is the only model with reduced accuracy after feature engineering. The accuracy of the data experiment without feature engineering was 77% in this model, and the accuracy was reduced to 73% after the data transformation of feature engineering. NB is the appropriate division on characterization and then calculate the frequency of each category in the training set and each feature classification of each category of conditional probability estimations. Our original input feature was 45-dimensions, and after its feature transformation, the feature was increased to 99 dimensions. The increase of feature dimension leads to the difference in the conditional probability of each category calculated by the NB algorithm. When the previous dimension is small, the probability value of the category is larger, the dimension is increased and the total sample size is not changed, so that the probability of each category becomes smaller, Since NB is mainly based on probability classification, the wrong classification will occur when the samples are dichotomous, resulting in a decrease in accuracy.

2) FEATURE RANK

According to the trained model, we can know the weight of each input feature, which reflects the impact of its corresponding features on the diagnostic results. We choose the RF model with the highest diagnostic accuracy. According to the weights reflected by the RF model, we sort the input features, Figure 4 shows the classification of the top 20 high-weight important features, we can see that 70% of the important features are mainly concentrated in the body measurements related to the two functional categories of kidney and liver. This is consistent with the clinically relevant influencing factors of diabetes, It reflects the potential correlation between kidney and liver-related sensitivities and DR.

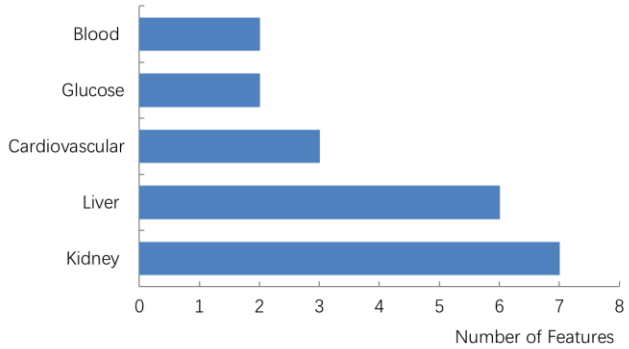

FIGURE 4. Multiple categories of histograms for the top 20 input features of DR.

TABLE 2. List of the top 10 sensitive features in rf classification.

Number	Name	Description	Weight
1	UIBC	unsaturated iron binding capacity	12.8
2	TBil	total bilirubin	3.7
3	GSP	glycosylated serum protein	3.6
4	CO_2	carbondioxide	2.8
5	LDH	lactate dehydrogenase	2.4
6	CK	creatine kinase	2.1
7	Ca	calcium	2.0
8	TG	triglyceride	1.9
9	GGT	γ-glutamyl transpeptadase	1.7
10	CHL	chloride	1.7

Table 2 lists the top ten important features, the top 10 features of interest include unsaturated iron bindings associated with protein synthesis in the blood, glycated serum proteins, and total bilirubin, triglycerides, and chlorides associated with the vital signs of human organs such as the kidneys and the liver.

IV. RELATED WORK

DR is the most studied field, mainly based on image processing techniques. Gardner et al used Neural Networks and pixel intensity values to achieve sensitivity and specificity results of 88.4% and 83.5% respectively for yes or no classification of DR [17]. They used a small dataset of around 200 images and split each image in to patches and then required a clinician to classify the patches for features before SVM implementation. Adarsh et al [18] also used image processing techniques to produce an automated diagnosis for DR. The area of lesions and texture features were used to construct the feature vector for the multiclass SVM. This achieved accuracies of 96% and 94.6% on the public image databases. Harry Pratt et al [19] propose a CNN approach to diagnosing DR from digital fundus images and accurately classifying its severity. Their network be able to learn the features required to classify the fundus images, accurately classifying the majority of proliferative cases and cases with no DR. The CNN achieves a sensitivity of 95% and an accuracy of 75% on Kaggle images databases. The retinal fundus images are successfully classified by the fuzzy classifier [20]–[22] with accuracy up to 95.63%.

To the best of our knowledge, diagnostic studies for DR are mostly based on images, not on EHRs. There has also been some progress in the study of disease or diabetes analysis from EHR data. Zheng *et al.* [11] propose a data informed framework for identifying subjects with and without Type 2 Diabetes Mellitus from EHR via feature engineering and machine learning. They select effective, relevant features from the EHR, then the redundancy of these features, repetitive reorganization and integration. They also evaluate and contrast the identification performance of widely-used machine learning models. Framework can identify subjects with and without Type 2 Diabetes Mellitus at an average AUC of around 0.98. Miotto *et al.* [23] present an unsupervised representation to predict the future of patients from the electronic health records. They present a patient's presentation and process the patient data in the electronic medical records. This method create a general-purpose set of patient features that can be effectively used in predictive clinical applications. Then, use this patient dataset to evaluate the disease predictions in two applicative clinical tasks: disease classification (i.e., evaluation by disease) and patient disease tagging (i.e., evaluation by patient).

The above research work is still focused on the diagnostic of DR images, but the data sample format has not been moved to a broader category (i.e., EHR data). Diagnostic the DR of potential patients through unconventional forms of data (images), diagnosing the complication of diabetes with EHR as a supplementary medicine, is an extremely significant study.

V. CONCLUSIONS

In this paper, a machine learning model is used to diagnose potential DR in patients with EHR data. We organized EHR data to reduce and transform a large number of unstructured, irregular and noisy feature data. Then five kinds of machine learning models were used to predict the patient samples respectively. The accuracy of the RF model was the highest and stable, reaching 92%. By analyzing its underlying principles, we find that RF shows better performance for this problem than other classification models and is easier to adjust. According to Ockham's Razor, from the analysis of the experimental results, the tedious work of eigenproject can be discarded and satisfactory predictions can be obtained. From the trained RF model, the weight of each input feature is obtained, and the weight reflects the degree of influence of its corresponding feature on the classification result. By analyzing the top 20 sensitive features, it is found that the human organs associated with the more affected features are mainly concentrated in the two major categories of kidney function and liver function. Then, the first 10 sensitive factors were analyzed and found that one of the most important characteristics of influence was the binding of unsaturated iron, which was the index of protein synthesis in blood. There were 3 renal related and 4 liver functional. These features are in line with the parameters of the clinical diagnosis based on the indicators, the first three features indicate that retinopathy

and nephropathy are also related. This result reflects the agreement between the important features of experimental analysis and clinical a priori knowledge and the correlation with the remaining complication or disease.

This study combined EHR data with good expectations for the diagnosis of DR disease, achieved high diagnostic accuracy (> 90%), get 92% accuracy. Slightly higher than the 91% accuracy rate that human doctors diagnosed through domain knowledge. For people in modern society to pursue efficient and convenient services, such as smart medical, precision medical, mobile medical and other means, Our proposed method, which uses EHR data to bring convenience, low cost, low threshold, and high accuracy compared to traditional DR diagnostic methods (via human fundus images). For people in modern society to pursue efficient and convenient services, such as smart medical, precision medical, mobile medical and other means. Our proposed method, which uses EHR data to bring convenience, low cost, low threshold, and high accuracy compared to traditional DR diagnostic methods (via human fundus images). Alleviates the cost burden of people going to professional medical institutions, is a kind of mobile medical and rapid medical treatment in life.

REFERENCES

[1] J. Benbassat and B. C. P. Polak, "Reliability of screening methods for diabetic retinopathy," *Diabetic Med.*, vol. 26, no. 8, pp. 783–790, 2009.
[2] M. J. Sculpher, M. J. Buxton, B. A. Ferguson, D. J. Spiegelhalter, and A. J. Kirby, "Screening for diabetic retinopathy: A relative cost-effectiveness analysis of alternative modalities and strategies," *Health Econ.* vol. 1, no. 1, pp. 39–51, 2010.
[3] E. Dhiravidachelvi and V. Rajamani, "A novel approach for diagnosing diabetic retinopathy in fundus images," *J. Comput. Sci.*, vol. 11, no. 1, pp. 262–269, 2014.
[4] Y. Xu, L. Wang, J. He, Y. Bi, M. Li, T. Wang, L. Wang, Y. Jiang, M. Dai, J. Lu, and M. Xu, "Prevalence and control of diabetes in Chinese adults," *JAMA*, vol. 310, no. 9, pp. 948–959, 2013.
[5] *Centers for Disease Control and Prevention, National Diabetes Statistics Report: Estimates of Diabetes and Its Burden in the United States*, US Dept. Health Hum. Services, Atlanta, GA, USA, 2014.
[6] I. Kocur and S. Resnikoff, "Visual impairment and blindness in Europe and their prevention," *Brit. J. Ophthalmology*, vol. 86, no. 7, pp. 716–722, 2002.
[7] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: Towards better research applications and clinical care," *Nature Rev. Genet.*, vol. 13, no. 6, pp. 395–405, 2012.
[8] D. Zhang, J. Sui, and Y. Gong, "Large scale software test data generation based on collective constraint and weighted combination method," *Tehnicki Vjesnik-Tech. Gazette*, vol. 24, no. 4, pp. 1041–1049, 2017.
[9] N. G. Weiskopf, G. Hripcsak, S. Swaminathan, and C. Weng, "Defining and measuring completeness of electronic health records for secondary use," *J. Biomed. Inform.*, vol. 46, no. 5, pp. 830–836, 2013.
[10] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research," *J. Amer. Med. Inform. Assoc.*, vol. 20, no. 1, pp. 144–151, 2013.
[11] T. Zheng, W. Xie, L. Xu, X. He, Y. Zhang, M. You, G. Yang, and Y. Chen, "A machine learning-based framework to identify type 2 diabetes through electronic health records," *Int. J. Med. Inform.*, vol. 97, pp. 120–127, Jan. 2017.
[12] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: A methodology review," *J. Biomed. Inform.*, vol. 35, nos. 5–6, pp. 352–359, 2002.
[13] D. Zhang, "High-speed train control system big data analysis based on fuzzy RDF model and uncertain reasoning," *Int. J. Comput., Commun. Control*, vol. 12, no. 4, pp. 577–591, 2017.
[14] J. L. Lustgarten, V. Gopalakrishnan, H. Grover, and S. Visweswaran, "Improving classification performance with discretization on biomedical datasets," in *Proc. AMIA Annu. Symp.*, vol. 2008, 2008, p. 445.
[15] W. Li, H. Liu, P. Yang, and W. Xie, "Supporting regularized logistic regression privately and efficiently," *PLoS ONE*, vol. 11, no. 6, 2016, Art. no. e0156479.
[16] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. London, U.K.: Pearson, 2006.
[17] C. S. Vinod, *Artificial Intelligence and Machine Learning*. New Delhi, India: PHI, 2014.
[18] G. G. Gardner, D. Keating, T. H. Williamson, and A. T. Elliott, "Automatic detection of diabetic retinopathy using an artificial neural network: A screening tool," *Brit. J. Ophthalmology*, vol. 80, no. 11, pp. 940–944, 1996.
[19] P. Adarsh and D. Jeyakumari, "Multiclass SVM-based automated diagnosis of diabetic retinopathy," in *Proc. Int. Conf. Commun. Signal Process. (ICCSP)*, Apr. 2013, pp. 206–210.
[20] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional neural networks for diabetic retinopathy," *Procedia Comput. Sci.*, vol. 90, pp. 200–205, Jan. 2016.
[21] R. Afrin and P. C. Shill, "Automatic lesions detection and classification of diabetic retinopathy using fuzzy logic," in *Proc. Int. Conf. Robot., Elect. Signal Process. Techn. (ICREST)*, Jan. 2019, pp. 527–532. doi: 10.1109/ICREST.2019.8644123.
[22] D. Zhang, D. Jin, Y. Gong, S. Chen, and C. Wang, "Research of alarm correlations based on static defect detection," *Tehnicki Vjesnik*, vol. 22, no. 2, pp. 311–318, 2015.
[23] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," *Sci. Rep.*, vol. 6, May 2016, Art. no. 26094.

YUNLEI SUN received the M.S. degree from the College of Computer and Communication Engineering, China University of Petroleum (East China), Qingdao, China, and the Ph.D. degree from the Institute of Network Technology, Beijing University of Posts and Telecommunications, Beijing, China. He is currently a Lecturer with the China University of Petroleum (East China). His research interests include cloud computing, sensor networks, service computing, machine learning, deep learning, pattern recognition, and distributed computing. He is a member of China Computer Federation (CCF).

DALIN ZHANG received the bachelor's degree in computer science, the master's degree in science, and the Ph.D. degree in computer science from the Beijing University of Posts and Telecommunications, in 2008, 2010, and 2014, respectively. In 2017, he was a Postdoctoral Researcher with the School of Electronics and Computer Engineering, Purdue University, USA. He is currently an Associate Professor of computer science and software engineering with Beijing Jiaotong University. At the same time, he led the Railway Software and Information Security Laboratory, National Research Center of Railway Safety Assessment (NRC-RSA), which is affiliated with Beijing Jiaotong University. His current research interests include railway information technology, software engineering, and information security. In the field of railway information technology, he mainly applies technologies such as big data analysis, data mining, text recognition, and business flow management to improve the efficiency of railway operation and maintenance and monitoring. His research results have been successfully deployed in China's high-speed railway operations. In the field of software engineering, his research focuses on developing applications of program analysis and software testing for improving software reliability, security, and performance. These research areas are mainly in software engineering and also data mining and programming languages. He is a member of CCF.